On the Implications of the General Data Protection Regulation on the Organisation of Evaluation Tasks

Francisco Rangel & Paolo Rosso

Autoritas Consulting, S.A., Spain & Universitat Politècnica de València, Spain

Abstract. *Evaluation campaigns allow for the creation of a common framework for research, making possible comparability and reproducibility in science. Furthermore, the huge amount of publicly available data in the different social platforms (social big data) favours evaluation tasks proliferation, for example in forensic linguistics. However, due to the implications that the release of the data may have on the privacy of people, rules for their protection must be laid down. These norms have been defined by the European Commission in the General Data Protection Regulation (GDPR) of April 27, 2016. Moreover, for the collection and distribution of data, each social media platform defines its legal base to use its data. In this paper, we describe the GDPR articles that apply for the organisation of evaluation tasks. Moreover, we propose a methodology to follow at the time of the organisation of evaluation tasks. Finally, we show a case study about the organisation of the PAN forensic linguistic tasks on author profiling at CLEF that we have been organising since 2013, showing how both GDPR and Twitter Terms of Service have been met when creating and distributing the corpora.*

Keywords: *GDPR, Corpora, Evaluation Tasks, Author Profiling.*

Resumo. *As tarefas de avaliação permitem a criação de um enquadramento de avaliação comum, permitindo a comparabilidade e reproducibilidade na ciência. A enorme quantidade de dados disponíveis publicamente nas diferentes plataformas sociais (social big data) contribui para a proliferação das tarefas de avaliação, por exemplo na área da linguística forense. Contudo, decorrente das possíveis implicações da divulgação dos dados para a privacidade das pessoas, são necessárias regras para sua proteção. Estas normas foram definidas pela Comissão Europeia no Regulamento Geral de Proteção de Dados (RGPD) de 27 de abril de 2016. Além disso, para efeitos de recolha e distribuição dos dados, cada plataforma de rede social define a sua base jurídica para utilizar os seus dados. Neste artigo, descrevemos os artigos do RGPD aplicáveis à organização de tarefas de avaliação. Propomos, ainda, uma metodologia a seguir para organização de tarefas de avaliação. Finalmente, apresentamos um estudo de caso sobre a organização das tarefas de linguística forense do PAN no CLEF para determinar o perfil dos autores, que organizamos desde 2013, mostrando de que modo observamos, quer o RGPD,* *quer os Termos e Condições do Twitter, na criação e distribuição dos corpora.*

Palavras-chave: *GDPR, Corpora, Tarefas de avaliação, Perfil dos autores.*

Introduction

It might be said that the main objective when organising evaluation tasks is to provide with a common framework where researchers can experiment and evaluate their results under the same conditions. Namely, a framework where both the data and the evaluation methodology are common to all the researchers. This evaluation framework allows for comparability and reproducibility.

The existence and publicly availability of big amounts of data in social platforms (namely social big data) favours the proliferation of evaluation tasks. This is also true in case of forensic linguistics Coulthard *et al.* (2016). In this vein, there are several evaluation tasks organised around the globe related to forensic linguistics. For example PAN,[1] the lab at CLEF[2] on digital text forensics focuses on different forensics linguistics aspects: author identification Kestemont *et al.* (2018), profiling Rangel *et al.* (2018), and obfuscation Hagen *et al.* (2018), whose aims, given a document, are respectively: to infer who wrote it, what are its author's demographic traits and to hide it.

When organising evaluation tasks, textual data (as well as multimedia one) should be labelled with information related to its content (e.g., irony, sentiment) or its author (e.g., gender, age, personality traits). In some cases, these data may be considered personal data (or personal data can be inferred from them). Therefore, the General Data Protection Regulation (GDPR),[3] the European regulation concerning the protection of individuals from the inappropriate use of their personal data Voigt and Von dem Bussche (2017), is of direct application. This regulation contains 99 (very restrictive Zarsky (2016)) articles, albeit we will focus only on those which directly apply to the scientific activities of organising evaluation tasks.

Likewise, before the download and reuse of data in the aforementioned evaluation tasks, the particular terms of use of the social platform from where the data is going to be collected must be taken into account. We will use Twitter as case study to illustrate its conditions, being the microblog platform that in most cases we used to collect the data for the PAN author profiling tasks, even though the presented methodology can (and must) be applied also to other platforms such as Facebook. In particular, the following should be considered when dealing with data for evaluation purposes:

- General Data Protection Regulation, mandatory when working with personal data (or from which personal data can be inferred) in/from/of the European Union.
- Particular terms of use of the specific social platform from where the data is collected. Concretely:
    - Legal base that allows the data treatment.
    - Permitted and prohibited behaviours related to collection, use and distribution of data.
    - The way to share and distribute data.
    - Other considerations that might reinforce the legal base for its utilisation.

The rest of the paper is structured as follows. In Section 2, we describe the legal framework of GDPR, focusing on the articles that directly apply to the organisation of evaluation tasks[4]. In Section 3 we illustrate the methodology to follow when organising an evaluation task. In Section 4 we present a case study. Concretely, we explain how we have applied the proposed methodology for the organisation of the Author Profiling task at PAN, showing the particularities of the social platform Twitter. In Section 5, we overview the created author profiling corpora and how GDPR was applied. Finally, in Section 6 we draw the conclusions of this study.

Overview of the General Data Protection Regulation

The General Data Protection Regulation was approved on April 27, 2016 with the aim at protecting natural persons with regard to the processing of personal data and on the free movement of such data. The GDPR is applicable as of May 25th, 2018 in all member states to harmonize data privacy laws across Europe.[5]

It is noteworthy that the GDPR has been developed on the basis of the principle of proactive responsibility. This principle assumes *the necessity that the responsible of the treatment applies technical and organisational measures to guarantee and demonstrate that the data treatment is according to the Regulation.*

This principle requires a conscious, diligent and proactive attitude regarding the processing of personal data. It requires to analyse what data is treated, for what purpose and what type of treatment operations are carried out. To *guarantee* and *demonstrate* mean that it must be explicitly determined how the required measures will be implemented, that these measures are adequate to comply with the Regulation and that this fact can be demonstrated to all the interested parties and to the supervisory authorities.

Bearing in mind with this principle, from the 99 articles that make up the legal text, we focus only on those that directly affect the organisation of evaluation tasks.

Article 4. Definitions

This article defines the needed concepts for the purpose of the Regulation. The first point defines personal data as any information that identifies or can be used to identify a natural person. This definition is of high interest since it determines whether the Regulation must be complied.

> 1. ’personal data’ means any information relating to an identified or identifiable natural person (’data subject’); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

Article 6. Lawfulness of processing (legal base)

One of the keys of the law is to identify the legal base that allows the personal data treatment. In the case of evaluation tasks, the only possibility is defined in Article 6 (1) a).

> 1.a) the data subject has given consent to the processing of his or her personal data for one or more specific purposes.

Article 7. Conditions for consent

If the legal base is the express consent of the subject, we should demonstrate such consent according to Article 7 (1).

> 1. Where processing is based on consent, the controller shall be able to demonstrate that the data subject has consented to processing of his or her personal data.

Article 8. Conditions applicable to child's consent in relation to information society services
This article regulates the conditions of consent when dealing with minors. For example, when a minor sign up in a social network, this article is mandatory.

> 1. Where point (a) of Article 6 (1) applies, in relation to the offer of information society services directly to a child, the processing of the personal data of a child shall be lawful where the child is at least 16 years old. Where the child is below the age of 16 years, such processing shall be lawful only if and to the extent that consent is given or authorised by the holder of parental responsibility over the child.
> Member States may provide by law for a lower age for those purposes provided that such lower age is not below 13 years.

Article 9. Treatment of special categories of personal data
Article 9 (1) refers to personal 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*" and says that "*shall be prohibited.*"

However, in (2) there are some exceptions that may apply:

> e) the treatment refers to personal data that the interested party has made manifestly public.
> j) the treatment is necessary for the purposes of archiving in the public interest, scientific or historical research purposes, or statistical purposes, in accordance with Article 89, paragraph 1 [. . .]

Article 17. Right of suppression
This article refers to the right of users to delete their data at anytime. Nevertheless, there is an exception to this rule that may apply:

> 3. d) It will not apply when the treatment is necessary for the purposes of archiving in the public interest, scientific or historical research purposes, or statistical purposes, in accordance with Article 89, paragraph 1 [...]

Article 22. Automated individual decision-making, including profiling
This article is the most controversial one since it prohibits the automated profiling of users (one of the aims of forensic linguistics). Nonetheless, there is a nuance that may allow the organisation of evaluation tasks since they do not produce legal effects:

> 1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

Article 24. Responsibility of the controller
This article (and subsequent Arts. 25, 30, 32, and 89) regulates the principle of proactive responsibility since we not only must apply technical and organisational measures, but also to be able to demonstrate them:

> 1. Taking into account the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for the rights and freedoms of natural persons, the controller shall implement appropriate technical and organisational measures to ensure and to be able to demonstrate that processing is performed in accordance with this Regulation. Those measures shall be reviewed and updated where necessary.

Article 25. Data protection by design and by default

Two principles should be followed (*data minimisation* and *pseudonymisation*) to difficult, among others, the inverse identification of people:

> 1. 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 shall, both at the time of the determination 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.
> 2. The controller shall implement appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons.

Article 30. Records of processing activities

An organisational measure to be taken into account is to record all the processing activities, such as for example when data is released to the research community. In this article is also described the information that should be registered:

> 1. Each controller and, where applicable, the controller's representative, shall maintain a record of processing activities under its responsibility. That record shall contain all of the following information:
> a) the name and contact details of the controller and, where applicable, the joint controller, the controller's representative and the data protection officer;
> b) the purposes of the processing;
> c) a description of the categories of data subjects and of the categories of personal data;
> d) the categories of recipients to whom the personal data have been or will be disclosed including recipients in third countries or international organisations;
> e) where applicable, transfers of personal data to a third country or an international organisation, including the identification of that third country or international organisation and, in the case of transfers referred to in the second subparagraph of Article 49(1), the documentation of suitable safeguards;
> f) where possible, the envisaged time limits for erasure of the different categories of data;
> e) where possible, a general description of the technical and organisational security measures referred to in Article 32(1).

Article 32. Security of processing

Besides *data minimisation* and *pseudonymisation* described in Article 25, data processing must be ensured with technical measures such as *encryption*:

> 1. Taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons, the controller and the processor shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk, including inter alia as appropriate:
> a) the pseudonymisation and encryption of personal data;

Article 89. Safeguards and derogations relating to processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes

Although Article 89 describes safeguards to be implemented, it is worth to mention some derogations that may apply in case of scientific research purposes, such as the organisation of evaluation tasks:

> 1. Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subject to appropriate safeguards, in accordance with this Regulation, for the rights and freedoms of the data subject. Those safeguards shall ensure that technical and organisational measures are in place in particular in order to ensure respect for the principle of data minimisation. Those measures may include pseudonymisation provided that those purposes can be fulfilled in that manner. Where those purposes can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects, those purposes shall be fulfilled in that manner.
> 2. Where personal data are processed for scientific or historical research purposes or statistical purposes, Union or Member State law may provide for derogations from the rights referred to in Articles 15, 16, 18 and 21 subject to the conditions and safeguards referred to in paragraph 1 of this Article in so far as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfillment of those purposes.

Methodology

In this section we propose a methodology to follow when organising evaluation tasks in order to ensure that GDPR, as well as the platform particular rules, are fulfilled when collecting, processing and distributing corpora *that contain personal data, or may contain identifiable personal data,* for scientific research purposes. It is noticeable the need to determine whether the corpora contain personal data as defined in GDPR Article 4 in order to apply (or not) the Regulation. The proposed methodology follows the schema represented in Figure 1 and it can be summarised in the following steps:

- To identify the legal base and to be able to demonstrate it.
- To consider special cases such as minors, special categories of data, or automatic profiling, whether some of them apply.
- To implement appropriate technical and organisational measures to ensure data protection.
- To distribute data according with both the social platform rules and the right of suppression.

- To record all the activities carried out with the data.
- Other considerations that may reinforce the legal framework to use the data in evaluation tasks.

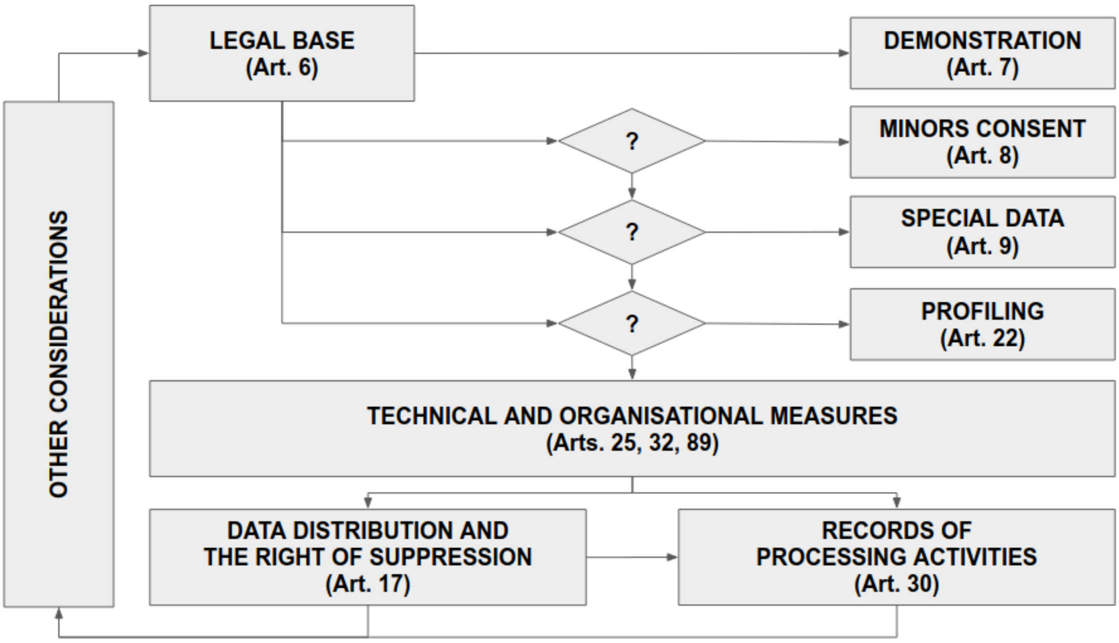

LEGAL BASE (Art. 6)
DEMONSTRATION (Art. 7)
?
MINORS CONSENT (Art. 8)
?
SPECIAL DATA (Art. 9)
?
PROFILING (Art. 22)
OTHER CONSIDERATIONS
TECHNICAL AND ORGANISATIONAL MEASURES (Arts. 25, 32, 89)
DATA DISTRIBUTION AND THE RIGHT OF SUPPRESSION (Art. 17)
RECORDS OF PROCESSING ACTIVITIES (Art. 30)

Figure 1. Methodology to accomplish GDPR when organising evaluation tasks, including automatic profiling.

The legal base and its demonstration

Following the GDPR principle of proactive responsibility, the first step is to determine the legal base that allows the use of the data in the evaluation task (Art. 6), as well as its demonstration (Art. 7). In case of evaluation tasks where the data is collected from social platforms the unique legal base that applies is *subject consent*. In such a case, it shall be demonstrated that the subjects gave their consent to use their data, in particular to use their data in evaluation tasks. This consent should be found in the terms of service of the social platform where the data is collected from. If this consent cannot be found, the data should not be used in the evaluation task.

Special cases: minors, special categories, automatic profiling

More attention should be paid when dealing with special cases such as minors (Art. 8), special categories of data (Art. 9), or automatic profiling (Art. 10). With respect to minors, the European Commission fixes the minimum age to consent at 16, albeit it allows the Member States to reduce that age as much as 13. In such cases, the consent shall be given by the legal guardian of the minor. Whether data from minors may be collected and used in the evaluation task, the organisers must ensure that the consent by the legal guardian was given. To do so the organisers should investigate how the social platform deals with minors and how it obtains the appropriate legal consent.

According to GDPR Article 9, the processing of special categories of personal data *shall be prohibited*. The first step is to determine whether the evaluation task needs or uses this kind of data. If it is needed, there are two exceptions (Section 2 of the Article) to the rule that may allow the use of this kind of data:

- *j)* Data is used for *specific research purposes*, which is the main purpose of evaluation tasks.
- *e)* Data *made manifestly public*. For each kind of special data, the organisers must ensure that the user made it manifestly public (e.g., giving public permissions to the reported birthday).

Automatic profiling is prohibited according to GDPR Article 22, but there is a nuance that may allow it in case of *non-commercial research purposes*.

Technical and organisational measures

GDPR urges to implement adequate technical and organisational measures to ensure that the data is secured, and to be able to demonstrate them. It should be followed the principles of *data minimisation* (Art. 25.1) and the *difficult to inverse identification of people* (Art. 89.1). To accomplish these principles, measures such as *encryption* (Art 32.1) and *pseudonymisation* (Art. 25.1) should be implemented.

Data distribution and the right of suppression

Data distribution must follow both the social platform rules and GDPR. In this regard and by applying the aforementioned technical and organisational measures, data should be released to the community encrypted and avoiding extra information that may allow the identification of personal data. This must be combined with the particular terms of service of the social platform which sometimes requires the release only of unique identifiers (e.g., Twitter). This situation should be analysed in each particular case.

In a similar vein, the right of suppression (Art. 17) allows users to delete their data at anytime. Deletion of the original data in the social platform should imply the automatic deletion of the data in the dataset of the evaluation task, albeit it might difficult the research activity (Art. 89.2) and the reproducibility of the experiments (Art. 17.3.d).

Records of processing activities

According to GDPR Article 30 all processing activities must be recorded. A special case is when data is released to third parties (e.g., to the participants of the evaluation task). It is imperative to implement the following measures:

- To register, at least, who is given access to the data, when, by whom, and what data in particular. It is recommendable to maintain a shared record (e.g., Google Sheet) with all the organisers, although only one of them should be the responsible to modify the register.
- To inform the researchers who receive the data that the only allowed purpose is *non-commercial scientific research*.

Other considerations

Depending on the task and the data to be used, other considerations may be extracted from the GDPR or the social platform terms of service. For example, if working with special categories of personal data such as (presumed) pedophiles that should not be available publicly, it may activate the *public interest* section in many GDPR articles that reinforce the legal base to use this data in the evaluation task.

Case Study: Author Profiling shared task at PAN

Since 2013 we have been organising at PAN an evaluation task on Author Profiling Rangel *et al.* (2013, b,a,c, 2017, 2018). With the exception of some years where data was collected also from other sources, we have mainly focused on Twitter data due to its availability, freedom of their users to express themselves and its idiosyncrasy for forensic linguistics.

In this section we describe how the proposed methodology has been applied to the organisation of the aforementioned evaluation campaigns, emphasising specific particularities of the task (e.g., dealing with special categories of data such as users personality traits or (presumed) pedophiles) and the social media platform (Twitter). Regarding the latter, besides GDPR we must fulfil the particular terms of the social media platform the data is collected from. In case of Twitter this information can be found in:

- Twitter Terms of Service[6], where the legal base for the data treatment is provided.
- Twitter Developer Policy[7], that indicates how data can be shared and distributed.
- Twitter Rules[8], that manifests prohibited behaviours for Twitter users, such as harassment or incitement to hatred, that allow us to make other considerations that reinforce our legal arguments.

To obtain the legal base and to be able to demonstrate it

As previously mentioned, according to GDPR Article 6, the unique legal base that applies is the *subject consent.* Furthermore, according to GDPR Article 7 we must be able to demonstrate that the subject consented. From the Twitter Terms of Service we can extract the needed legal base and its demonstration since Twitter is ensuring that the users consent, among others, the use of their data by third parties. Concretely, in Article *3. Content of the services*, in *Your rights and grants of rights in the contents*, Twitter users agree with the following (this must be accepted when a Twitter account is created):

> By submitting, posting or displaying Content on or through the Services, you grant us a worldwide, non-exclusive, royalty-free license (with the right to sublicense) to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute such Content in any and all media or distribution methods (now known or later developed). This license authorizes us to make your Content available to the rest of the world and to let others do the same. You agree that this license includes the right for Twitter to provide, promote, and improve the Services and to make Content submitted to or through the Services available to other companies, organizations or individuals for the syndication, broadcast, distribution, promotion or publication of such Content on other media and services, subject to our terms and conditions for such Content use. Such additional uses by Twitter, or other companies, organizations or individuals, may be made with no compensation paid to you with respect to the Content that you submit, post, transmit or otherwise make available through the Services.

The consent in case of minors

When organising evaluation tasks with Twitter data we should take into account the possibility of using personal data from minors. In this regard and according to GDPR Article 8 regarding the consent of minors, explicitly this responsibility is derived to the holder of the parental responsibility.

In the Twitter Terms of Service, in Article *1. Who may use the services?*, it is stipulated that:

> [...] you must be at least 13 years [...]

GDPR stipulates the minimum age at 16, even though it allows the Member States to lower it:

> Member States may provide by law for a lower age for those purposes provided that such lower age is not below 13 years.

Hence, depending on the Member State, this age can be ranged between 13, that Twitter requires, and 16, required by the Regulation. In such cases, Twitter should be obligated to obtain the consent from the legal guardian of the minor, according to the aforementioned article. For example, in the adaptation of the GDPR that is being processed in Spain, in the Report of the Presentation on the Organic Law Project on Personal Data Protection 121/000013[9], of October 9, 2018, in its Article 7 on the Consent of minors, in its section 1, stipulates:

> 1. The treatment of personal data of a minor may only be based on his consent when he is older than 14 years.

In this case, when the national law is effective, if the minor is between 13 and 14, Twitter shall ensure that the consent to use its services was given by the holder of the parental responsibility at the moment of the account creation. In conclusion, this nuance reinforces the argument of the legal base (the consent), no matter the data might come from minors.

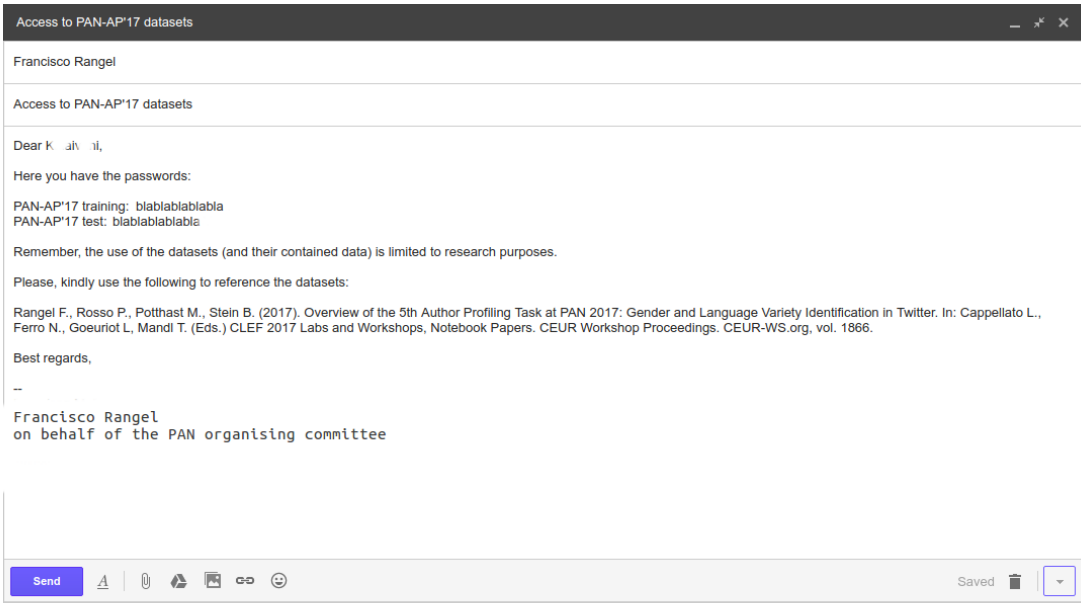
Dealing with special categories of personal data

According to GDPR Article 9 (1), the processing of special categories of personal data *shall be prohibited*. In linguistic forensics evaluation tasks we use to work with some of these special categories. For instance, when working on author profiling (e.g., personality traits) or stance detection (e.g., stance in favour or against some political matter). However, both exceptions e) (*data made manifestly public*) and j) (*scientific research purposes*) from Section 2 of the above article allow us to work with these kinds of data. Furthermore, Twitter Terms of Service, Section *3. Content of the services* reinforces the aforementioned exception e):

> You are responsible for your use of the Services and for any Content you provide, including compliance with applicable laws, rules, and regulations. You should only provide Content that you are comfortable sharing with others.

Automatic profiling

According to GDPR Article 22 profiling is prohibited. However, as we showed previously, there is a nuance that may allow our scientific activities since they do not produce legal or similarly significantly effects. Due to that, we inform researchers that the only allowed processing is for non-commercial research purposes (see Figure 2).



Access to PAN-AP'17 datasets

Francisco Rangel

Access to PAN-AP'17 datasets

Dear K ai i,

Here you have the passwords:

PAN-AP'17 training: blablablablabla
PAN-AP'17 test: blablablablabla

Remember, the use of the datasets (and their contained data) is limited to research purposes.

Please, kindly use the following to reference the datasets:

Rangel F., Rosso P., Potthast M., Stein B. (2017). Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In: Cappellato L., Ferro N., Goeuriot L, Mandl T. (Eds.) CLEF 2017 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1866.

Best regards,

--

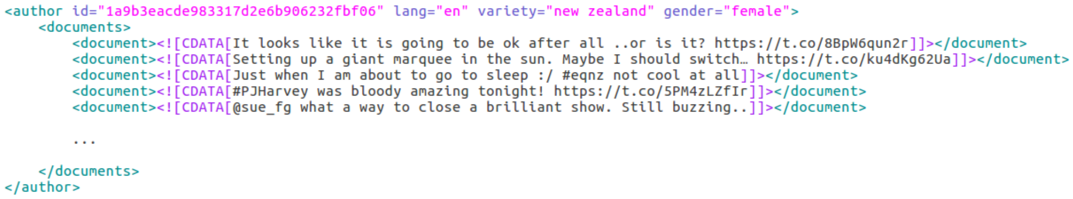
```
Francisco Rangel
on behalf of the PAN organising committee
```

Figure 2. Email to give access to the dataset.

Technical and organisational measures

According to GDPR Articles 24, 25, 32 and 89, it is mandatory to implement the appropriate technical and organisational measures to ensure and be able to demonstrate that the data is secured. Concretely, we have implemented the following measures:

- To ensure that data is *pseudonymised* (Arts. 25, 32, and 89), we remove user mentions and other personal information (e.g., replacing mentions by @mention)[10].
- To ensure *data minimisation* principle (Arts. 25 and 89), we only distribute texts written by the authors and the corresponding labels (e.g., gender, age, etc.). An example of data format is shown in Figure 3.



```
<author id="1a9b3eacde983317d2e6b906232fbf06" lang="en" variety="new zealand" gender="female">
    <documents>
        <document><![CDATA[It looks like it is going to be ok after all ..or is it? https://t.co/8BpW6qun2r]]></document>
        <document><![CDATA[Setting up a giant marquee in the sun. Maybe I should switch… https://t.co/ku4dKg62Ua]]></document>
        <document><![CDATA[Just when I am about to go to sleep :/ #eqnz not cool at all]]></document>
        <document><![CDATA[#PJHarvey was bloody amazing tonight! https://t.co/5PM4zLZfIr]]></document>
        <document><![CDATA[@sue_fg what a way to close a brilliant show. Still buzzing..]]></document>

        ...

    </documents>
</author>
```

Figure 3. Data minimisation principle distributing only textual contents and labels.

- To ensure that data cannot be accessed freely without intervention (Art. 25 (2) and 32), data:
    - is *encrypted* when stored and distributed. We compress it with a 16 random generated characters.
    - is distributed only to known people that contacted us to ask for the password (as shown in the next subsection, this allows us to track processing activities).

Data distribution and the right of suppression

In the Twitter Developer Policy, in *F. Be a Good Partner to Twitter* is explicitly said how we should distribute the tweets. According to the original text shown below, Twitter only allows the distribution of its contents (tweets, users or direct messages) via its unique identifier (ID):

> 2. If you provide Twitter Content to third parties, including downloadable datasets of Twitter Content or an API that returns Twitter Content, you will only distribute or allow download of Tweet IDs, Direct Message IDs, and/or User IDs.

However, there are some exceptions that may favour and ease the organisation of evaluation tasks. Basically, it can be downloaded other information than IDs via non-automated means, as well as it can be surpassed both the distribution limit and the storage time limit for non-commercial research purposes:

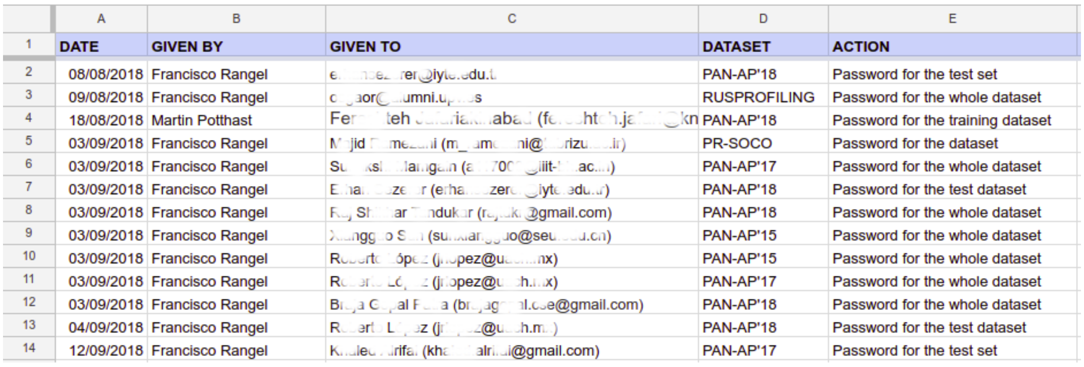
> a) You may, however, provide export via non-automated means (e.g., download of spreadsheets or PDF files, or use of a "save as" button) of up to 50,000 public Tweet Objects and/or User Objects per user of your Service, per day.
> b.i) You may not distribute more than 1,500,000 Tweet IDs to any entity (inclusive of multiple individual users associated with a single entity) within any given 30 day period, unless you are doing so on behalf of an academic institution and for the sole purpose of non-commercial research or you have received the express written permission of Twitter.
> b.ii) You may not distribute Tweet IDs for the purposes of (a) enabling any entity to store and analyze Tweets for a period exceeding 30 days unless you are doing so on behalf of an academic institution and for the sole purpose of non-commercial research or you have received the express written permission of Twitter, or (b) enabling any entity to circumvent any other limitations or restrictions on the distribution of Twitter Content as contained in this Policy, the Twitter Developer Agreement, or any other agreement with Twitter.

GDPR Article 17 refers to the right of users to suppress their data. In this regard, Twitter users can delete their account or some of their tweets, and they also should be deleted from the datasets. This will occur if Twitter general rule of distributing only IDs is followed. However, GDPR Article 17 contains the exception (3) d) that allows to not applying the right of suppression in case of scientific research purposes. We can argue in favour of providing pseudonymised texts than tweets IDs taking into account the exception a) from the Article 2 of the Twitter Developer Policies, as well as GDPR Articles 17, 25, 32 and 89, in order to:

- maintain the reproducibility of the experiments, according to Article 17 (3) d).
- ease the research activity, according to Article 89 (2).
- difficult the inverse identification of people, according to Article 89 (1).
- follow the principle of data minimisation, according to Article 25 (1).
- apply technical and organisational measures such as encryption and pseudonymisation, according to Articles 32 (1) and 25 (1) respectively.

Records of processing activities

GDPR Article 30 compels to maintain a record of all processing activities regarding personal data, for example, when the data is distributed to a research team. At the PAN lab, we maintain a list with all the people we send the data to, as well as we inform them about the only allowed purpose for the data (non-commercial research purposes).



	A	B	C	D	E
1	DATE	GIVEN BY	GIVEN TO	DATASET	ACTION
2	08/08/2018	Francisco Rangel	[redacted]	PAN-AP'18	Password for the test set
3	09/08/2018	Francisco Rangel	[redacted]	RUSPROFILING	Password for the whole dataset
4	18/08/2018	Martin Potthast	[redacted]	PAN-AP'18	Password for the training dataset
5	03/09/2018	Francisco Rangel	[redacted]	PR-SOCO	Password for the dataset
6	03/09/2018	Francisco Rangel	[redacted]	PAN-AP'17	Password for the whole dataset
7	03/09/2018	Francisco Rangel	[redacted]	PAN-AP'18	Password for the test dataset
8	03/09/2018	Francisco Rangel	[redacted]	PAN-AP'18	Password for the whole dataset
9	03/09/2018	Francisco Rangel	[redacted]	PAN-AP'15	Password for the whole dataset
10	03/09/2018	Francisco Rangel	[redacted]	PAN-AP'15	Password for the whole dataset
11	03/09/2018	Francisco Rangel	[redacted]	PAN-AP'17	Password for the whole dataset
12	03/09/2018	Francisco Rangel	[redacted]	PAN-AP'18	Password for the whole dataset
13	04/09/2018	Francisco Rangel	[redacted]	PAN-AP'18	Password for the test dataset
14	12/09/2018	Francisco Rangel	[redacted]	PAN-AP'17	Password for the test set

Figure 4. Excel sheet recording all processing activities regarding PAN datasets.

In Figure 4 an example of this record is shown in the form of an Excel sheet. Similarly, in Figure 2 we show an example of the informative email sent to the requester of the data. In this email we provide with the dataset passwords, inform about the unique allowed purpose of its use and kindly request the researcher to cite the overview paper where the dataset is described.

Other considerations

In the Authorship Attribution task at PAN 2012[11] Inches and Crestani (2012), a subtask on Sexual Predator Identification was organised. In the Author Profiling task at PAN 2013[12] Rangel *et al.* (2013) a subset of the previous data was also included. At present, we are organising the SemEval 2019 Shared Task 5 on Multilingual detection of hate speech against immigrants and women in Twitter (hatEval)[13]. In all these cases we work with very special categories of data (namely (presumed) pedophiles, misogynists, and racists). Twitter Rules do not allow users to behave abusively, such as for example sharing abusive, hateful or unwanted sexual contents. Twitter defines abusive behaviour as:

> Abuse: You may not engage in the targeted harassment of someone, or incite other people to do so. We consider abusive behavior an attempt to harass, intimidate, or silence someone else’s voice.
> Unwanted sexual advances: You may not direct abuse at someone by sending unwanted sexual content, objectifying them in a sexually explicit manner, or otherwise engaging in sexual misconduct.
> Hateful conduct: You may not promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. Read more about our hateful conduct policy.

In case there are tweets containing this kind of abusive contents because they have not been deleted by Twitter, according to GDPR Article 6 (1) e), if we try to identify them we would be working for the *public interest*:

> e) processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;

Author Profiling Corpora

As mentioned before, we have been organising the Author Profiling task at PAN forensic linguistics Lab from 2013, both at CLEF[14] (Conferences and Labs of the Evaluation Forum) and FIRE[15] (Forum for Information Retrieval Evaluation). Every year we focus on different aspects of the authors (e.g., gender, age, personality traits, language variety) as well as on different languages (e.g., Arabic, Dutch, English, Italian, Portuguese, Spanish, Russian, or even computer languages such as Java). In this section we describe each of these corpora and how the GDPR was applied when created, processed and distributed (a summary can be seen in Table 1).

CORPUS	PERSONAL	CONSENT	MINORS	SPECIAL	MEASURES			DISTRIB.
	DATA			CAT.	DM	EN	PS	
				PAN AT CLEF				
PAN-AP'13	✗	✗	✓	✓	✓	✗	✗	📄 + 🏷
PAN-AP'14	✓	?	✓	✗	✓	✗	✗	📄 + 🏷
PAN-AP'15	✓	✓	✗	✓	✗	✓	✗	🆔 + 🏷
PAN-AP'16	✓	?	✗	✗	✓	✗	✗	📄 + 🏷
PAN-AP'17	✓	✓	?	✗	✓	✓	✗	📄 + 🏷
PAN-AP'18	✓	✓	?	✗	✓	✓	✓	📄 + 🖼 + 🏷
				PAN AT FIRE				
RusProf'17	✓	✓	?	✗	✓	✓	✗	📄 + 🏷
PR-SOCO'16	✗	✓	✗	✓	✓	✓	✓	📄 + 🏷
				LEGEND				
Data Min.		YES	✓	ID	🆔			
ENcryption		NO	✗	Text	📄			
PSeudonym.		UNKNOWN	?	Image	🖼			
				Labels	🏷			

Table 1. Summary with the GDPR measures applied to the different Author Profiling corpora, identified in the first column. The second column reports whether the corpus may contain personal data and, in such a case, if the users consented. In the third and fourth columns the occurrence of minors and special categories of data are represented respectively. Columns five to seven show the technical and organisational measures applied, whereas column eight indicates the type of data distributed within the corpus. A legend is given at the bottom of the table.

Age and Gender Identification in Social Media (PAN-AP'13 at CLEF)

The focus of the 2013 evaluation task was on age and gender identification in social media. We tried to emulate a realistic big data scenario looking for open and public online repositories such as Netlog[16] with posts labelled with author demographics (gender and age). Following pioneer investigations Schler *et al.* (2006), we considered three age groups: 10s (13-17), 20s (23-27), and 30s (33-47). We also incorporated a small number of samples of adult-adult conversations about sex together with conversations of sexual predators Inches and Crestani (2012) with the aim of investigating the robustness of the state-of-the-art of age identification systems to unveil the age of sexual predators (usually pretending to be minors). In Table 3 we show the statistics of the English and Spanish corpora[17]. The corpus was balanced by gender and imbalanced by age group. More information can be found in the evaluation task overview paper Rangel *et al.* (2013).

		ENGLISH		SPANISH	
Age	**Gender**	**No. of Authors**		**No. of Authors**	
		Training	Test	Training	Test
10s	male	8 600	888	1 250	144
	female	8 600	888	1 250	144
20s	male	(72) 42 828	(32) 4 576	21 300	2 304
	female	(25) 42 875	(10) 4 598	21 300	2 304
30s	male	(92) 66 708	(40) 7 184	15 400	1 632
	female	66 800	7 224	15 400	1 632
Σ		236 600	25 440	75 900	8 160

Table 2. Distribution of the number of authors per class in PAN-AP'13 corpus.

Data were collected from the Netlog social platform that is no longer available. Hence, personal information cannot be inferred from the contents distributed in the corpus and, therefore, the GDPR does not apply. The corpus contains texts written by minors in the range of 10s (13-17), and texts from users labelled as sexual predators that can be considered special categories of data. We applied data minimisation by distributing only texts and labels corresponding to the author's age and gender. We did not encrypted data since the information was publicly available. Moreover, we did not applied pseudonymisation because we considered mentions to other people as significant for the task (however the sexual predators subset is anonymised). The first row of Table 1 summarises the described measures.

Multi-Genre Age and Gender Identification (PAN-AP'14 at CLEF)

The aim of the 2014 evaluation task was investigating how the author profiling approaches would perform on different genres: social media, blogs, Twitter and hotel reviews. The corpus covers English and Spanish languages (see Table 5), except in case of hotel reviews that are in English. That year, age ranges considered the following groups: 18-24, 25-34, 35-49, 50-64, and 65+. More information about the collection of the corpus can be found in the overview paper of the evaluation task Rangel *et al.* (b).

		ENGLISH		SPANISH	
Age	**Gender**	**No. of Authors**		**No. of Authors**	
		Training	Test	Training	Test
10s	male	8 600	888	1 250	144
	female	8 600	888	1 250	144
20s	male	(72) 42 828	(32) 4 576	21 300	2 304
	female	(25) 42 875	(10) 4 598	21 300	2 304
30s	male	(92) 66 708	(40) 7 184	15 400	1 632
	female	66 800	7 224	15 400	1 632
Σ		236 600	25 440	75 900	8 160

Table 3. Distribution of the number of authors per class in PAN-AP'13 corpus.

As there are several social media, we must determine whether each of them may contain personal data. The case of social media was discussed previously, and in case of blogs, personal data should not be inferred from contents unless the users explicitly published them. Thus, the GDPR does not apply for these social media.

In case of Twitter or reviews, personal data can be inferred from the contents and therefore they may contain personal data as defined in the Article 4 of GDPR. Due to the fact that in 2014 GDPR did not exist, the explicit consent was not mandatory and we cannot know if these platforms required it at that time. Nowadays, the social platforms must obtain the consent of the users in case they did not already give it. The users can revoke this consent or exercise the right of suppression described in Article 17. In such cases, we shall appeal to the exception 3.d) of the same article to maintain the data for scientific research purposes. In any case, we do not know whether the consent was given.

The corpus contains texts written by minors in the range of 10s (13-17) and it does not contain special categories of data. We applied data minimisation by distributing only texts and labels with age and gender information. We did not encrypted data since it was publicly available, as well as we did not applied pseudonymisation because we considered mentions to other people as significant for the task. The described measures are summarised in the second row of Table 1.

Age, Gender and Personality Recognition in Twitter (PAN-AP'15 at CLEF)

The author profiling evaluation task at PAN 2015 focused on age, gender and personality recognition of Twitter users. The most widely theory in psychology to define personality is Five Factor Theory Costa and McCrae (1985, 2008). This theory defines five traits (OCEAN): openness to experience (O), conscientiousness (C), extroversion (E), agreeableness (A), and emotional stability / neuroticism (N). To annotate the data we created an online questionnaire asking for age, gender and personality traits following the BFI-10-test Rammstedt and John (2007). Personality scores were normalised between -0.5 and +0.5, and we used the following age groups: 18-24, 25-34, 35-49, 50+. Except for age, the corpus covers English, Spanish, Italian and Dutch. The corpus statistics are shown in Table 4 and more information can be found in the overview paper of the evaluation task Rangel *et al.* (a).

	Training				Test			
	EN	ES	IT	DU	EN	ES	IT	DU
Users	152	110	38	34	142	88	36	32
18-24	58	22			56	18		
25-34	60	56			58	44		
35-49	22	22			20	18		
50+	12	10			4 8	8		
Male	76	55	19	17	71	44	18	16
Female	76	55	19	17	71	44	18	16
E (mean)	0.16	0.18	0.17	0.24	0.17	0.16	0.15	0.24
S (mean)	0.14	0.07	0.20	0.21	0.13	0.09	0.20	0.22
A (mean)	0.12	0.14	0.22	0.13	0.14	0.14	0.19	0.15
C (mean)	0.17	0.24	0.18	0.14	0.17	0.21	0.21	0.17
O (mean)	0.24	0.18	0.23	0.29	0.26	0.19	0.25	0.28

Table 4. Distribution of the number of authors per class in PAN-AP'15 corpus.

As Twitter users can be identified from their contents, the tweets should be considered as personal data. Although the Regulation should apply from 25 May 2018, in 2016 entered into force. Thus, we followed its Article 6 and requested the explicit consent of the users to process their data for research purposes. The users had to consent before filling out the aforementioned questionnaire.

This corpus does not contain data from minors since the lowest age is 18. It may be considered the existence of special categories of data regarding personality traits. We followed Twitter rule of distributing tweet IDs, thus we could not apply the data minimisation criteria nor the pseudonymisation. The applied measures are summarised in the third row of Table 1.

Cross-Genre Age and Gender Identification (PAN-AP'16 at CLEF)

In the 2016 evaluation task, we aimed at investigating the effect of the cross-genre evaluation: how the models perform when they are trained on one genre and evaluated on another different genre. In this regard, the training corpus was collected from Twitter for the three languages: Dutch, English, and Spanish. In case of Spanish and English, we merged the training and test sets from PAN-AP'14 Twitter corpus Rangel *et al.* (b), whilst in case of Dutch, the training corpus was mined as a precursor of TwiSty Verhoeven *et al.* (2016). The test corpus for English and Spanish was obtained from the test partition of the PAN-AP'14 blog subcorpus. Furthermore, as in previous years we provided with an early bird evaluation. However, unlike in previous years where early birds used a subset from the test set, this year we took advantage of this early evaluation to evaluate another genre. In concrete, early birds data in English and Spanish was collected from the social media subset of the PAN-AP'14 corpus. The test set (both early and final tests) for Dutch combined reviews from the CSI corpus Verhoeven and Daelemans (2014) and student essays. As shown in Table **??**, in case of Dutch only gender information is provided, whereas for English and Spanish the following age groups are covered: 18-24, 25-34, 35-

49, 50-64, 65+. More information about the corpora can be found in the overview paper of the evaluation task Rangel *et al.* (c).

PAN-AP'16 corpus was created from PAN-AP'14, thus what was discussed there it also applies here. The only exception is that there are no minors in 2016 corpus since the lowest age was increased to 18. A summary of measures can be seen in the fourth row of Table 1.

	ENGLISH								SPANISH					
	SocialMedia		Blog		Twitter		Reviews		SocialMedia		Blog		Twitter	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
18-24	1 550	680	6	10	20	12	360	148	330	150	4	4	12	4
25-34	2 098	900	60	24	88	56	1 000	400	426	180	26	12	42	26
35-49	2 246	980	54	32	130	58	1 000	400	324	138	42	26	86	46
50-64	1 838	790	23	10	60	26	1 000	400	160	70	12	10	32	12
65+	14	26	4	2	8	2	800	294	30	28	4	2	6	2
Σ	7 746	3 376	147	78	306	154	4 160	1 642	1 272	566	88	56	178	90

Table 5. Distribution of the number of authors per class in PAN-AP'14 corpus.

Gender and Language Variety Identification in Twitter (PAN-AP'17 at CLEF)

The focus of the 2017 evaluation task was on gender and language variety identification in Twitter. The corpus included four languages: Arabic, English, Portuguese and Spanish. We retrieved tweets geolocated in the capital cities where the target language variety is used. Unique users were selected and annotated with the corresponding variety. A dictionary with proper nouns was used to annotate the users' gender, as well as a manual inspection of their photo profiles was carried out to improve the annotation quality. Finally, for each user a hundred tweets were collected from her/his timeline. The corpus was divided into training/test in a 60/40 proportion, with 300 authors for training and 200 authors for test. The corresponding languages and varieties are shown in Table 6 along with the total number of authors for each subtask. More information about this corpus is available in the evaluation task overview paper Rangel *et al.* (2017).

(AR) Arabic	(EN) English	(ES) Spanish	(PT) Portuguese
Egypt	Australia	Argentina	Brazil
Gulf	Canada	Chile	Portugal
Levantine	Great Britain	Colombia	
Maghrebi	Ireland	Mexico	
	New Zealand	Peru	
	United States	Spain	
		Venezuela	
4,000	6,000	7,000	2,000

Table 6. Distribution of the number of authors per class in PAN-AP'17 corpus.

In the fifth row in Table 1 the applied GDPR measures when building and distributing the PAN-AP'17 corpus are summarised. As data was collected from Twitter, the consent was given to the social platform. It is not possible to know whether there are minors in the corpus because age was not verified. There are no data belonging to special categories since the unique provided label refers to users' gender. We applied data minimisation, since only texts and labels were distributed, as well as encryption since data was distributed compressed with password. We did not pseudonymised texts because nouns might contribute to the task.

Multi-Modal Gender Identification in Twitter (PAN-AP'18 at CLEF)

In 2018 we aimed to investigate the effect of multi-modal information on the gender identification task in Twitter. Multi-modal means that besides textual information, also images could be used. The corpus included three languages: Arabic, English and Spanish. This corpus was created as a subset of the PAN-AP'17 corpus. For each author, we collected all the images shared in her/his timeline. We discarded users who deleted their account as well as users with less than 10 images in their timeline. Each author contains exactly 100 tweets and 10 images. The corpus is completely balanced per gender and split in training/test sets as shown in Table 7.

	(AR) Arabic	(EN) English	(ES) Spanish	Total
Training	1,500	3,000	3,000	7,500
Test	1,000	1,900	2,200	5,100
Total	2,500	4,900	5,200	12,600

Table 7. Distribution of the number of authors per class in PAN-AP'18 corpus.

The sixth row of Table 1 summarises the applied measures. The only differences with PAN-AP'17 lie in the following: the distributed corpus contains also images, and this year we applied pseudonymisation by removing user mentions.

Cross-Genre Gender Identification in Russian (RUSPROFILING'17 PAN at FIRE)

Slavic languages have been less investigated from an author profiling standpoint and have never been addressed at PAN before. This task aimed at investigating gender identification in Russian from a cross-genre perspective. That is, we provided tweets as a training corpus and Facebook posts, online reviews, texts describing images or letters to a friend, as well as tweets as test corpus. In Table 8 a summary of the number of authors per genre is shown. More information on the corpus construction can be found in the overview paper of the evaluation task Litvinova *et al.* (2017).

RusProfiling'17 corpus contains data from different sources, even though we can group them into two types: social media platforms and students' essays. In case of social media platforms, as seen previously, personal data may be inferred from contents, coercing the application of the Regulation. In case of students' essays, although personal information should not be identifiable from their contents, the ease to obtain their consent worth it.

Table 1 summarises the GDPR measures that we applied to build and distribute the corpus. In case of social media platforms the consent was given when the account was

Dataset	Genre	Number of authors
Training	Twitter	600
Test	Essays	370
	Facebook	228
	Twitter	400
	Reviews	776
	Gender-imitated	94

Table 8. Distribution of the number of authors per genre in RusProfiling'17 corpus.

created, as well as in case of students' essays, the students gave their consent when participated. We cannot know whether there are minors in the data collected from social platforms since we did not verified the age, but we can ensure that there are no minors in the subsets of essays and gender-imitated since the authors were university students. There are no special categories of data because we only provided gender as labels. We applied both data minimisation and encryption to distribute only texts and gender labels, and we compressed the corpus with password. Pseudonymisation was not applied because mentions might contribute to the task.

Personality Recognition in SOurce COde (PR-SOCO'16 PAN at FIRE)

Finally, in the PR-SOCO evaluation task we aimed at investigating whether personality traits could be inferred from the way Java programming language is used by computer science students. Students were asked to write source code responding to some functional requirements of different programming tasks. In addition each student answered a Big Five personality test. The dataset consists of 2,492 source code programs written by 70 students (49 for training, 21 for test). The scores for the personality traits range between 20 and 80. More information about the corpus can be found in the overview paper of the evaluation task Rangel *et al.* (2016).

Despite the fact that natural persons should no be identifiable from the PR-SOCO'16 corpus, we applied GDPR measures because they were identifiable when collecting the data. Data was collected from students who explicitly expressed their consent. There are no minors since the subjects were university students of Computer Science, but the corpus does cover the special category of data regarding personality traits. We applied data minimisation, encryption and pseudonymisation: data minimisation since only source code and personality scores were distributed, encryption because the corpus was distributed compressed with password, and pseudonymisation in case some students incorporated personal nouns for instance in the source code comments. The corpus is distributed as plain text containing source code in Java language together with the labels corresponding to the five personality traits. In the last row of Table 1 we summarise the applied measures when the corpus was created and distributed.

Conclusions

The organisation of evaluation tasks allows the creation of a common framework for research, fostering comparability and reproducibility. Moreover, social data allows for investigating forensic linguistics aspects in a big data scenario. However, due to the implications that the release of the data may have on the privacy of people, the European law for its protection must be contemplated. These norms are defined in the General Data Protection Regulation (GDPR) of April 27, 2016, as well as in the legal base of use of the particular social platform from where data are collected.

In this paper, we have proposed a methodology to follow when creating corpora for the organisation of an evaluation task. Firstly, we have described the GDPR articles that apply. For each article, we have highlighted the principal aspects as well as the plausible exceptions that may help in the organisation of the task. GDPR principle of proactive responsibility assumes that the responsible of the treatment, in this case the organiser of the evaluation task, applies technical and organisational measures to guarantee and demonstrate that the data treatment is according to the Regulation. Therefore, the first step is to identify (Art. 6) and demonstrate (Art. 7) the legal base for the treatment (i.e., subject consent). A special attention must be paid when dealing with special cases (Art. 8) (i.e., minors), special categories of data (Art. 9) (i.e., political options, religious of philosophical beliefs, sexual orientation, etc.), or the treatment implies (automatic) profiling (Art. 22). In such cases, the organiser must investigate whether the possible exceptions may apply (i.e., research purposes, data made manifestly public, etc.). Furthermore, the organiser must apply technical and organisational measures (Arts. 25, 32, 89) (i.e., data minimisation, encryption, pseudonymisation, etc.) to difficult the inverse identification of people. Finally, the organiser must distribute data according with both the social platform rules and the right of suppression (Art. 17) and to record all the processing activities carried out with the data (Art. 30). At least, to register who is given access to the data as well as to inform that the only allowed purpose is non-commercial scientific research.

With the aim at guiding researchers in the application of the GDPR to the organisation of shared tasks, we have presented a case study about the organisation of the forensic linguistic task on author profiling at the PAN Lab at CLEF, that we have been organising since 2013, showing how both GDPR and Twitter Terms of Service have been complied. Finally, we have described the different corpora created at PAN and how the Regulation was observed in these cases.

Acknowledgements

This article was made possible by NPRP grant 9-175-1-033 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

Notes

[1] https://pan.webis.de/

[2] http://www.clef-initiative.eu

[3] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

[4] We use italic when text is extracted from the legal source, and underline when we want to highlight something.

[5] It is worth to mention that the GDPR must be adapted to the local legislation of each Member State. This implies to translate the Regulation, at least, to 24 official languages. Furthermore, it shall be adapted to the cultural, social and legal particularities of each of the States Sosoni and Biel (2018).

[6] https://twitter.com/en/tos

[7] https://developer.twitter.com/en/developer-terms/policy.html

[8] https://help.twitter.com/en/rules-and-policies/twitter-rules

[9] http://www.congreso.es/public_oficiales/L12/CONG/BOCG/A/BOCG-12-A-13-1.PDF

[10] We do it in those cases where we consider that this information is not valuable for the specific task.

[11] https://pan.webis.de/clef12/pan12-web/author-identification.html

[12] https://pan.webis.de/clef13/pan13-web/author-profiling.html

[13] https://competitions.codalab.org/competitions/19935

[14] http://clef2018.clef-initiative.eu/

[15] http://fire.irsi.res.in

[16] https://www.netlog.com

[17] In the training part of the English collection, numbers inside parentheses for male 20s and 30s correspond to the number of samples of sexual predator conversations while numbers inside parenthesis for female 20s correspond to the adult-adult sexual conversation samples. The final collection includes samples from sexual predator conversations for male 20s and 30s, and samples from adult-adult conversations for female 20s.

References

Costa, P. T. and McCrae, R. R. (1985). *The NEO personality inventory: Manual, form S and form R.* Psychological Assessment Resources.

Costa, P. T. and McCrae, R. R. (2008). The revised neo personality inventory (neo-pi-r). *The SAGE handbook of personality theory and assessment*, 2, 179–198.

Coulthard, M., Johnson, A. and Wright, D. (2016). *An introduction to forensic linguistics: Language in evidence.* Routledge.

Hagen, M., Potthast, M. and Stein, B. (2018). Overview of the Author Obfuscation Task at PAN 2018. In *Working Notes Papers of the CLEF 2018 Evaluation Labs*, CEUR Workshop Proceedings: CLEF and CEUR-WS.org.

Inches, G. and Crestani, F. (2012). Overview of the International Sexual Predator Identification Competition at PAN-2012. In P. Forner, J. Karlgren and C. Womser-Hacker, Eds., *CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers.*

Kestemont, M., Tschuggnall, M., Stamatatos, E., Daelemans, W., Specht, G., Stein, B. and Potthast, M. (2018). Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection. In *Working Notes Papers of the CLEF 2018 Evaluation Labs*, CEUR Workshop Proceedings: CLEF and CEUR-WS.org.

Litvinova, T., Rangel, F., Rosso, P., Seredin, P. and Litvinova, O. (2017). Overview of the rusprofiling pan at fire track on cross-genre gender identification in russian. In *FIRE (Working Notes)*, 1–7.

Rammstedt, B. and John, O. (2007). Measuring personality in one minute or less: A 10 item short version of the big five inventory in english and german. In *J. Research in Personality*, 203–212.

Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B. and Daelemans, W. In L. Cappellato, N. Ferro, G. Jones and E. San Juan, Eds., *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 2015.*

Rangel, F., González, F., Restrepo, F., Montes, M. and Rosso, P. (2016). Pan@ fire: Overview of the pr-soco track on personality recognition in source code. In *Forum for Information Retrieval Evaluation*, 1–19: Springer.

Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B. and Daelemans, W. In L. Cappellato, N. Ferro, M. Halvey and W. Kraaij, Eds., *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 2014.*

Rangel, F., Rosso, P., Koppel, M., Stamatatos, E. and Inches, G. (2013). Overview of the Author Profiling Task at PAN 2013. In P. Forner, R. Navigli and D. Tufis, Eds., *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers.*

Rangel, F., Rosso, P., Potthast, M. and Stein, B. (2017). Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In L. Cappellato, N. Ferro, L. Goeuriot and T. Mandl, Eds., *Working Notes Papers of the CLEF 2017 Evaluation Labs*, CEUR Workshop Proceedings: CLEF and CEUR-WS.org.

Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M. and Stein, B. In K. Balog, L. Cappellato, N. Ferro and C. Macdonald, Eds., *CLEF 2016 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org.*

Rangel, F., Rosso, P., y Gómez, M. M., Potthast, M. and Stein, B. (2018). Overview of the 6th author profiling task at pan 2018: Multimodal gender identification in twitter. In *CLEF 2018 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org.*

Schler, J., Koppel, M., Argamon, S. and Pennebaker, J. W. (2006). Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, 199–205: AAAI.

Sosoni, V. and Biel, Ł. (2018). Eu legal culture and translation. *International Journal of Language & Law (JLL)*, 7.

Verhoeven, B. and Daelemans, W. (2014). Clips stylometry investigation (csi) corpus: a dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In *9th International Conference on Language Resources and Evaluation (LREC 2014).*

Verhoeven, B., Daelemans, W. and Plank, B. (2016). Twisty: a multilingual twitter stylometry corpus for gender and personality profiling. In *10th International Conference on Language Resources and Evaluation (LREC 2016).*

Voigt, P. and Von dem Bussche, A. (2017). *The EU General Data Protection Regulation (GDPR)*, volume 18. Springer.

Zarsky, T. Z. (2016). Incompatible: The gdpr in the age of big data. *Seton Hall L. Rev.*, 47, 995.