Legal aspects of text mining publicly available data*

Benjamin Bremert[1]

[1] Unabhängiges Landeszentrum für Datenschutz (ULD, Independent Centre for Privacy Protection) Schleswig-Holstein, Kiel, Germany
bbremert@datenschutzzentrum.de

Abstract. The paper assesses the admissibility of data and text mining under the upcoming GDPR. It will then develop the specific risks and difficulties in compliance with transparency and information obligations under GDPR in a big data context. In the end the paper will present a possible solution for the transparency requirements in the form of an open standard for communicating allowed forms of reuse and notifying the data subject of this processing. The paper will also discuss the issues of using copyright protected material and introduce a possible legal basis for processing.

Keywords: Big Data, Data Mining, Data Protection, EU law, General Data Protection Regulation, Publicly Accessible Data, Text Mining Data Protection, Transparency

Over the last years, the internet evolved from a medium of information consumption into a medium of informatory participation. Users not only use web services to consume information on a topic of interest, but interactively use web services to access dynamic content or instead of consuming produce their own content for others to access. A study shows that in the U.S. 86% of adults between 18-29 years use social media as of November 2016.[1] In Germany the latest figures from 2014 show, that using these social networks users have not only begun to regularly share their personal life with friends on these networks online, but also to share personal information with the whole world. On social networks such as Twitter, Youtube or Instagram the focus shifts from interacting with friends to interaction with a larger audience without personal ties. These platforms made it possible for anyone to publish content to anybody else without having to know how to setup and operate the otherwise needed infrastructure and software. The possibility of publishing information virtually boundlessly also comes with the danger of information being used by individuals or companies for purposes the author never intended.

* This work is partially funded by the German Ministry of Economics and Energy within the project iTESA (intelligent Traveller Early Situation Awareness) which is embedded in the "Smart Data – Innovation from data" programme, http://www.smart-data-itesa.com/en/index.html

1 Statista: Percentage of adults in the United States who use social networks as of November 2016, by age group, https://www.statista.com/statistics/471370/us-adults-who-use-social-networks-age/

The stored and publicly accessible information within those networks and services is not only a valuable asset to the respective provider, but also to third parties who are able to exploit this information for their own services or scientific studies. These services or studies often use and analyse the information to obtain derived data by text mining. Data and text mining is considered the process of pulling and generating information of or identifying (commercially) useable patterns within structured and unstructured information. Apart from issues of copyright, the text mining of publicly available data mostly affects the data subjects rights granted within national and supranational data protection regulation. As the General Data Protection Regulation (GDPR)[2] will replace the Data Protection Directive on 25 May 2018 and the Regulation on Privacy and Electronic Communications (ePrivacy Regulation)[3] is also planned to become effective in 2018, the conditions under which the use of publicly available information is permitted by data protection regulation is going to change in contrast to former national regulations. But besides the new data protection rules, data and text mining practice is also regulated by copyright regulation which is not harmonised whereby the legal situation is very dependent from the respective legislation. This paper will first assess the legal situation in data protection law and develop solutions for exemplary difficulties and then will briefly illustrate the prospect of copyright law within the European Union regarding data and text mining.

1. General Data Protection Regulation & ePrivacy Regulation

The legal requirements with respect to data protection are currently defined by the national data protection legislation and the upcoming GDPR as well as the ePrivacy Regulation. According to Article 2 (1) the GDPR applies to the wholly or partly automated processing of personal data or the storage of personal data within a filing system. Data and text mining is the extraction and derivation of information and patterns from (often large amounts of) structured and unstructured information by applying machine learning methods and algorithms. Machine learning is the software-based generation of knowledge and prediction of information based on previously learned experience. Without doubt the GDPR will be directly applicable in these cases where personal data is being processed. When using publicly available data the presence of personal data should be expected, at least if the crawled sites are not somehow limited

[2] Regulation EU 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Official Journal L 119/1, online: http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679

[3] Proposal for a regulation of the European Parliament and of the Council concerning the respect for private life and the protection of personal data in electronic communications and repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications), online: http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=41241

to known sites without any information relating to an identified or identifiable natural person.

The ePrivacy Regulation however will only be applicable so far as data and text mining could be seen as an electronic communication service, Article 2 (1) ePrivacy Regulation, and is not covered by the exceptions in Article 2 (2) ePrivacy Regulation. According to Article 2 (4) of the Directive establishing the European Electronic Communications Code[4] an electronic communication service is a service normally provided for remuneration via electronic communications networks, which encompasses internet access service and/or interpersonal communications and/or services consisting wholly or mainly in the conveyance of signals such as transmission services used for the provision of machine-to-machine services and for broadcasting. Based on this definition the ePrivacy Regulation is generally applicable where a transmission of information occurs, whether through a telecommunications provider or a content provider offering interpersonal communication services.

In conclusion, this definition does not apply to data and text mining; hence those activities do not fall within the scope of the ePrivacy Regulation.

2. Personal data

For the GDPR to be applicable personal data must be subject to the data processing. In accordance to Article 4 (1) GDPR this is the case, if the data can be referred to as information relating to an identified or identifiable natural person. In a big data context and if data is collected from publicly available sources, it can be assumed that at least partially personal data is being processed. This is for example even then the case, if for the purpose of text mining only public text messages (such was Twitter) are being used and any user information is disregarded, because an identification of the user might still be possible from his personal style of phrasing or specific personal details within the messages. An anonymised message can only then be presumed, if the message is deprived of a specific style of writing and only the semantic content remains. Therefore data protection regulation in most cases will still be applicable.

3. Legal basis for data processing

From a data protection point of view and based on Article 5 (1) (a) GDPR it is required that personal data shall be processed lawfully, fairly and in a transparent manner in relation to the data subject. First of all that means there has to be a legal basis for the processing of personal data. Possible legal bases can be found in Article 6 (1) GDPR.

[4] Proposal for a Directive of the European Parliament and of the Council establishing the European Electronic Communications Code, online: http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=comnat:COM_2016_0590_FIN

3.1 Consent, Article 6 (1) (a) GDPR

A primary legal basis for data processing is consent, Article 6 (1) (a) GDPR. In many cases where data for text mining is obtained from commercial services such as social networks, the provider of these services reserves the right to pass user generated content to third parties. At this point the question is, if these clauses can be seen as a valid legal basis for transmitting data to a third party and if this third party is then able to use the obtained data based on the presumed consent.

The social network Twitter for example, reserves the right to "*share and disclose public information*" within their privacy police[5] and their terms of service[6] state, that "*by submitting, posting or displaying*" content through Twitter, the user also grants Twitter a "*worldwide, non-exclusive, royalty-free license (with the right to sublicense) to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute*" user generated content within "*any and all media or distribution methods (now known or later developed)*". Other services use clauses similar as this one.

The requirements for a lawful consent are widespread within the GDPR. Article 4 (11) GDPR defines consent of the data subject as a free, specific and informed unambiguous indication of the data subject's wishes by which he or she signifies agreement to the processing of personal data relating to him or her, which can occur by statement or clear affirmative action. Further requirements can be found in Article 7 GDPR and within the general principles in Article 5 GDPR.

Under those conditions, it is more than doubtful, that a clause, hidden within the terms of service or the privacy policy of an online service is a clear indication of the data subject's wish nor can these terms specify legal consequences for plain actions such as publishing content on a social media platform. Because this would presume, that the data subject, by clicking the button within the registration form accepts the services terms of service and not only agrees to comply with a certain set of rules, but also puts out a statement regarding the abstract use of a mostly abstract amount of personal data. A different evaluation could be appropriate for cases where the user must specifically agree to data processing on a (limited and distinct) per case base and where he is not only aware of such an outcome, but a certain legally binding act is aimed at this consequence. This leads to the conclusion that a consent to data processing cannot be seen in clauses that are covert within terms of service or the privacy policy of a service. Then again, the GDPR doesn't prohibit free services to stipulate, that users must agree in the processing of data and transmission to third parties in exchange for not having to pay for service usage. But such a business model requires service providers to transparently disclose the basic terms under which the service is offered to the user before and it would have to be considered, that the drafters of the GDPR have expressed their doubt whether consent is freely given, if there "is a clear

[5] Twitter Privacy Police, online: https://twitter.com/en/privacy.
[6] Twitter Terms of Service, online: https://twitter.com/en/tos#intlContent.

imbalance between the data subject and the controller" or "the provision of a service is dependent on the consent despite such consent not being necessary for such performance"[7].

This leads to the conclusion that most data processing that is done with connection with possible sources of publicly available data (e.g. social networks) and widely considered a consent based data processing might be resting upon sham consent which neither is an appropriate legal basis for the original data processing nor a possible legal basis for data and text mining by a third party.

3.2 Article 6 (1) (b) GDPR

Another legal basis for data processing, which can be found in Article 6 (1) (b) GDPR, states that data processing shall be lawful, if it is necessary for the performance of a contract to which the data subject is party or to take steps at the request of the data subject prior to entering a contract.

Article 6 (1) (b) GDPR applies in cases where the data processing is necessary to comply with provisions of a contract. More controversial are cases, where the data subject virtually pays for the use of a free service by handing over personal data and allowing the provider use the data for marketing purposes. This would mean that the data controller hands over personal data to the text and data miner in return for money and this business finances the service that the data controller provides to the data subject. As this would be a major obligation of the underlying contract regulating the use of the service, this consideration would have to be clearly communicated at the time the contract is concluded. This requirement in most cases will not be met, especially if these obligations are regulated deep within the respective Terms of Service or the Privacy Police. If, however these obligations are transparently communicated with the data subject and the data and text mining is done within the basic principles of Article 5 GDPR, Article 6 (1) (b) GDPR could be a possible legal basis for data processing in a big data context.

3.3 Article 6 (1) (f) GDPR

As central legal basis for the data processing in a context of data and text mining could be seen in Art. 6 (1) (f) GDPR. Under this clause the processing shall be lawful, if it is necessary for the purposes of the legitimate interests pursued by the controller or a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data.

In a first step a legitimate interest in the data processing from the data controller's point of view must be established. The European lawmakers did not legally define

[7] Recital 43 GDPR.

which specific interests should count as legitimate, therefore the term should be interpreted in European and data protection context. At this point the only limitation is done by requiring the interest to be legitimate. This could be understood in a way that the interests must be somehow connected to a legally recognized and protected right or claim.[8]

In any case the data controller is not able to refer to an interest as legal basis for data processing that is in no possible way protected by law or opposes fundamental principles of the legal system, as there is no legal protection of interests which themselves oppose the law.[9]

In the case of data or text mining in general the data controller might invoke a scientific or economic interest as legal basis for data processing, which in this respect means these interests are resting upon fundamental rights and more specific on Articles 13, 15 and 16 of the Charter of Fundamental Rights of the European Union. There also will be little doubt, that the exercise of one's freedom of science, freedom to conduct a business and the right to engage in work is indeed a legitimate interest. As far as a scientific interest is invoked as a legitimate interest for the data processing, which might regularly be the case in a data and text mining context, one should also consider that this might also affect common interests in contrast to being in the sole interest of the data controller. This could be the case if the achievement of a specific scientific goal or the solution of a problem is – more than in the normal case – in public interest. Under these circumstances the public interest might have to be taken into consideration when weighing the opposing interests. On the other hand, this could lead to situations where the data controller pretends to be pursuing public interests just so the weighing of opposing interests is being decided in his favour. Another problem could be the pursuit of public interests by private parties whose pursuit normally is a public duty. This would bear the risk of making a private party ultimately the advocate for public duties. In contrast to these objections, does it have to be considered if a data controller not only pursues his own interests but these interests indirectly also serve the public interest. Even if the motive of the data controller is not altruistic, a possible solution could be that he would have to primarily pursue his own legitimate interests, but if these interests also happen to serve the public, this should be taken into consideration. This should be limited to cases where the data controller is knowingly also pursuing public interests and if he does so consciously. This might be the case where the scientific research aims at the solution of important social problems.[10]

[8] Kühling/Buchner, Article 6 GDPR, mn. 146.

[9] Article 29 Data Protection Working Party, Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC ("WP 217"), p. 25 online: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf.

[10] WP 217, p. 28-29.

The next step would be to establish if the data processing is necessary to reach the legitimate interest; therefore, it must be suitable in serving the legitimate interest. The condition of necessity is a direct implementation of the principles of data minimisation found in Article 5 (1) (c) GDPR as well as purpose limitation found in Article 5 (1) (b) GDPR.[11] This implies that specific data processing is necessary, if the legitimate interests cannot be realised by data processing with a lesser gravity of interference and simultaneously maintaining the level of realisation.

At this point the individual case is to be assessed. In a data and text mining context only those data will be considered necessary, that are required for the specific mining case. That would not be the case for unspecific (purpose of processing) or unneeded collection and processing of data, which might or might not be needed for an unspecific use case in the future. Especially in connection with metadata the question of anonymization or fuzziness could be raised. Meaning that if this information could be blurred or anonymized without considerably aggravating the reason for the data processing, the information should be anonymized or blurred at the time of crawling. Only this approach would take the principle of data minimisation found in Article 5 (1) (c) GDPR into account.

Then the affected interests, fundamental rights and freedoms of the data subject are to be determined. Fundamental rights are all rights derived from the European Convention on Human Rights (ECHR) as well as the Charter of Fundamental Rights of the European Union. If personal data are being processed at least the data subject's rights as per Article 7 and 8 of the Charter of Fundamental Rights of the European Union will be affected. In contrast to the data controller the data subject can claim "interests" and is not limited to "legitimate interests". There are different opinions on how this difference in wording affects the determination of interests. Some commentaries believe that the data subject should also be able to claim interests that are based upon illegitimate motives.[12] While illegitimate is not specified further, it becomes apparent that this understanding might also include motives that oppose the legal system which clearly is not intended by the GDPR. This opinion also misjudges the recital 47 first sentence, which states that the reasonable expectations of data subjects based on their relationship with the controller have to be considered. This is another argument, that an interest that is based upon reasonable expectations cannot be deemed illegitimate and vice versa. In summary, it can be concluded that not all theoretically possible interests are to be considered, but that this reverse exception is comparable comprehensive as the exception in form of the data controller's legitimate interest.

After having determined the opposing interests for and against the data processing these interests are to be weighed against each other. The standard for the weighing of interests is, that the more one interest is affected, the higher the importance of the opposite interest has to be. The standard is also modified by Article 6 (1) (f) GDPR,

[11] Gola, Article 6 GDPR, mn. 60.
[12] Gola, Article 6 GDPR, mn. 52.

which as a rule allows the data processing if the interests of the data subject do not prevail the legitimate interests of the data controller. This rather wide leeway in favor of data processing also speaks against a rather narrow interpretation of Article 6 (1) (f) GDPR. The weighing has to be done from an objective point of view, as recital 47 references by specifying that reasonable expectations of data subjects should be taken into account. That also means that the actual expectations of the individual data subject do not matter in this context, but are rather considered if the data subject makes use of his right to object, Article 21 GDPR. At this point it is not significant, if the data controller is compliant with the obligations imposed on him within the GDPR, because he cannot derive positive effects just from complying with obligations he is bound to anyway. It is in contrast to consider, to which extent and what type of personal data is being processed and how data is used.[13] A data processing to a rather small extent, e.g. to derive generic information from personal data, after which the personal data is being discarded will affect the interests of the data subject considerably less than a comprehensive processing of data which afterwards is kept.

In summary, the weighing of interests depends on the specific purpose and process of data and text mining with the consideration of possible risks during the data processing. The controller will not be able to first just collect the data and later specify a purpose for the (further) processing of data. The implementation level of data protection by design and default according to Article 25 GDPR is also to be taken into account.

3.4 Processing of special categories of personal data, Article 9 GDPR

If the data controller is processing special categories of personal data as specified in Article 9 (1) GDPR, such as personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and 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, the data controller has to additionally fulfil the requirements of Article 9 (2) (e) GDPR, which requires the data to be manifestly made public by the data subject. The first group of special categories of personal data within Article 9 (1) GDPR not only applies to personal data containing these information directly, but furthermore cases where these information could be derived ("personal data revealing …") from "regular" personal data, such as a name revealing information about the data subject's ethnicity. If a publication can be classified as manifestly made public, should in this context be assessed in the perspective of an objective observer.

Even with special categories of personal data, Article 6 (1) (f) GDPR is the legal basis for data processing and is just complemented by Article 9 (2) (e) GDPR, which in

[13] Gola, Article 6 GDPR, mn. 53 - 59; Kühling/Buchner, Article 6 GDPR, mn. 149 – 154.

contrary to some commentaries[14] is no legal basis on its own and therefore cannot be considered lex specialis[15].

For publications on social networks that means, that the observer must be left in the impression that the account is affiliated with the real person and the data connected to this person has been published manifestly. If in doubt, the requirements of Article 9 (2) (e) GDPR are not met.[16]

When using publicly available data, special categories of personal data are particularly problematic, as many information found when crawling the internet will contain special categories of personal data and the controller often is not able to verify, if the person responsible for the publication or distribution of the personal data also is the data subject and / or is permitted to publish or distribute this information. The first might e.g. be the case if a service validates the identity of users (such as the blue verified badge shown on certain accounts on Twitter), but in this case doesn't say anything about what information is shared on such an account and whether a possibly different data subject of this shared data is connected to the account holder or has permitted such a data usage.

In general, the processing of special categories of personal data in connection with crawling should be avoided, as the indicated hurdles are hard to overcome. A possible solution could be to limit the crawler to websites containing knowingly no special categories of personal data. Another solution could be the limitation of processing if special categories of personal data are found during the process of crawling. Even though there already must be a legal basis for the crawling itself, one could argue that in context with the freedom of information the controller should be able to assort unwanted information (personal information in general or special categories of personal data), at least if this is done directly during the process of crawling, the data is only stored volatile and is immediately deleted upon identification. As another approach it could be assumed that for the purpose of legitimate data processing, any found special categories of personal data is initially believed to be published lawful. But as this approach de facto imposes an obligation on the data subject to constantly monitor the internet for published personal data, this would be a considerable disregard for the data subject's rights.

4. Data subject's rights & transparency requirements

Articles 13 und 14 GDPR impose significant obligations regarding transparency and information about the data processing on to the data controller. As the data within a data and text mining context regularly is not obtained directly from the data subject, Article 14 DGPR is applicable.

[14] Gola, Article 9 GDPR, mn. 1.
[15] Kühling/Buchner, Article 9 GDPR, mn. 4.
[16] Kühling/Buchner, Article 6 GDPR, mn. 80.

Article 14 (1) GDPR states, the data controller has to provide the information stated in Article 14 (2) GDPR not later than specified in Article 14 (3) GDPR to the data subject, except if one of the exemptions in Article 14 GDPR is applicable. Information the data controller must provide to the data subject include the identity of the data controller (Article (1) (a) GDPR), the contact details of the data protection officer (Article 14 (1) (b) GDPR), the purposes of the processing and the legal basis for processing (Article 14 (1) (c) GDPR), the categories of personal data concerned (Article 14 (1) (d) GDPR), the recipients or categories of recipients of the personal data (Article 14 (1) (e) GDPR), the period for which the personal data will be stored or criteria to determine that period (Article 14 (2) (a) GDPR), the legal interests pursued by the data controller with the processing (Article 14 (2) (b) GDPR) and the source the personal data originate, and if applicable, whether it came from publicly accessible sources (Article 14 (2) (f) GDPR).

At this point, there are two issues regarding the data controller's transparency obligations. First the compliance could possibly result in the processing of additional and otherwise discarded information, just for the sole purpose of compliance. Alongside with the difficulties of the actual implementation notifying the data subject about the data processing and the disclosure of the required information about the data controller.

Article 14 (5) (b) GDPR provides an exception for Article 14 (1) – (4) GDPR and therefore could exempt the data controller from the information obligations mentioned above. The exception is granted under the condition, that provision of such information proves impossible or would involve a disproportionate effort, in particular in connection with processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes. The provision of information is impossible, if the data controller doesn't know the data subject, which complies with Article 11 (1) GDPR that states the data controller does not have to process personal data for the sole reason to comply with the GDPR. The second case states that the data controller can abstain from informing the data subject, if the provision of this information would result in a disproportionate effort. This exception requires a weighing of interests. The more of the data subject's interests are affected by the processing, the longer a higher effort of the data controller is still considered proportionate. Specifically mentioned is the processing for archiving purposes in public interest, scientific or historical research purposes or statistical purposes, which in this context are considered privileged purposes of data processing. In case the exception is applicable the last sentence of Article 14 (5) GDPR orders the data controller to take appropriate measures to protect the data subject's rights and freedoms and legitimate interests and specifically designated the public release of the above-mentioned information as an example.

A possible solution to the described issues of transparency could be the design of an open standard for the specification and distribution of content re-use rights. This

standard should be designed in a way like robots.txt or Privicons[17], so it could be used for content that is published through social media platforms as well as content which is directly published by the author through his website. An open and machine readable format shall be used for the exchange of this metadata. And the party which is crawling the information should be able to trigger a pingback mechanism, which notifies the author of the use of his content. For use in a big data and crawling context it should be machine readable as well as sufficiently adjustable.

An actual implementation of this open standard could be the development of an open data protection or "reuse" interface which could be implemented by the provider of the service or the website. This would allow crawlers to query the interface in order to retrieve the legal settings for a specific piece of data. The current possibility in the form of robots.txt might constitute a solution for smaller or private websites, but fails to provide the needed degree of customisation (for possibly thousands of users) for the specific purpose of crawling content. Apart from that robots.txt was primarily designed to be used in a search engine specific context[18] and the development did not consider the technical means and in 1994 unknown applications that are possible nowadays.

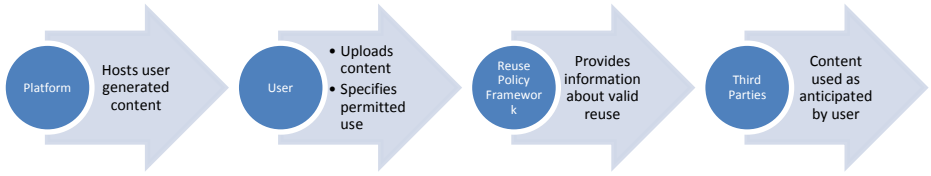

Figure 1. Reuse Policy Framework

The proposed standard consists of a dynamic backend or static file. For smaller or private websites this might be a single "reuse.txt" which contains the static reuse instructions in machine readable form. For bigger services with thousands of users and

[17] L.-E. Holtz, H. Zwingelberg, M. Hansen, Privacy Police Icons, Privacy and Identity Management for Life, p. 284.

[18] Wikipedia: Robots exclusion standard, online: https://en.wikipedia.org/wiki/Robots_exclusion_standard

possibly millions of different data sets this might be a dynamic backend such as a REST API or can be incorporated into an existing API. Using this system the user has the possibility to define specific reuse policies and apply these policies to existing content. Not only is it possible to specify policies for different types of content or a single item, but the user is also able the individually specify the legitimate kind or purpose of usage. This enables the user to unlock specific content only for scientific purposes and prohibit any other reuse. Dependent upon the actual implementation the system might require the operator of the crawler to specify the type of usage and on this basis only to return content specifically intended for the crawler's type of use or the system returns all reusable data with the corresponding reuse policy flags.

Another possibility is the use of pingbacks, to inform the user about the reuse of his content. Like the XML-RPC request being sent from one weblog that links to another weblog, a pingback could be sent to the user upon retrieval of data. This system can either be implemented using the outlined API or using a third party service. This would allow the data controller to inform the data subject about the data processing and send him the required information or a link to a website where the data controller has stored the required information in accordance with Article 14 GDPR.

If the data controller stores the information according to Article 14 GDPR on a dedicated website, the data subject should also have a possibility to contact the data controller and exercise his rights as a data subject.

Another option would be to use a third party to provision these services. The user would sign up for the corresponding service and could implement a similar API service into his website as illustrated above. This provider could also provide services for the data and text miner and therefore be considered a neutral third party. A major drawback of this option is, that a third party would receive personal data and information about the data subject's use of services and the service provider hypothetically would be able to connect identities the data subject uses across different types of services, i.e. if the data subject has two different private weblogs where he publishes content under different aliases and uses the API services under one account for both weblogs.

As an initial draft the "reuse policy framework" therefore might be designed in a way that allows the user to place a small robots.txt-like file on his webserver or a service provider to implement a corresponding policy flag within his API. The framework should allow users to specify different policy classes (i.e. "public_posts", "private_posts", "science_posts") for use with different resources. Resources contain information about the policies for a specific content item (i.e. a static site, an item on a weblog or a posting on a social network) or a group of items (i.e. a folder containing different static sites, multiple items on a weblog or multiple postings on a social network) which may be assigned policies or specific reuse settings. Reuse settings for use within policies or resources should contain information on whether a resource may be used (all/none/single = policies are set for every single item), for which pur-

pose (all/nc = non-commercial/sci = scientific/priv = private), how long a item may be used for the specified purpose, how long a crawler should wait until it attempts to access the next resource and how many crawlers are allowed to access the page simultaneously.

Example of static reuse.txt:

```
{public_posts}
allowed        all
policy         none
expire         7d
wait           5s
limit          1

{private_posts}
allowed        none

{science_posts}
allowed        all
policy         sci
wait           1s
limit          5

[/*]
ruleset        private_posts

[/blog/*.html]
(crawler: bad_crawler)
allowed        none

ruleset        public_posts
```

An implementation within a REST API could result in the following return values for a single item accessed (equal to the above “public_post”):

```
"reuse" :{
       "allowed":"all",
       "policy":"none",
       "expire":"7d",
       ...
}
```

5. Copyright

When crawling and analysing publicly available information there is a chance that work protected by copyright is also being processed. Comparable to data protection law there must be a legal basis for any technical processing that is not considered plain consumption of the copyrighted material. During the technical process of crawling and mining the information, the digital file is duplicated at least once, as the file is accessed on the remote webserver and transmitted to the machine running the crawler (and the data and text mining software). This transmission (and any further duplication, depending on the specific technical implementation) could infringe the copyright of the copyright holder if the person accountable is not able to invoke a specific legal basis.

As described in the introduction copyright law is generally not harmonised within the European Union which leads to a reasonable amount of uncertainty regarding the use of content that is protected by copyright law. In Germany, the one applicable legal basis for data processing (and duplication of a copyright protected work) in this context is Para. 44a Urhebergesetz (UrhG) which allows a temporarily volatile duplication which must be an integral part of a technical process as far as the pursued purpose is lawful use and has no own commercial value. This exception is the implementation of Art. 5 of the Directive 2001/29/EC[19] which means that this regulation should be implemented in national copyright laws throughout the European Union. The question is, if data and text mining really is only a volatile part of a technical process that has no own commercial value. Considering that the results of data and text mining are in fact commercially utilisable this might constitute a problem. This might also be questionable in cases where the controller offers a commercial service of data and text mining to third parties. But since most cases of data processing within a data and text mining context not just happen "on the fly", but also require the data to be stored, this, if the information is protected by copyright law, must be deemed a duplication which would require a (different) legal basis that is not currently available.

In a scientific context, the European Commission is currently working applicable exception in the form of a copyright exception for data and text mining of research organisations acting in the public interest[20].

6. Conclusion

In summary, the controller looking to process publicly available data for data and text mining will face major challenges within the scope of GDPR. The data controller must observe the general principles of the GDPR in Article 5 GDPR as well as data protection by design and default, as described in Article 25 GDPR. It is still unknown

[19] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, online: http://eur-lex.europa.eu/legal-content/DE/ALL/?uri=CELEX:32001L0029

[20] Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market.

how specific implementation of the GDPR within member states affect data and text mining, but especially relating to the raised issues of special categories of personal data, the possibilities enabled by opening clauses are heavily limited. But especially considering the possibility that publicly available data contain special categories of personal data the controller must ensure that a legal basis is applicable, sources for the data used are narrowed towards "safe" sources or special categories of personal data are not being processed. In context with copyright law there currently is no practical experience whether the current exceptions are sufficient to justify a use of copyright protected material for data and text mining. Furthermore it also depends on the specific data processing implementation. At least for scientific use the European Commission is currently working on a specific exception. The implementation of a standard for communicating allowed use would not only provide legal certainty but also benefit the data subject as it could provide a tool to efficiently exercise ones right to data protection.

References

1. Kühling, J./Buchner, B., DS-GVO, Munich 2017.
2. Ehmann, E./Selmayr, M., Datenschutz-Grundverordnung, Munich 2017.
3. Gola, P., DS-GVO, München 2017.
4. Paal, B./Pauly, D. A., Datenschutz-Grundverordnung, Munich 2017.
5. Schaffland, H.-J./Wiltfang, N., DS-GVO, Berlin 2017.
6. Camenisch, Jan/Fischer-Hübner, Simone/Rannenberg, Kai, Privacy and Identity Management for Life, Berlin 2011.
7. Zarsky, Tal Z, Incompatible: The GDPR in the Age of Big Data, Seton Hall Law Review: Vol. 47 : Iss. 4 , Article 2, online: http://scholarship.shu.edu/shlr/vol47/iss4/2.