Project no. 732027

VIRT-EU

Values and ethics in Innovation for Responsible Technology in EUrope

Horizon 2020

ICT-35-2016
Enabling responsible ICT-related research and innovation
Start date: 1 January 2017 – Duration: 36 months

D6.4
Prototype for ethical data research practices

Due date: 30 November 2019

Actual submission date: 1 December 2019

Number of pages: 39

Lead beneficiary: UU

Authors: Matteo Magnani, Davide Vega, Fredrik Jonasson, Luca Rossi, Irina Shklovski

Project Consortium

Beneficiary no.	Beneficiary name	Short name
1 (Coordinator)	IT University of Copenhagen	ITU
2	London School of Economics	LSE
3	Uppsala Universitet	UU
4	Politecnico Di Torino	POLITO
5	Copenhagen Institute of Interaction Design	CIID
6	Open Rights Group	ORG

Dissemination Level

PU	Public	X
CO	Confidential, only for members of the consortium (including the Commission Services)	
EU-RES	Classified Information: RESTREINT UE (Commission Decision 2005/444/EC)	
EU-CON	Classified Information: CONFIDENTIEL UE (Commission Decision 2005/444/EC)	
EU-SEC	Classified Information: SECRET UE (Commission Decision 2005/444/EC)	

Dissemination Type

R	Document, report	X
DEM	Demonstrator, pilot, prototype	X
DEC	Websites, patent filling, videos, etc.	
O	Other	
ETHICS	Ethics requirement	

Contents

Executive summary

The objective of this deliverable is to present *an extension of a Twitter data collection tool to support participant notification and other compliance measures.* This task was the responsibility of the quantitative unit of the VirtEU project, led by Matteo Magnani at Uppsala University. This report contains:

- an analysis of the implications of the General Data Protection Regulation (GDPR) on online social network research;
- the identification of three possible extensions of the functionality of typical Twitter data collection tools to simplify compliance with the GDPR;
- an experiment to test the limits imposed by the Twitter platform with respect to the aforementioned extensions, including the resulting decisions about possible implementations;
- a list of system requirements and design considerations to add the new functionality to an existing data collection tool (DMI-Tcat);
- an overview of the new functionality of the tool.

The code of the extended Twitter data collection tool is available at:
https://bitbucket.org/uuinfolab/dmi-tcat-plugin

1 Introduction

1.1 Context and motivation

The main role of the quantitative unit of the VirtEU project was to develop tools (software and methods) for the analysis of online social network data to be used by other project members. In addition, the quantitative unit was tasked to perform some of the analyses. The results of these activities have been described in previous deliverables, in particular Deliverable 2.2 Section 2, where the software platform developed to collect and explore the data is described, and Deliverable 3.1.

A few months before the intermediate project review, on the 25th of May 2018, the General Data Protection Regulation[1] (GDPR) came into force. Therefore, the reviewers suggested that the quantitative unit should reallocate the resources remaining after the completion of the research described in Deliverable 3.1 to investigate the implications of the GDPR on the kind of online social network research performed in the project. This would include the design and implementation of an extended data collection tool providing new functionality to simplify compliance with the GDPR, and the present deliverable that includes a description of the tool (including the research performed to design it) in addition to its downloadable source code.

When the authors of this deliverable started looking at the scientific literature to identify typical measures to be implemented in social network data collection tools to simplify compliance with the GDPR, it was soon clear that some original research was necessary. Several papers had already been published to examine the impact of the GDPR on research in general [12, 21, 24] and on specific research fields [23]. However, none had addressed the implications of the GDPR for online social network analysis.

Therefore, this activity did not only consist of the design and implementation of a new extended data collection tool, but also required theoretical reflections and practical experiments to identify and prioritize appropriate extensions.

This report, its related research papers and the software have not been used during the collection of the project data, that started at the beginning of the project and was performed using the tool described in Deliverable 2.2.

1.2 Structure of this report

The remainder of the report is organized as follows. In Section 2 (The GDPR and online social network research) we start with an analysis of the implications of the GDPR on online social network research. This section provides a broad set of reflections that are not constrained by considerations on software implementation. This broad analysis led to the identification of possible extensions of the functionality of typical Twitter data collection tools to simplify compliance

[1] 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.

with the GDPR. These are listed in Section 3 (Selected features for the extended tool). Section 4 (Experimental feasibility analysis) describes an experiment to test the limits imposed by the Twitter platform with respect to the aforementioned extensions. In fact, a Twitter data collection tool is only a part of a complex system, where both the social media platform and the researchers who are supposed to use the tool may introduce limitations that make some possible implementations unfeasible or unlikely to be adopted. This section also mentions the choices made about what features to implement based on the results of the experiment.

The rest of the report focuses on the actual extended tool we developed. Section 5 (Software requirements and design) provides system requirements and design considerations to add the new functionality to DMI-Tcat, which is a popular existing data collection tool. One of the reasons to choose DMI-Tcat was that it has an active community of developers, who we could approach to receive feedback about our proposal. Section 6 (Overview of the tool) shows an overview of the new functionality provided by the tool. The code is available through Git at: https://bitbucket.org/uuinfolab/dmi-tcat-plugin

1.3 Related publications and contributors

Part of the material contained in this deliverable is included in recent and ongoing research papers. The discussion presented in Section 2 constitutes part of a larger paper on the GDPR and social network research we prepared after the intermediate project evaluation, authored by Andreas Kotsios, Matteo Magnani, Luca Rossi, Irina Shklovski, Davide Vega [18]. This work, currently accepted for publication on the ACM transactions on social computing, also contains a section about the need to develop new social network software providing GDPR-related functionality. The experiment we performed to test the feasibility of our planned extensions has become part of a paper on ethical aspects of Twitter research, authored by Irina Shklovski, Luca Rossi, Matteo Magnani and Davide Vega and currently under submission. The design of the tool has been described in [16], authored by Fredrik Jonasson under the supervision of Matteo Magnani and Davide Vega; part of the content of the deliverable comes from this work. The overview of the extended tool is currently only presented in this deliverable, and has been produced by Matteo Magnani and Davide Vega. We will consider including part of this in the tool documentation in the future. The unit coordinator, Matteo Magnani, has written the remaining parts and integrated material from the papers mentioned above. Davide Vega and Fredrik Jonasson have written the code of the demonstrator.

2 The GDPR and online social network research

The GDPR introduces seven general principles to be followed when processing personal data[2][3]. In this section we discuss the meaning of these principles when they regulate the processing of online social network data, emphasizing the cases where ambiguities arise. A summary of the main GDPR-related aspects that should be considered during a social network analysis process, including a list of exemptions that can be applied in research, is presented in Tables 1 and 2.

2.1 Lawful bases for data processing

The first basic principle of GDPR states that the data must be processed in a lawful, fair and transparent way. In the case where a controller is a university, as in the VirtEU project, it may be most suitable to use as a lawful basis that the processing is necessary "for the performance of a task carried out in the public interest"[4]. The definition of such tasks is left to Union or Member States law[5]. There is, however, no need for an explicit statutory provision as long as there is a clear basis in law[6]. Even in cases where no national legislation is introduced with regards to it, it should be accepted that pubic actors, such as universities, may use this lawful basis for processing of personal data[7]. Since in many countries universities – often even private ones – are considered to be public authorities by law and they act on carrying out tasks of public interest, such as conducting research[8], the public task basis for processing personal data seems to be the appropriate lawful basis for a social network research project, as long as the processing is necessary for that project[9]. This lawful basis puts the onus of ensuring that the rights of the data subject are balanced against the public interest goals of institutions, whose aims presumably are oriented towards

[2] Art 5 GDPR

[3] (P1) The data must be processed in a lawful, fair and transparent way. (P2) Personal data may only be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes. (P3) The data may be processed only if they are adequate, relevant and limited to what is necessary with regards to the purpose of processing. (P4) Only data that are accurate and up to date, to the level that it is possible, may be processed. (P5) Personal data may only be processed for a period that is necessary for the processing and therefore the controllers must create criteria to determine what retention periods are suitable for their purposes. (P6) The controllers must apply technical and organisational measures in order to protect personal data they control against unauthorised and unlawful processing as well as accidental loss, destruction or damage. (P7) The data controllers have the responsibility to be compliant and to be able to demonstrate compliance when needed, which implies that written records must be kept on whether and how the controller is compliant.

[4] Art 6.1(e) GDPR. See also SOU 2017:50

[5] Art 6.3 GDPR

[6] Rec 41 GDPR

[7] SOU 2017:50, p.18

[8] See for example in the UK the Freedom of information Act 2000 and in Sweden the Higher Education Act 1992:1434

[9] According to art 6.2 and 6.3 GDPR as well as rec. 45 GDPR it is stated that Union or Member State law shall define whether the controller performing a task of public interest can be a legal person governed by public law or by private law.

	General rule	Exem.	Details
1	Identify the roles w.r.t. the GDPR ecosystem (data subjects, controllers, processors, DPO. . .) and the data flows.	no	This can be challenging in some cases, consult the DPO if uncertain.
2	Identify the nature of the data (personal / non personal / sensitive).	no	In case of sensitive data, we can process it (1) if we have explicit consent, (2) if the data was manifestly made public by the data subject (use this carefully), or (3) in case of research purposes, if there are suitable safeguards (e.g., pseudonymization, approval from ethics committee).
3	Identify explicit and legitimate purposes for the processing.	yes	The specification in case of research can be a bit more general (such as the general research area or part of the project, not specific analytical tasks). Some specification of the intended purpose is however necessary.
4	Identify the lawful basis for data processing.	no	Based on national legislation, that is still being produced, some actors conducting research (e.g. universities) might be assumed to operate in the public interest and therefore the public task basis may primarily be used. Otherwise the consent and legitimate interests bases should be examined.
5	Define clear temporal limits for data processing. Non-anonymized data can be kept for no longer than is necessary for the purposes of the processing.	yes	More extended periods may apply in case of research as long as appropriate safeguards are implemented.
6	Put in place technical and organizational measures to protect the data, e.g., ensure privacy by design and by default, pseudonymize the data as soon as possible.	no	The measures should be proportionate to the aim pursued.
7	In case of profiling perform a DPIA.	no	Consider with the DPO whether a DPIA is necessary.

Table 1: A summary of general rules and exemptions to be considered during the social network analysis process. Column Exem. indicates whether explicit exemptions exist for research, and exemptions (if any) and other considerations are indicated under Details. Abbreviations used in the table: Data Protection Officer (DPO), Data Protection Impact Assessment (DPIA). (Part 1, from [18])

	General rule	Exem.	Details
8	Inform the data subjects about the collection, purposes and their rights at the time the data is obtained (if obtained directly from the data subject) or within a reasonable period after the data is obtained and no later than a month (if the data is obtained indirectly).	yes	For secondary data, providing information is not necessary if the provision of such information proves impossible or would involve a disproportionate effort, if this is likely to render impossible or seriously impair the achievement of the objectives of the processing.
9	Collect only adequate, relevant and limited data to what is necessary to achieve the purposes of the processing.	yes	As the purpose may be specified in less precise terms (see the exception to Rule 3), this rule is also affected. Consider deleting unwanted data as soon as possible, acknowledging and documenting the process.
10	Data subjects have the right to check if there is data concerning them, and the right to obtain these data.	no	Even if not part of the GDPR, national laws may still restrict this right, e.g., secrecy acts.
11	Data subjects have the right to have the data concerning them erased.	yes	Not necessary if it is likely to render impossible or seriously impair the achievements of the objectives of the processing. National laws may also restrict this right.
12	Keep data accurate and up to date.	no	
13	If a new purpose emerges, new legal bases for data processing should be identified.	yes	If the new purpose is research, further processing is considered to be compatible to the initial purpose.
14	If the controller changes the purpose of the processing, information must be provided to the data subject prior to this processing.	yes	See the exception to Rule 3 about the increased flexibility in the specification of the purpose in case of research.
15	Keep written records to demonstrate compliance.	no	

Table 2: A summary of general rules and exemptions to be considered during the social network analysis process. Column Exem. indicates whether explicit exemptions exist for research, and exemptions (if any) and other considerations are indicated under Details. (Part 2, from [18])

the greater good. This basis is not available at all to commercial organizations and research labs – at least as long as no law provides for that – who must rely on consent or the legitimate interest basis to process personal data.

With regards to the use of consent[10] as a lawful basis for the processing of data in research, there are some things that have to be taken into consideration. The first one is that even though this lawful basis can also be used for the processing of personal data by a research project, an entity may use this lawful basis only "if a data subject is offered control and is offered a genuine choice with regard to accepting or declining the terms offered or declining them without detriment."[11] If this is not possible, something that in online social network research can be the case, then this lawful basis should not be used [12]. Additionally, for public universities since they are public authorities, researchers must always assess whether or not the consent provided by the data subjects is valid, namely if it is indeed freely given or it is given as a product of imbalance in powers between the university and the data subjects[13]. Lastly, one should make a distinction regarding the term consent as developed in the GDPR and as an "ethical standard and procedural obligation"[14]. That means that it can be so that the lawful basis for processing is the public task basis, art 6.1(e) GDPR, but consent is used as an additional safeguard. In this case it is not two lawful bases used for the processing of personal data but only one, the public task base; consent is only a procedural obligation and not the lawful basis provided for in art 6.1(a).

One last thing that we would like to add here is that if the personal data processed are of sensitive character, an entity conducting research — at least an entity, such as a university, that bases their research activities on some piece of legislation — may primarily base the lawful processing of such data on the fact that the processing is necessary for scientific research purposes as long as appropriate measures are deployed according to art 89.1 and the research is based on a law "which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject" according to art 9.2(j) GDPR[15]. Following the same argumentation as

[10] It is not the goal of this paper to make an analysis on consent as a lawful basis in general – for a better understanding we refer to the Article 29 Working Party Guidelines on consent under Regulation 2016/679 – but it is worth reminding here that a consent for processing of personal data by a data subject has to be freely given, specific, informed and unambiguous.

[11] Article 29 Working Party Guidelines on consent under Regulation 2016/679, p. 3

[12] Here it is also important to consider that, as we will argue later in this paper, it can be difficult to provide information to the data subjects of a network research project and therefore it can similarly be challenging to provide the possibility for an informed consent.

[13] In most research projects, this should not be a great issue since data subjects in a network research project do not normally have a direct connection to a university, but it is still worth considering possible problems that may arise.

[14] Art 29 Working Party Guidelines on consent under Regulation 2016/679, p 28

[15] Art 9.2(g), namely that the processing is necessary for reasons of "substantial public interest" could also be the basis for lawful processing of sensitive personal data but since art 9.2(j) specifically refers to scientific research purposes, processing that takes place for scientific purposes should be based on the legal ground of art 9.2(j)

above we could, however, claim that if the processing is not necessary or if there is still no specific legislation with regards to processing for research purposes, consent could also be used as a lawful ground for such processing, according to art 9.2(a) GDPR[16].

2.2 Secondary data collection and transparency

In the VirtEU project we collected data through a third actor, i.e., online social networks obtained from APIs without the direct involvement of the data subject. The difference with respect to more traditional ways of collecting social network data, e.g., using questionnaires, is not only in the scale or the nature of the data but in the relation between the data subject and the data controller: two different articles are concerned with providing information to the data subject when the data are collected directly from them[17] and when data about them have not been obtained from them[18].

In essence, these articles detail some of the ways in which the principle of transparency must be put into action. Transparency addresses the right of the data subject to know and understand how the data are being used; it "requires that any information addressed to the public or to the data subject be concise, easily accessible and easy to understand, and [in] clear and plain language [in particular] in situations where the proliferation of actors and the technological complexity of practice make it difficult for the data subject to know and understand whether, by whom and for what purpose personal data relating to him or her are being collected [...]." If personal data are collected, the data subjects should be informed about the collection and its purposes in order to enable them to exercise their rights. Note that this is different from consent but instead refers to the information that must be made available about data processing activities. Essentially, data subjects should be able to easily find out who might be using their data and for what purposes.

While making the data subjects aware of the processing and of their rights may seem straightforward when data are collected directly from them, this can become very difficult to accomplish when large networks are obtained from APIs. The potential difficulties to provide information under specific circumstances are acknowledged in the GDPR, where exceptions for research in particular are introduced. Article 14 states that providing information is not necessary if 1) "the data subject already has the information"; or 2) "the provision of such information proves impossible or would involve a disproportionate effort, in particular for [...] scientific or historical research purposes", subject to some safeguards[19], if providing information "is likely to render impossible or seriously impair the achievement of the objectives of that processing". Article 14 then continues

[16] Worth mentioning here that in many countries such processing by a university, even if consent is given by the data subject, could take place only after an ethics committee permits it. See also SOU 2017:50 s. 160.

[17] Art 13 GDPR

[18] Art 14 GDPR

[19] Art 89 GDPR

stating that ”[i]n such cases the controller shall take appropriate measures to protect the data subject’s rights and freedoms and legitimate interests, including making the information publicly available”.

These are some examples of the kinds of research exemptions that are embedded in the GDPR, codifying and specifying research conduct. Both those exemptions apply to social network research based on online data collected from social media platforms assuming that social media platforms have already informed their users through appropriate Terms of Services that their data will be shared with third parties (eg. through APIs) or assuming that the large scale of collected data will require a disproportionate effort to inform all affected data subjects. This is an example of balancing research needs against the derogation of the rights of the data subject. Technically termed ”proportionality of the effort”, this is a relatively vague concept. The controller, in order to determine whether it is going to be disproportionately difficult to provide the information, must take into consideration the number of data subjects, the age of the data and if there are any appropriate safeguards already adopted[20]. If, after this assessment, the controller finds that the effort will be disproportionate, then she has to assess once again whether the effort involved to provide the information to the data subject exceeds the impact and effects on the data subject in the case where the information is not provided. This assessment has to be documented and depending on the outcome the controller may have to take extra measures (such as pseudonymisation or anonymization if possible and appropriate).

As an example, this means that although the research exceptions may not technically require that every single Twitter user of the millions involved in any large-scale Twitter network research be notified that their data are used for research, the logic involved in deciding to collect data and to skip the notification must be formally documented. This documentation must also demonstrate that appropriately storage, security and pseudonymization techniques have been considered. In addition, it is unclear whether providing information to these users should be considered an impossible or very difficult task. In any case, the disproportionate effort it would require to provide information to the data subjects shall be demonstrated by the data controller, and is not something that should just be taken for granted. To address this uncertainty we set up an experiment aimed at quantifying the limits and efforts in notifying Twitter users about an ongoing data collection, described in Section 4.

The concept of transparency is particularly relevant in the context of social network research, as previously highlighted e.g. by Borgatti and Molina [5], and as such it requires a more extensive discussion. In particular, some additional details should provide a better description of the obligations of the data controller with regard to the provision of information. There are three points that are important here.

First, the data controller must always provide information within a reasonable period after the data is obtained and no later than a month (if the data is obtained indirectly) as long as this is possible given the appropriate adherence

[20] Rec 62

to the research exemptions detailed above.

Second, if the controller changes the purpose of the processing, she must provide the information to the data subject prior to this processing[21]. For example, research data may have been collected for one purpose but the research question has shifted in the course of the data analysis and these data will now be used for a different purpose. This then speaks to how precisely the information about processing must be specified. Looking at rec 33, even though referring to consent, we can conclude that the specification in case of research can be a bit more general (such as the general research area or part of the project, not specific analytical tasks). Therefore changing data analysis approaches and even research questions may not require informing the data subject anew.

Related to the above is the fact that if the change leads to further processing that is incompatible to the initial purposes, mere information of the change does not ”whitewash” other obligations of the controller. According to art 5.1(b) GDPR processing should comply to the purpose limitation principle. That means that as soon as the new processing is incompatible to the initial, the controller should either avoid the new processing or find a new lawful basis for it. There is, however, an exception with regards to research purposes, since in such case the further processing for such a purpose is considered to be compatible to the initial purpose.

Third, the general principle does not assume that the methods and the analysis are known in details at the moment of the data collection. However, the common practice in many areas of research where data is often collected with no specific hypothesis/evaluation framework becomes problematic because at least a limited explanation for the purposes of data processing is always necessary. The GDPR recognizes that it is not always possible to know from the beginning the entire scope of the research until the data is collected and used. Rec 33 (in case of consent) states that data subjects should be able to ”consent only to certain areas of research or parts of research projects to the extent allowed by the intended purpose”. Thus some specification of the intended purpose is necessary, limiting but not entirely eradicating exploratory forms of data collection.

2.3 Online social network data and data minimization

Where some network data can be collected directly in the form of network information, that is, nodes and edges, many network datasets are obtained through processing of other types of data. For example this is often the case in research based on social media such as Twitter. Network studies of Twitter can be based on the user-articulated following/followers structure, that can be considered direct network information. At the same time, we can build networks mapping communication processes, either explicit (replies, mentions)

[21] Rec 61. See also Opinion where it is stated that in case the change is related to an incompatible further processing informing about the change does not ”whitewash” other obligations of the controller, such as finding another lawful basis for the changed processing or asking for new consent

or implicitly specified for example by the usage of common hashtags [15]. To build this second type of network, researchers collect the content of users' posts and then extract and infer relational information. The problem arises if we consider the implications of collecting the content of the posts to build the network. Depending on the topic of posts, the type of content that is likely collected may vary but could include data revealing information that is not only identifying of natural persons but also includes sensitive data such as political affiliation, religious belief, etc.

The GDPR makes a distinction between different types of personal data, such as data with regards to ethnicity and sexual preferences (the so-called sensitive personal data[22]), and in order for the processing to be considered lawful the controller must respect the essence of data protection rights and follow suitable safeguards[23]. Notice that data which in combination with other data can lead to revealing sensitive data may also be considered as sensitive data. For example name in combination with phone number, where each piece of data is not sensitive, may constitute sensitive data together if they probably reveal the ethnicity of a person. It is easy to see how the average stream of messages written by an average user might easily contain sensitive personal data or data that can be combined to reveal sensitive personal data about the data subject. Further, such data can be derived about persons simply from information produced by their connections. For example, it may be possible to ascertain a person's political affiliation if the majority of his connections explicitly communicate theirs.

Handling sensitive data is not forbidden, but before starting the data collection researchers need to plan some safeguards. Under the GDPR, controllers may not process sensitive personal data except if the subject has provided her "explicit consent"[24] or the data "was manifestly made public by the data subject"[25], or in case of research purposes[26]. While one may consider using the concept of "manifestly made public" for special cases such as online social networks, where the information is publicly posted online by the users, we advise against this interpretation. In fact, in the context of social media, as a consolidated body of literature has made clear, assuming when something is "manifestly public" is problematic [8] and a potentially serious breach of standard ethical research practices. On the contrary, the exemption in case of research purposes can be used, but only if processing is necessary, in accordance to Article 89(1), based on Union or Member State law which shall be "proportionate to the aim pursued, respect the essence of the data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject." Moreover, it seems that profiling on the basis of personal data is

[22] In the context of GDPR sensitive personal data is defined as "Personal data which are, by their nature, particularly sensitive in relation to fundamental rights and freedoms merit specific protection as the context of their processing could create significant risks to the fundamental rights and freedoms."

[23] Art 9 GDPR

[24] Art 9(2)(a) GDPR

[25] Art 9(2)(e) GDPR

[26] Art 9(2)(j) GDPR

forbidden unless there are ”suitable safeguards”[27]. For example, in Sweden, it was recommended that one such security measure can be considered the decisions of the relevant ethics committee[28].

Finally, even if the data are not sensitive, the data minimization principle should still apply. Using again Twitter data as an example, when researchers collect information based on a hashtag they can fetch data using the hashtag with another meaning, and so not related to the study, or data using the hashtag as was intended, but still including additional unwanted information. This means that researchers must put in place mechanisms that will effectively strip out unwanted data and delete it as soon as possible, acknowledging and documenting the process.

2.4 Data analysis and profiling

Social network analysis includes a wide range of data analysis tasks. Sometimes whole-network statistics are important, for example to correlate the communication/interaction structure of a team or organization to its performance. Sometimes meso-level structures are of interest, for example if we want to identify communities [14, 13, 7] or other relevant sub-structures such as online conversations [19, 29] inside a larger network. The identified groups can then also be used to classify individual actors, for example assigning them to a given community or role. Other types of micro-level analysis involve the characterization of single actors, for example when the most central or prestigious actors are identified [30]. When individuals are the object of the analysis, which is the case for most of the tasks listed above, an important concept to be considered is profiling.

The GDPR puts a special emphasis on the concept of profiling by specifying the definition and codifying acceptable practices. Accordingly, in the GDPR profiling is composed of three main stages ”a) collection of personal data; b) automated analysis to identify correlations; c) applying the correlation [the result of b)] to an individual to identify characteristics of present or future behaviour”[29].

Note that the notion of ”automated analysis” is used in the GDPR in opposition to ”manual”. Although both types of processing are under the purview of the GDPR, profiling is necessarily automated. However, automated here would mean both the use of a statistical software for conducting any form of data analysis as well as the use of more complex approaches such as machine learning algorithms. Thus any data analysis that includes computational assistance from software falls under automated analysis and thus can be classified as forms of profiling.

Given the above, many (but not all) social network analysis tasks can be classified as profiling. All centrality measures are clear examples, as they associate

[27] Rec 51 GDPR

[28] SOU 2017:50

[29] Art 29 Data Protection Working Party, WP251rev.01, ”Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679”

results of the network analysis to specific individuals. Any analysis that singles out individuals based on the identification of positions, roles and communities is similarly a form of profiling.

What is the researcher to do if their activities constitute profiling of the data subject? This does not mean that the particular data analysis is disallowed. However, this may require the performance of a data protection impact assessment (DPIA), for which the advice of the appointed data protection officer should be sought. Although the GDPR states that profiling has to be systematic and extensive to require a DPIA, many authorities have made a broader implementation and if profiling may affect individuals in general (e.g. it provides custom access to services, it includes sensitive data, is related to vulnerable individuals, and in general the processing can lead to a high risk to the rights and freedoms of the data subject) and if it is conducted in a large scale combining sensitive data, then a DPIA is in general necessary. The question of whether a DPIA is necessary is clearly a very important one, because a very strict approach leading to an assessment for every possible case of social network analysis can become practically problematic for the researchers. While we wait for more guidelines[30] and other legal specifications, the role of the researchers together with the DPOs deciding on whether an assessment is needed or not (following the law but also being practical) is of even higher importance.

Alongside profiling, DPIAs are also applicable to systematic monitoring of individuals and locations. An interesting question arises with respect to what constitutes locations and public spaces. For example, the GDPR mentions a ”systematic monitoring of a publicly accessible area on a large scale” as a reason for a DPIA[31]. We are not aware of existing legal interpretations of whether e.g. Twitter is a publicly accessible area, but the WP29 interprets ”publicly accessible area” as being any place open to any member of the public, for example a piazza, a shopping centre, a street or a public library. Clearly these are examples of physical places but Twitter is also a place that is open to any member of the public provided they have the means to access it (an Internet connection and access to an email address). Such questions will likely be decided later on as the regulation stands the test of time and litigation, but it is an important item to consider for researchers conducting large-scale collection and processing of ostensibly ”public” data.

2.5 Data storage and storage limitation

We now discuss what happens after the research is concluded, in case the researchers want to store the collected networks. If the data are still personal, e.g., they still contain identifiers or have been pseudonymized, then the data controller must guarantee some rights to the data subjects if she wants to keep the network data. On a general level we can organize these rights along three lines: a) temporal duration of personal data storage, b) the accessibility of the

[30] https://www.ucl.ac.uk/legal-services/research/data-protection-impact-assessment

[31] Art 35(3)(c) GDPR

stored data to the data subject, c) the right of the data subject to withdraw his/her data. All these tasks are in general strictly regulated by the GDPR, but with significant exemptions for research, discussed in the following. Under the assumption that the networks have been anonymized, then there is no problem because the GDPR no longer applies: the data are no longer personal. However, network anonymization is a complex issue, that we also discuss below.

When it comes to temporal storage limitation, the GDPR states that in general data can be ”kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed”, but more extended periods may apply in case of research as long as appropriate safeguards are followed[32].

No exemption because of research is instead mentioned regarding the data subjects’ right to check if there is data concerning them, and the right to obtain these data[33]. This means that when requested the controller should provide the data, in a ”commonly used and machine-readable format”[34] (even if there are possibly other national laws that may restrict this right of a data subject, such as for example secrecy acts[35]). Considering the average amount of data represented by a single node in a typical social network project, this should not be a problem. Nevertheless, as for other parts of this section, the size of the network may constitute a practical difference, and for large networks researchers should probably consider implementing an automated data filtering functionality.

Finally, the right to erasure, also known as right to be forgotten, grants to the data subject ”the right to obtain from the controller the erasure of personal data concerning him or her without undue delay”[36]. However, also in this case the GDPR contains an exemption to this obligation if the erasure ”is likely to render impossible or seriously impair the achievements of the objectives of that processing”[37]. Many SNA measures are not so sensitive to a small amount of missing data [17] and the discipline has developed a set of techniques to handle missing data. Nevertheless, it should be acknowledged that a significant number of subjects requesting their data to be removed might seriously impair the research objectives, thus researchers would have the right to legally object to the data removal.

While these are the general guidelines emerging from the GDPR, according to art 89(2) Member States may further limit the data subjects right to access, rectification, restriction and to object in case of research if there are appropriate safeguards in place, and as long as the derogation is necessary for the fulfillment of the research.

[32] Art 5(e) GDPR

[33] Art 15 GDPR

[34] Art 20 GDPR

[35] SOU p. 223

[36] Art 17 GDPR

[37] Art 17 GDPR

2.6 Data anonymization

The GDPR asks for appropriate safeguards. The safeguards that are named in the GDPR are technical and organizational, e.g. data minimization, pseudonymization and anonymization. In addition, there can also be legal safeguards, such as contractual clauses between the controller and the processor, ethical vetting etc. [20]. Here we focus on anonymization, which should result in the data not being re-identifiable by the controller or any other person. In social network analysis, the typical approaches to anonymization are based on clustering, graph modification or network perturbation [31].

Data anonymization approaches in general are part of a considerable debate where some researchers argue that anonymization is impossible while others contend that it is in some cases [11, 1, 22]. Social network data is far more difficult to anonymize than other types of data and research on appropriate anonymization techniques is still in its relative infancy. Many of the simpler and more traditional approaches such as replacing node identifiers as well as more recent and complex approaches have been critiqued as insufficient [2, 1]. The knowledge of research being conducted in a particular location by a specific research group may be enough to reveal the identities of individuals encoded in the network to those who are familiar with these people more directly. In a small social network, such as for example a company division, it may be simple for the people in the network to recognize others based on just the revealed relational patterns [5]. Such an issue is not specific of social networks, but has been amply documented in qualitative and ethnographic research [25, 28]. As another case, if the data are public and indexed (e.g., by Web search engines), it can be very easy to find the original data using a part of it as a search key, such as finding the authors of a social media post based on the text of the post.

Whether anonymization or even just pseudonymisation are generally possible in a social network context is a difficult question. The GDPR states the necessity for privacy by design and by default but does not request specific privacy-preserving solutions: the controller should select and apply the appropriate measures for each case. In the GDPR, pseudonymisation requires the ”additional information” to be ”kept separately” and to be ”subject to technical and organisational measures”[38], which is not really possible when the data source is public: if one removes the user identifier but keeps the text of the post (e.g., the tweet), a simple search on a search engine or on the social media platform can easily lead to the original, complete information. In this case, a possibility to be considered by the researchers (but not explicitly required by the GDPR) is to transform the text so that the analysis can still be performed but it becomes more complicated to fetch it from the Web, such as replacing it with a bag of words. The relevance of this discussion is that according to rec 26 pseudonymized data is identifiable, so the GDPR applies to that, while anonymized data is not, so the GDPR is no longer relevant. However, given the difficulty in fully anonymizing the data we should often assume that the GDPR is still the relevant regulation.

[38] Art 4(5) GDPR

Even when we do not need identifiers to process social network data, because for example we are only interested in the structure of the network and its relationship with some indicators, we still need the identifiers if we want to extend the network, to know to which nodes the newly available information refers to. According to the GDPR we should at the very least pseudonymise the data ”as soon as possible” (rec 78). However, it is not unusual in online network studies to keep collecting data for months or even years, which means that ”as soon as possible” may be as late as the end of the study. One solution here is to develop or extend data collection systems with built-in network pseudonymisation functions, for example automatically removing identifiers and separately storing a mapping to user accounts in a location that requires special access credentials. Such solutions may seem overly onerous given the current accepted practices, but the GDPR forces us to rethink our attitudes towards data collection and the impacts of our practices more broadly. In addition, the idea of designing ethically-related features in social network analysis software has already appeared in the literature [5].

3 Selected features for the extended tool

For our extension we decided to start from the widely used Twitter Capture and Analysis Toolset system (Tcat), developed within the context of the Digital Methods Initiative running at the University of Amsterdam [6].

DMI-Tcat supports both capturing and analysis of Twitter data[39]. When it comes to capturing, the software uses the Twitter API and is thereby bound to the limitations of that API. We can only capture the data that Twitter is offering to their API users. The analysis part of the tool offers a wide range of options for visualizing and exporting data. These options can be used to get different insights during research.

The program is able to export to different common file formats as CSV and GEXF. GEXF, an acronym for Graph Exchange XML Format, is a common format for describing advanced networks and their structures. The format is supported by multiple software programs such as GEPHI.

Based on the general analysis of the impact of the GDPR on online social network analysis research, presented in the previous section, we prioritized the following specific features to be implement as an extension of DMI-Tcat: automated pseudonymization, automated user notification and user accessibility to personal data.

3.1 Automated pseudonymization

The person responsible for data collection (called administrator in the next sections) should have the choice, before data collection begins, to request the automated pseudonymization of the data. The tool should then for example replace screen name and user identifiers with random codes.

[39] https://github.com/digitalmethodsinitiative/DMI-Tcat/wiki [Accessed: 12-Jun-2019]

The tool should also store a mapping containing information to undo the pseudonymization, only available to the administrator but not to the data analysts.

3.2 Automated user notification

This is based on the GDPR principles of *fair and transparent processing*, that *"require that the data subject be informed of the existence of the processing operation and its purposes. The controller should provide the data subject with any further information necessary to ensure fair and transparent processing"*[40].

In practice, the tool shall be able to inform Twitter users that their tweets are being collected. It should also provide the users with information on how they can express their rights.

3.3 User accessibility to personal data

The system shall be able to provide data about specific Twitter users upon request. The tool shall be able to compile this data into a format that is portable, for example a CSV file or PDF.

4 Experimental feasibility analysis

In the previous section we have identified three specific enhancements to make it easier for online social network researchers to be compliant with the GDPR. However, in Section 2 we have also highlighted how the application of GDPR principles to online data is not straightforward. Therefore, before starting to build the extensions object of this deliverable we have tested the feasibility of some alternative design choices.

Not surprisingly, existing research tools to collect online Twitter data are modeled after accepted ethical practices and make it difficult to introduce different research practices. In particular, the tools act as black boxes and do not provide the possibility of customizing the data collection process without modifying their code, but only provide the collected data as output. There are two problems that come up here. First, there is no way to anonymize data output directly from the tools even if we wanted to. Second, there are no facilities here to provide notification of data collection to users.

Therefore, in this section we focus on two aspects. First, we tested the limits imposed by the Twitter API to the process of informing Twitter users about an ongoing data collection. Then, we briefly discuss our interactions with the community developing the tool we chose to be extended, which was deemed necessary to check if our extensions would be accepted by the community.

[40] rec 60

4.1 Limitations from the Twitter API

Twitter data collection can happen completely unnoticed by the users. No matter if one is collecting tweets trough an hashtag or lists of following/followers, the existing APIs do not notify the users when their data is shared. In the usual scenario, when tweets are collected on Twitter the only contact information we have are the Twitter identifier and screen name of the accounts whose tweets were collected. This means that should we wish to at least inform users about their data being part of a project, we can only do so via Twitter. It is typically not possible to send direct (private) messages to generic Twitter users (for example, users not following us who have not explicitly allowed this in their privacy settings). Thus the only possible way to inform them is through a public mention that would also be visible by others.

In this way, to inform data subjects in order to enable them to protect their privacy we have to release additional information about them: our public message implies that those accounts have posted tweets with the hashtag we were monitoring. This kind of necessity to make something public in order to enable privacy is how Carol Tavris and Susan Sadd [27] defined the privacy paradox in 1975, in contrast to contemporary definitions. That is, in order to activate the privacy guarantee of the research subject in question, we must first cancel it by revealing or exposing the fact of their participation to others [9].

With a public mention being our only option, then, we had to make two important decisions: (1) how many users to mention in the same tweet and (2) whether we should check what their current screen name is. Both decisions have an impact on the time needed to send the notifications.

First, including more accounts in the same Tweet would reduce notification time, but would also again release more information as each notified user would see the other user names in the same tweet. Knowing that they have also used the same hashtag. While none of this discloses ostensibly private information (any user could search for the hashtag and see who else posted on the topic), this action does highlight participation in the conversation to other people in a different way with potentially unwanted outcomes. Second, checking the current screen name would require more accesses to the Twitter API but would avoid that we mention the wrong account because screen names can change in time.

As an indication, the Twitter API currently allows us to send 2400 tweets per day, meaning that we would need around one year and two months to notify one million users (using a single notification account). This, under the assumption of using the Twitter screen-name, which is not a user identifier on Twitter and risks to be directed to the wrong account – the alternative being making additional API calls to fetch the current Twitter screen name of the accounts in the data.

Our decision was to minimize mentions' visibility by not including more than a single user per Tweet, and to maximize the API calls at our disposal by not checking for changes in the screen names:

> @*userscreenname*: some of your tweets have been collected by Uppsala University for academic research. More info and opt out: <link to info page>

Other decisions also made the practical information process less trivial than one may think. For example, should we also notify accounts mentioned in the collected tweets, even if they were not producing tweets themselves? Should we notify accounts retweeting other accounts' tweets? In our case, we were interested in the retweet network (so we planned to notify the former) but we would only build a mentioning network between those who had authored at least one tweet (so we planned not to notify the latter).

The effort to inform the users of their data being collected did not work satisfactorily. Our notification account was quickly stopped, after 22 tweets, by Twitter itself that interpreted such activity as an infringement of its terms of services. According to their interpretation the account had been marked as having a *spamming behaviour*.

This led to an interesting exchange with Twitter where we tried to defend our activity and stress the ethical reasons for informing the users. In the conversation with @TwitterSupport we mentioned that despite the behaviour being compatible with their definition of spamming, the account was an attempt to enforce the rights of the users to know that their tweets had been collected and why.

> Hi, we are processing public tweets in the context of a European research project and according to (some interpretation of) the GDPR we shall notify the users about this (art. 14). To do so, we are sending to the list of users whose tweets we have processed a tweet (more precisely, a @ mention) with a link to the project information page for data subjects, with instructions on how to receive the data, opt out, etc. However, our account has been temporarily locked — we do not know why, but we guess because of the frequent tweets with the same text (apart from the different user screen name). We only need to produce one single mention for each user involved in the study. How can we proceed? Is it possible to allow this behavior from this account? Regards,

A bot promptly replied:

> Thank you for taking the time to report this to us. We'll take a look and will follow up if we need additional info from you. Have you checked out our Help Center for troubleshooting tips? It's a great resource for instant answers to the most common questions: https://support.twitter.com.

The account was then unblocked, but without offering any exception, and blocked again after some more tweets (52 in total). The conversation with Twitter was then moved to email with Twitter Platform Operations who, in the last email, stated:

> Thanks for that additional information. We can confirm that sending unsolicited @mentions is prohibited by our Automation Rules

> and the Twitter Rules on spam. if you are sending automated @mentions the recipient or mentioned user(s) must have requested or have clearly indicated an intent on Twitter to be contacted by you. We can only consider a request to reactivate your app after you agree to stop this behavior.

4.2 Interactions with the community

While the planned strategy for notifying the users had to find a way through the limits imposed by Twitter's API (e.g. the impossibility of sending direct messages to users who are not following you), the pseudonymization of the collected data was more directly under our control. Dmi-Tcat's code is available on git-hub and released under a very permissive Apache 2.0 license. Moreover, the tool is largely adopted in digital media research and there is an active community of users and developers. For this reason the goal was not just to produce a local version of Tcat that could pseudonymize users' name as soon as possible, but to provide an ethically valuable contribute to the platform and the community. For this reason we opened an "enhancement" issue on git-hub with the title *Enhancement of DMI-Tcat making it facilitate GDPR-Compliance* where we suggested some possible changes. Among other measures we proposed the following scenario to be implemented:

> *An admin-user with access to the capture part decides to collect data. Due to he or she being careful not to use more personal identification data than necessary the pseudonymize check box is ticked when creating the capture bin. The program collect tweets as usual. But now the table tcat_captured_bins contains an attribute (a column) stating that the bin shall be pseudonymized for non-admin users. We then have a non-admin user who only has access to the analysis page. Since admin-user did not feel comfortable with exposing personal identification data the non-admin user can only access the data without the personal identification information. The data is pseudonymized. If a situation occurs where the non-admin user needs or wants access to the pseudonymized part, then he or she will need to take some action involving contact with the admin.*

This would have created a context in which the person responsible for the data collection should have reflected beforehand on the potential personal nature of the Twitter data and decide if pseudonymizing the data for the researchers accessing and exporting the data or not.

Currently the issue is still open on git-hub but, after several months from the last comment, it is fair to assume that it will never gather enough interest to be actually implemented. The reasons behind this failure were several. In the discussion tool developers questioned the inherent usefulness of this proposal on two accounts. First, they discussed whether the system administrator ought to be the one carrying responsibility for ethical decisions with respect to data collection eventually landing on the agreement that the researcher (here conceptualized as the end-user) ought to be the one taking on this responsibility. Thus the envisioned technical administrator does not need this facility in the first place. As one user noted: ”this creates a lot of coding overhead and maintainability issues.” Second, participants in this discussion quickly pointed out the futility of pseudonymization since Twitter’s search interface would allow anyone to quickly deanonymize the tweet anyway. Since the point of the data collection is to get the original textual content of the tweet (thus making anonymization of the tweets themselves inappropriate), pseudonymization of Twitter users was seen as superfluous.

There are two elements that are worth noting here. On the one side the responsibility for pseudonymization is attributed to the researcher (the end user) rather than to the person technically in charge of the data collection. This disregards the fact that many researchers run their own tcat instances and are, *de facto*, admin and end users simultaneously. This also raises the issue of how responsibility might or should be distributed in data-intensive research projects where not all researchers share the same level of technical expertise. Current research ethics guidelines do not address the issue of how such responsibility might need to be negotiated.

On the other side, there is the familiar argument that pseudonymization is pointless because a *simple* search on Twitter would immediately retrieve the original tweet with all the relevant information attached. The ethical concerns of the searchability of content have been discussed by social media researchers extensively. For the most part, however, these concerns have centered on unwanted disclosures perpetuated through publication. Some researchers have advocated for radical anonymization and modes of writing against search engines [26]. Others have proposed different levels of ‘disguise’ for publishing online content [10]. None of this discussion, however, has considered the ability to anonymize or pseudonymize data from the researchers themselves. While for some types of research the integrity of the tweet content may be required, there are growing trends in computational research where tweets are used as *bags-of-words*, such as for sentiment analysis of topic detection, rather than as meaningful messages [3]. Assuming that the text cannot be anonymized in any case so there is no reason to provide facilities for this, seems to impose an ethically lowest common denominator to all types of research.

The perspective here seems to echo some of the ethical considerations that we have seen emerging from the AoIR guideline where ethical research follows a procedural approach with specific ethical questions to be raised at specific moments during the research process. In our opinion, deeming an ethical decision (to not store deanonymized data) useless because malicious users could simply overcome it, resonates an idea of privacy focused on potential harm (and malicious users) rather than on users’ dignity [32, 4].

5 Software requirements and design

DMI-Tcat consists of a capturing and an analysis part. These parts work independently from each other and only have a database in common. An important detail about the different parts is that one needs to be an admin user to use the capturing part of the software. Hence only an admin user can collect data. This has the consequence that a ”regular” user will have to get settled with only having access to the analysis part of the software while the admin can use both the capture and analysis parts.

5.1 Pseudonymization

Efficiency and modularity were main criteria when planning for the pseudonymization functionality. We thus decided to have most of the pseudonymization code in one file. The plan was to force the original flow of the data through a file called *pseudonymization.php* where both pseudonymization and building and maintaining of a table consisting of the data needed for depseudonymization is done. By simply replacing the original data with pseudonymized data and then returning the pseudonymized data to the original flow of the program we would make sure that as little of the original code as possible was tampered with.

5.1.1 User stories

The user stories for the pseudonymization are the following:

- From the capturing interface the administrator should have the choice, before data collection begins, to pseudonymize the data. The tool should then mark the actual collection for pseudonymization. The tool should also create a *pseudonymization table* if it doesn't exist already. The pseudonymization table holds the pseudonymized value and its corresponding reference value. The pseudonymization table is the measure one needs to depseudonymize data.

- Whenever a user wants to do an analysis of the pseudonymized data, he or she can export the data as usual.

- During export of the data, the tool pseudonymizes the data by replacing the original identifiable data with a reference number. The original data is then stored in the pseudonymization table where the original data, the corresponding reference number and the type of original data are stored. The pseudonymized data is then given to the user.

- As an administrator, there exists a possibility to export the pseudonymization data from the analysis and export-interface. The tool delivers a CSV file to the user. The file contains all the information that was stored in the earlier step. With the help of the pseudonymization table the administrator can translate pseudonymized values and original data.

5.1.2 Specifications

To be able to fulfill these user stories, the following ideas and plans for the implementation were laid down. First, add a check box to the capture interface where you can choose if you want to pseudonymize the data you are about to capture. Then, make sure to have a way to be able to check which collections are pseudonymized or not. After suggestions from the community this was achieved by adding another column, pseudonymized, to the table tcat_query_bins.

Tcat_query_bins lists all the collections in the database and in our newly added column one can see if a collection is to be pseudonymized or not.

While a collection is started, there is already now a function that checks that all the necessary tables in the database exist. If any required table is missing the function creates it. By extending this function so it also checks and, if necessary, creates a pseudonymization table there will always exist a table when we need it. The table will contain the information that gives the possibility to depseudonymize the data residing in the query bin.

The pseudonymization table will have three columns:

- Reference value, also referred to as Pseudonymize value. This value will replace the original data in the pseudonymized table. This will be the primary key of the table. Therefore every value in this column needs to be unique.

- Original data is the data that we are replacing with a pseudonymized value. The original data is data that we have chosen to hide and instead show the corresponding pseudonymized value.

- Data type. This value describes what kind of data we are pseudonymizing. As an example, if we are pseudonymizing a screen name the data type field will contain the string 'screen name'. Since a lot of the values that we are pseudonymizing consist of numbers or more or less strange names that can appear confusing, this column makes the interpretation of the pseudonymization table easier.

The programming has been mainly done in the PHP language. This language has only one data type, the array. The array can be one dimensional or multidimensional with keys and values. When searching for a key in a multidimensional array it can be compared with a hash map when it comes to speed.

The tool creates a database and maintains some tables in the database. Some of the tables are the same for all of the collections, as an example the table that keeps track of all collections, with related information, of tweets.

Some tables are created exclusively for every collection of tweets that is initiated. Examples of these tables are the table containing all the tweets in a collection or the table containing all the hashtags in a collection. To illustrate the table structure there is an excerpt of the database tables used by the tool in listing 2. There one can see tables that are shared among all collections of tweets but also tables that are unique for the collection named ”small”.

```
+-----------------------------+
| Tables_in_twittercapture    |
+-----------------------------+
| small_hashtags              |
| small_media                 |
| small_mentions              |
| small_places                |
| small_tweets                |
| small_urls                  |
| small_withheld              |
| tcat_captured_phrases       |
| tcat_controller_tasklist    |
| tcat_error_gap              |
| tcat_error_ratelimit        |
| tcat_pseudonymized_data     |
| tcat_query_bins             |
| tcat_query_bins_periods     |
| tcat_query_bins_phrases     |
| tcat_query_bins_users       |
| tcat_query_phrases          |
| tcat_query_users            |
| tcat_status                 |
+-----------------------------+
```

Listing 1: Tables in a DMI-Tcat when there is one stored collection of tweets named *small*.

Listing 1 shows different tables containing different types of data. All the tables whose names start with *tcat_* are shared among all collections. An excerpt of the table small_tweets and its structure is shown in listing 2. To illustrate how the export and analysis page with its related functions interact, see Figure 1 and Figure 2.

```
+-------------------------------+----------------+
| Field                         | Type           |
+-------------------------------+----------------+
| id                            | bigint(20)     |
| created_at                    | datetime       |
| from_user_name                | varchar(255)   |
| from_user_id                  | bigint(20)     |
| from_user_lang                | varchar(16)    |
| from_user_Tweetcount          | int(11)        |
| from_user_followercount       | int(11)        |
| from_user_friendcount         | int(11)        |
| from_user_listed              | int(11)        |
| from_user_realname            | varchar(255)   |
| from_user_utcoffset          | int(11)       |
| from_user_timezone           | varchar(255)  |
| from_user_description        | varchar(255)  |
| from_user_url                | varchar(2048) |
| from_user_verified           | tinyint(1)    |
| from_user_profile_image_url  | varchar(400)  |
| from_user_created_at         | datetime      |
| from_user_withheld_scope     | varchar(32)   |
| from_user_favourites_count   | int(11)       |
| source                       | varchar(512)  |
| location                     | varchar(64)   |
| geo_lat                      | float(10,6)   |
| geo_lng                      | float(10,6)   |
| text                         | text          |
| reTweet_id                   | bigint(20)    |
| reTweet_count                | int(11)       |
| favorite_count               | int(11)       |
| to_user_id                   | bigint(20)    |
| to_user_name                 | varchar(255)  |
| in_reply_to_status_id        | bigint(20)    |
| filter_level                 | varchar(6)    |
| lang                         | varchar(16)   |
| possibly_sensitive           | tinyint(1)    |
| quoted_status_id             | bigint(20)    |
| withheld_copyright           | tinyint(1)    |
| withheld_scope               | varchar(32)   |
+------------------------------+---------------+
```

Listing 2: Columns containing information about a Tweet stored in the *small*_tweets table.

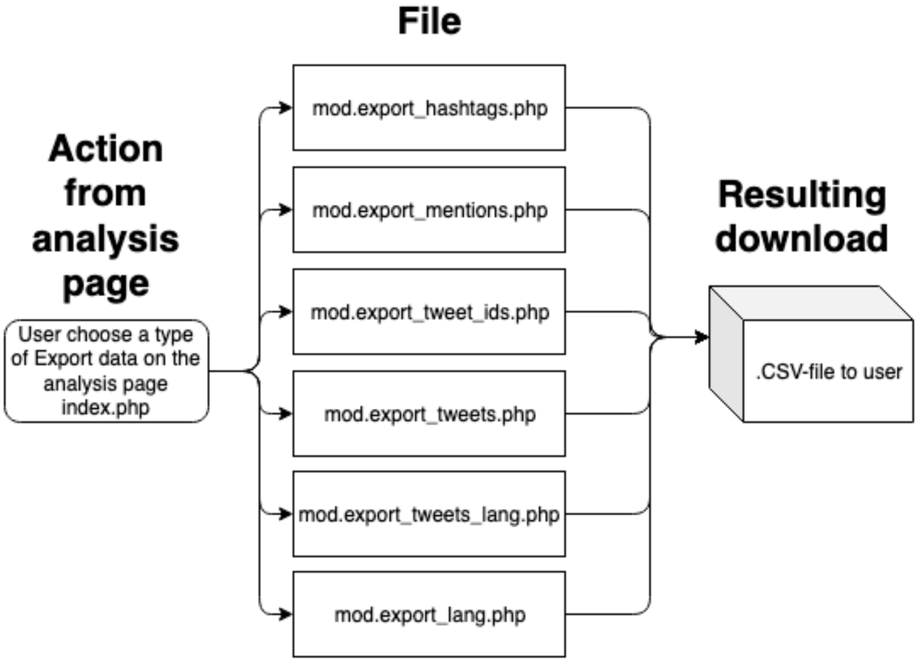


Figure 1: An illustration of how an export without pseudonymization works at the architectural level.

As seen in the illustration, if a user wishes to export data, he or she has multiple choices of what to export. For example a full export, using mod.export_tweets.php or just an export of hashtags using mod.export_hashtag.php.

When making a choice on the analysis page a request is sent to one of the analysis or export files tasked with retrieving the searched information and present it in form of an exportable CSV file.

```
while ($data = $rec->fetch(PDO::FETCH_ASSOC)) {
    $CSV->newrow();
    if (preg_match(”/_urls/”, $sql)
        || preg_match(”/_media/”, $sql)
        || preg_match(”/_mentions/”, $sql))
        $id = $data[’Tweet_id’];
    else
        $id = $data[’id’];
        $CSV->addfield($id);
        $CSV->addfield(strtotime($data[”created_at”]));
        $fields = array ( ’created_at’, ’from_user_name’,
        ’text’, ’filter_level’, ’possibly_sensitive’,
        ’withheld_copyright’, ’withheld_scope’, ’truncated’,
        ’reTweet_count’, ’favorite_count’, ’lang’, ’to_user_name’,
        ’in_reply_to_status_id’, ’quoted_status_id’, ’source’,
        ’location’, ’geo_lat’, ’geo_lng’, ’from_user_id’,
        ’from_user_realname’, ’from_user_verified’,
        ’from_user_description’, ’from_user_url’,
        ’from_user_profile_image_url’, ’from_user_utcoffset’,
        ’from_user_timezone’, ’from_user_lang’,
        ’from_user_Tweetcount’, ’from_user_followercount’,
        ’from_user_friendcount’, ’from_user_favourites_count’,
        ’from_user_listed’, ’from_user_withheld_scope’,
        ’from_user_created_at’ );
    foreach ($fields as $f) {
        $CSV->addfield(isset($data[$f]) ? $data[$f] : ’’);
    }
}
```

Listing 3: An extract from the file mod.export_tweets.php which is called when an export of all the tweets and their related data is chosen.

Of certain interest in listing 3 is the array named $fields on row 9. As shown this array consists of multiple keys, these keys correspond to the columns shown in the excerpt from the database in listing 2. On row 1 one can see that as long as there are tweets left that we want to export, these are saved in a variable called $data. The $data variable can be thought of as one row from the table ending with ”_tweets” and therefore represents one Tweet. On line 2 a CSV object gets a new row and in the loop at row 10 the fields that we wish to show as columns in the exported CSV table are written to the CSV object. Then we go back to row 1 to check if there are any tweets left to export.

For the pseudonymization to work effectively there is a need for a solution where a regular user just cannot access pseudonymized data. A regular user should thereby not be able choose whether the pseudonymization shall be activated or not. There has to be some kind of limitation, or an access policy where the pseudonymization is guaranteed and thereby forced by someone responsible for the data processing. Any other option would leave the pseudonymization completely useless since the point of protecting the data subject's data would be lost.

As mentioned above, with the ambition of reusing as much code as possible and to keep changes at the same place a new file called pseudonymization.php was created. The files purpose was to contain the code responsible for this feature.

5.1.3 Cooperation with other files

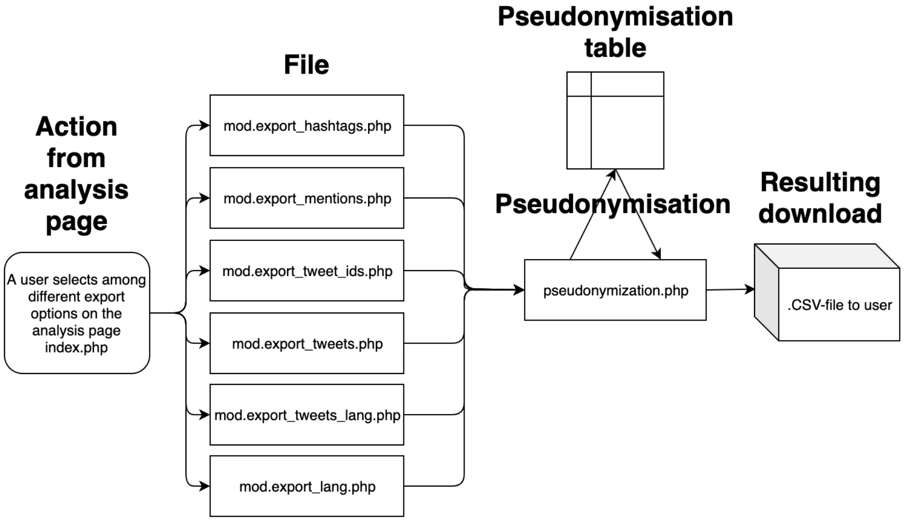


Figure 2: An illustration of how an export with pseudonymization works at a per-file-level.

Fortunately, the original design of DMI-Tcat is already very modular, which makes it fairly easy to implement the call to our module in the different original files.

```
while ($data = $rec->fetch(PDO::FETCH_ASSOC)) {

        $last_index = pseudonymize($data, $pp);

    $CSV->newrow();
    if (preg_match(”/_urls/”, $sql)
        || preg_match(”/_media/”, $sql)
        || preg_match(”/_mentions/”, $sql))
        $id = $data['Tweet_id'];
    else
        $id = $data['id'];
        $CSV->addfield($id);
        $CSV->addfield(strtotime($data["created_at"]));
        $fields = array ( 'created_at', 'from_user_name',
        'text', 'filter_level', 'possibly_sensitive',
        'withheld_copyright', 'withheld_scope', 'truncated',
        'reTweet_count', 'favorite_count', 'lang',
        'to_user_name', 'in_reply_to_status_id',
        'quoted_status_id', 'source', 'location', 'geo_lat',
        'geo_lng', 'from_user_id', 'from_user_realname',
        'from_user_verified', 'from_user_description',
        'from_user_url', 'from_user_profile_image_url',
        'from_user_utcoffset', 'from_user_timezone',
        'from_user_lang', 'from_user_Tweetcount',
        'from_user_followercount', 'from_user_friendcount',
        'from_user_favourites_count', 'from_user_listed',
        'from_user_withheld_scope', 'from_user_created_at' );
        foreach ($fields as $f) {
            $CSV->addfield(isset($data[$f]) ? $data[$f] : '');
        }
}
```

Listing 4: Excerpt from the same source code as in listing 3. Here with added pseudonymization functionality which is represented by the function call at line 3.

Since pseudonymization offers the possibility of identifying individuals if needed, there needs to be a data structure supporting depseudonymization. The structure is a table keeping track of what information that is masked by which reference number is needed. In this implementation, the table is the one visible in listing 5.

```
+---------------+--------------+------+-----+---------+----------------+
| Field         | Type         | Null | Key | Default | Extra          |
+---------------+--------------+------+-----+---------+----------------+
| pseudo_val    | bigint(11)   | NO   | PRI | NULL    | auto_increment |
| original_data | varchar(255) | NO   |     | NULL    |                |
| fieldtype     | varchar(255) | NO   |     | NULL    |                |
+---------------+--------------+------+-----+---------+----------------+
```

Listing 5: Structure of the pseudonymization table.

The different columns in listing 5 have the following functionality:

- pseudo_val is the reference value that will replace and thereby mask the original data in the table that is being pseudonymized. The column is set as primary key with auto incrementation, thus making sure that there is a unique replacement value for every unique original data that is masked.

- original_data consists of the data that we are masking. This column can consist of a screenname, a user id or any other data that we want to pseudonymize.

- The field type is a column with the purpose of explaining what kind of data is masked. For a pseudonymized id value, this column would simply contain the string ”id”.

To have the module pseudonymization.php acting as independently as possible there was a need for other supporting functionality in addition to direct pseudonymization functions. Since the module shall be able to take the data, process it and return it to the regular program the following tasks need to be taken care of:

- Fetch the existing pseudonymization table from the database and store it into an array thus making it accessible during the whole pseudonymization process.

- Check if a certain collection of tweets are marked for pseudonymization.

- Keep track of how many records exist in the pseudonymization table and how many are added during the pseudonymization process. Thus also keeping track of the pseudo_val:s increment and uniqueness.

- Save the added values of the pseudonymization table back to the database.

Algorithm 1 In this listing the procedure PSEUDONYMIZE is described using pseudocode.

```
procedure PSEUDONYMIZE($fields\_for\_pseudonymization$,$pseudonymization\_table$, $Tweet$, $startindex$)
  $fields\_for\_pseudonymization = an\ array\ consisting\ of\ all\ the\ fields\ that\ we\ would\ like\ to\ pseudonymize\ in\ a\ Tweet.$
  $pseudonymization\ table = a\ table\ where\ all\ pseudonymized\ original\ data\ are\ stored\ together\ with\ their\ pseudonymization\ value.$
  $Tweet\ =\ a\ Tweet\ and\ its\ fields\ of\ data.$
  $startindex = index\ of\ the\ last\ entry\ in\ the\ pseudonymization\ table.$
  for each field $f$ and its $value$ in $Tweet$ do
    if $f\ exists\ in\ fields\_for\_pseudonymization$ then
      if $value\ exists\ in\ pseudonymization\_table$ then
        $value = pseudonymization\_table[value]$
      else
        $startindex = startindex + 1$
        $pseudonymization\_table[startindex] = value$
        $value = startindex$
```

There exists one special case where this approach will not suffice and that is when there is a mention of a user in a text. Since a big use of Twitter is to retweet or in other ways refer to other users there are frequently appearing user names in text.

Since we want to pseudonymize only certain parts of the text and not the whole text we search the text with help of a regular expression that only matches the substrings containing mentioning of users that we want to replace.

Algorithm 2 In this listing we see the procedure PSEUDONYMIZE walking through textfields searching for substrings with aid from a regular expression.

```
procedure PSEUDONYMIZE($fields\_for\_pseudonymization$,$pseudonymization\_table$, $Tweet$, $startindex$)
  if $f ==' text'$ then

    $search\ f\ after\ any\ substrings\ starting\ with\ '@'$
    $string = appendstringwithallfindingsofsubstringsstartingwith'@'$
    for $every\ substring\ S\ in\ string$ do
      if $S\ exists\ in\ pseudonymization\_table$ then
        $value = pseudonymization\_table[value]$
        $replace\ occurences\ of\ substrings\ in\ string\ with\ S.$
      else
        $startindex = startindex + 1$
        $pseudonymization\_table[startindex] = value$
        $S = startindex$
        $replace\ occurences\ of\ substrings\ in\ string\ with\ S.$
```

With the above functionality in place, a pseudonymization of all the values is achieved.

5.2 Notifying the users

Based on the experiment described in Section 4, we decided to notify users by posting a tweet containing the hashtag that we are collecting (hereafter called beacon), with the help of Twitter's API. When it comes to collecting data based on a certain hashtag or location we assume that a user who is contributing to a hashtag-given discussion will probably also follow what others write with that hashtag. If the collector then publishes a beacon containing that hashtag and informs that a collection of tweets is done for that hashtag there is a probability that the data subject sees the message and hence knows about the collection and the fact that he or she is probably represented there.

To make sure that the beacon does not disappear too fast among all the other tweets containing the same hashtag we wanted a repeated publishing where we inform about the collection taking place.

To send the information we use a Cron job who runs the capturing script once every minute. Cron is a job-scheduler in Ubuntu where one can schedule repetitive tasks. In the case of DMI-Tcat, the repetitive task is calling a PHP file every minute to fetch data from the Twitter API and insert it to the database, this is taken care of by a Cron job.

To keep our impact on the architecture at a minimum we decided to take advantage of the existing Cron job as it was. Also, even if the Cron job run once per minute we only want to send one Tweet per day due to Twitter's rules for the API.

To *not* send a Tweet everytime the Cron job gets called we used the fact that the collections we are sending beacons for have a stored date and time when they where created. Every minute when the Cron job is called we check if it is the time on the day when the tracking of the collection begun. A hashtag that we started to track at 22:11 will therefore always have its beacon sent at 22:11 every 24 hours.

When posting tweets with Twitter's API one rule is that one cannot post identical tweets one after another. Since we wanted to repeatedly send our beacon Tweet this was an obstacle. We circumvented this issue by adding the time in days that had passed since the collection begun to our beacon message. Since we wanted to send one Tweet per day, the number representing days since start of the collection would always increase by one day for every sending, hence the tweets would not be identical. Our first beacon then mentions one day, the second beacon mentions two and so forth.

5.3 Giving access to the users' data

In addition to providing information, the beacon provides an opportunity for Twitter users to enter a page where they can get information regarding the storage of their tweets.

To be able to provide the data subject with information regarding the data collected from him or her we designed a page where the data subject can use Twitter credentials to log in. This is made possible by Twitter offering a solution called Twitter OAuth. Then we can use the Twitter user id to search our database for stored tweets belonging to the user.

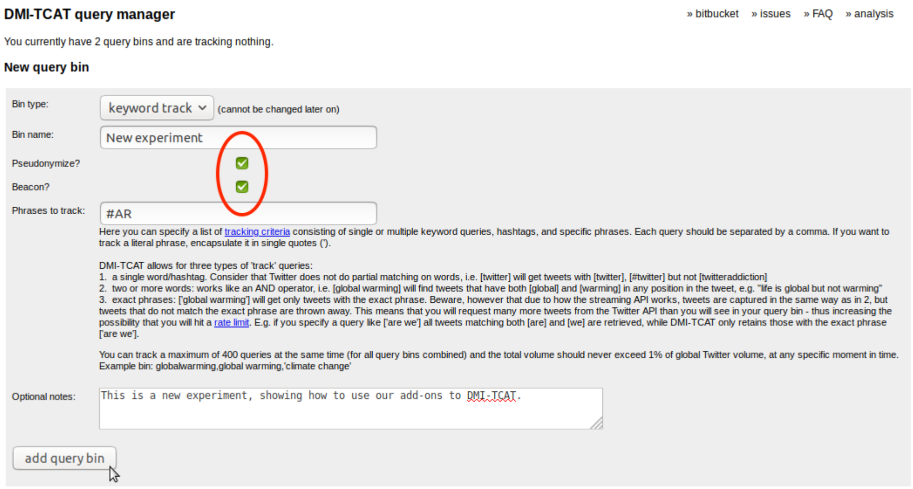
6 Overview of the tool

In this final section we provide an overview of the extended tool.

Figure 3 shows the interface used by the administrator of the installed tool to start new data collection processes. The interface is the same as in the original tool, with two additional options provided to the administrator: *Pseudonymize* and *Beacon* (marked by a red ellipsis in the figure). Notice that these are options that can be activated or deactivated: it should be a choice of the controller whether they are needed or potentially harmful for the project, based on a data privacy impact assessment. In the example, we have started a data collection of tweets containing the hashtag #AR (Augmented Reality), and asked the tool to pseudonymize the data and to inform Twitter users about the collection.

6.1 Pseudonymization

The result of the *Pseudonymize* option is that the analyst accessing the tool will only see pseudonymized data. For example, Table 3 shows the data provided to the analyst about the users whose tweets have been captured in the #AR data



DMI-TCAT query manager

You currently have 2 query bins and are tracking nothing.

New query bin

Bin type: keyword track (cannot be changed later on)

Bin name: New experiment

Pseudonymize? ☑

Beacon? ☑

Phrases to track: #AR

Here you can specify a list of tracking criteria consisting of single or multiple keyword queries, hashtags, and specific phrases. Each query should be separated by a comma. If you want to track a literal phrase, encapsulate it in single quotes (').

DMI-TCAT allows for three types of 'track' queries:
1. a single word/hashtag. Consider that Twitter does not do partial matching on words, i.e. [twitter] will get tweets with [twitter], [#twitter] but not [twitteraddiction]
2. two or more words: works like an AND operator, i.e. [global warming] will find tweets that have both [global] and [warming] in any position in the tweet, e.g. "life is global but not warming"
3. exact phrases: ['global warming'] will get only tweets with the exact phrase. Beware, however that due to how the streaming API works, tweets are captured in the same way as in 2, but tweets that do not match the exact phrase are thrown away. This means that you will request many more tweets from the Twitter API than you will see in your query bin - thus increasing the possibility that you will hit a rate limit. E.g. if you specify a query like ['are we'] all tweets matching both [are] and [we] are retrieved, while DMI-TCAT only retains those with the exact phrase ['are we'].

You can track a maximum of 400 queries at the same time (for all query bins combined) and the total volume should never exceed 1% of global Twitter volume, at any specific moment in time.
Example bin: globalwarming,global warming,'climate change'

Optional notes: This is a new experiment, showing how to use our add-ons to DMI-TCAT.

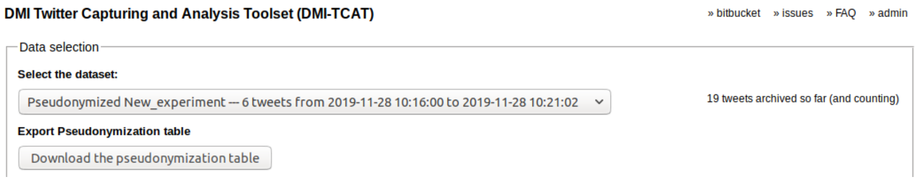
Figure 3: Interface to create new data collection processes, with two extensions

date	id	name	lang	tweets	followers	friend_count	listed	...
2019-11-28	1	2		26925	2993	2503	438	...
2019-11-28	3	4		904	205	405	8	...
2019-11-28	5	6		7716	392	141	29	...
2019-11-28	7	8		8120	686	768	6	...
2019-11-28	9	10		430	128	415	4	...
2019-11-28	11	12		59	1	0	0	...
2019-11-28	13	14		885	190	2099	3	...
2019-11-28	15	16		141	58	235	0	...

Table 3: Information about the users in the #AR data, pseudonymized

a few minutes after the collection has started. As it can be seen in the table, no user identifiers nor user names are visible, but they have been replaced by numbers.

The administrator has the possibility of recovering a table with the mapping from the codes to the original user identifiers and user names. This table is stored separately from the data and can only be accessed by the administrator, giving the possibility for the controller to set up a procedure for deanonymization. Figure 4 shows the interface of the administrator, with the option of downloading the pseudonymization table, and Table 4 shows an extract from the table where we only show one of our users with which we tweeted a message including the #AR hashtag to test the system.



DMI Twitter Capturing and Analysis Toolset (DMI-TCAT)

Data selection

Select the dataset: Pseudonymized New_experiment — 6 tweets from 2019-11-28 10:16:00 to 2019-11-28 10:21:02

19 tweets archived so far (and counting)

Export Pseudonymization table

Figure 4: Pseudonymization table

pseudo_val	original_data	field_type
...	...	...
3	81143606	id
4	dvladek	name
...	...	...

Table 4

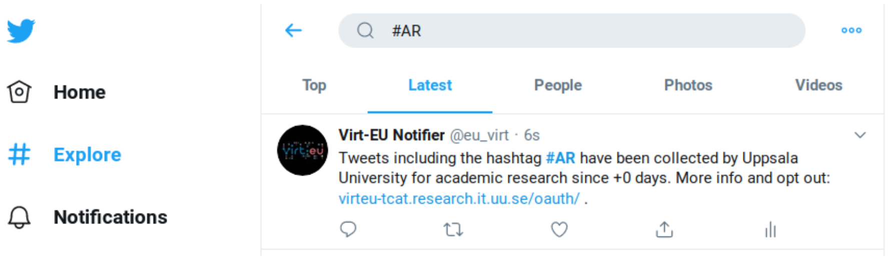
6.2 Notifying the users

As described in the previous section, when the option *Beacon* is activated then the tool starts regularly informing the users about the ongoing data collection. Figure 5 shows how a user interested in the #AR hashtag would also retrieve our information tweets, whose frequency can be decided in advance – at this time it is fixed for each installation.

6.3 Giving access to the users' data

If any Twitter user visits the page linked in the information message, s/he can authenticate to our system using Twitter credentials and see a list of the (active or inactive) data collection processes performed by the tool. For each process, the number of tweets *produced by the logged-in Twitter user* is also shown, as in Figure 6.

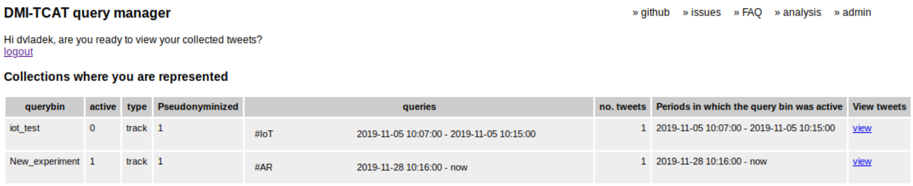
If some tweets of the user have been collected, then the user can also see the specific list of his/her tweets, as shown in Figure 7. In this way, the user may



Virt-EU Notifier @eu_virt · 6s

Tweets including the hashtag #AR have been collected by Uppsala University for academic research since +0 days. More info and opt out: virteu-tcat.research.it.uu.se/oauth/ .

Figure 5: A notification automatically posted by the data collection tool

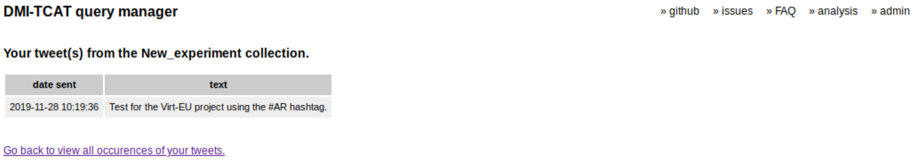


Hi dvladek, are you ready to view your collected tweets?

Collections where you are represented

querybin	active	type	Pseudonyminized	queries	no. tweets	Periods in which the query bin was active	View tweets
iot_test	0	track	1	#IoT 2019-11-05 10:07:00 - 2019-11-05 10:15:00	1	2019-11-05 10:07:00 - 2019-11-05 10:15:00	view
New_experiment	1	track	1	#AR 2019-11-28 10:16:00 - now	1	2019-11-28 10:16:00 - now	view

Figure 6: Checking if one's tweets have been captured



Your tweet(s) from the New_experiment collection.

date sent	text
2019-11-28 10:19:36	Test for the Virt-EU project using the #AR hashtag.

Go back to view all occurences of your tweets.

Figure 7: Getting one's own tweets

decide to ask the person responsible for the data collection to remove specific tweets. Notice that this option is not automated, because as specified in the GDPR there are situations where the right of the users to get their data removed is limited.

References

[1] Narayanan A. and E W Felten. No silver bullet: De-identification still doesn't work. Technical report, 2014.

[2] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *16th international conference on World Wide Web*, pages 181–190, 2007.

[3] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval.* Pearson, 2011.

[4] Edward J Bloustein. Privacy as an aspect of human dignity: An answer to dean prosser. *NYUL rev.*, 39:962, 1964.

[5] S.P. Borgatti and J.-L. Molina. Towards ethical guidelines for ethical research in organizations. *Social Networks*, 27, 2005.

[6] Erik Borra and Bernhard Rieder. Programmed method: Developing a toolset for capturing and analyzing tweets. *Aslib Journal of Information Management*, 2014.

[7] Cecile Bothorel, Juan David Cruz, Juan David, and Barbora Micenkova. Clustering attributed graphs: models, measures and methods. *Network Science*, 3(3):408–444, 2015.

[8] D. Boyd. *Social network sites as networked publics: Affordances, dynamics, and implications. A networked self.* Routledge, 2010.

[9] Alida Brill. *Nobody's business: paradoxes of privacy.* Addison-Wesley, 1990.

[10] Amy Bruckman. Studying the amateur artist: A perspective on disguising data collected inhuman subjects research on the internet. *Ethics and Inf. Technol.*, 4(3):217–231, November 2002.

[11] A. Cavoukian and K. El Emam. Dispelling the myths surrounding de-identification: Anonymization remains a strong tool for protecting privacy. Technical report, Information and Privacy Commissioner of Ontario, Canada, 2011.

[12] G. Chassang. The impact of the EU general data protection regulation on scientific research. *ecancer*, 11(709), 2017.

[13] Michele Coscia, Fosca Giannotti, and Dino Pedreschi. A classification for community discovery methods in complex networks. *Statistical Analysis and Data Mining*, 4(5):512–546, oct 2011.

[14] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[15] Obaida Hanteer, Luca Rossi, Davide Vega D'Aurelio, and Matteo Magnani. From Interaction to Participation: The Role of the Imagined Audience in Social Media Community Detection and an Application to Political Communication on Twitter. In *ASONAM*, pages 531–534. {IEEE} Computer Society, 2018.

[16] Fredrik Jonasson. A system for gdpr-compliant collection of social media data: from legal to software requirements (bachelor thesis).

[17] Gueorgi Kossinets. Effects of missing data in social networks. *Social networks*, 28(3):247–268, 2006.

[18] Andreas Kotsios, Matteo Magnani, Luca Rossi, Irina Shklovski, and Davide Vega. An Analysis of the Consequences of the General Data Protection Regulation (GDPR) on Social Network Research. *ACM Transactions on Social Computing*, mar 2019.

[19] Matteo Magnani, Danilo Montesi, and Luca Rossi. Conversation Retrieval from Social Media.

[20] C. Magnusson Sjöberg. Scientific Research and Academic e-Learning in Light of the EU's Legal Framework for Data Protection. In *New Technology, Big Data and the Law. Perspectives in Law, Business and Innovation.* 2017.

[21] L. Marelli and G. Testa. Scrutinizing the EU General Data Protection Regulation. *Science*, 360(6388), 2018.

[22] A. Narayanan, J. Huey, and E.W. Felten. A Precautionary Approach to Big Data Privacy. In *Data Protection on the Move*, pages 357–385. 2016.

[23] S. Penasa, I. de M. Beriain, C. Barbosa, A. Białek, T. Chortara, A.D. Pereira, P.N. Jiménez, T. Sroka, and M. Tomasi. The EU General Data Protection Regulation: How will it impact the regulation of research biobanks? Setting the legal frame in the Mediterranean and Eastern European area. *Medical Law International*, 2018.

[24] K. Schaar. What is important for Data Protection in science in the future? General and specific changes in data protection for scientific use resulting from the EU General Data Protection Regulation. In *Working Paper Series of the German Council for Social and Economic Data*. German Council for Social and Economic Data (RatSWD), 2016.

[25] I. Shklovski and J. Vertesi. Un-googling publications: the ethics and problems of anonymization. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 2169–2178, 2013.

[26] Irina Shklovski and Janet Vertesi. "un-googling" publications: The ethics and problems of anonymization. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, pages 2169–2178, New York, NY, USA, 2013. ACM.

[27] Carol Tavris and Susan Sadd. *The Redbook Report on Female Sexuality.* New York: Delacorte Press, 1977.

[28] W. C. van den Hoonaard. Is Anonymity an Artifact in Ethnographic Research? *Journal of Academic Ethics*, 1(2):141–151, 2003.

[29] Davide Vega and Matteo Magnani. Foundations of Temporal Text Networks. *Applied Network Science*, 3(1):25:1—-25:26, 2018.

[30] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*, volume 8 of *Structural analysis in the social sciences, 8.* Cambridge University Press, 1994.

[31] Jin Zhou, Xiaoke Xu, Jie Zhang, Junfeng Sun, Michael Small, and Jun-an Lu. Generating an Assortative Network With a Given Degree Distribution. *International Journal of Bifurcation and Chaos*, 18(11):3495–3502, 2008.

[32] Michael Zimmer. "but the data is already public": on the ethics of research in facebook. *Ethics and information technology*, 12(4):313–325, 2010.