Deliverable Number: 6.3

Deliverable Title: Report on legal and ethical framework and strategies related to access, use, re-use, dissemination and preservation of social media data

Work Package: 6 New forms of data: legal, ethical and quality matters

Deliverable type: Report

Dissemination status: Public

Submitted by: CESSDA, NSD

Authors: Sunniva Hagen (CESSDA, NSD), Elizabeth Lea Bishop (CESSDA, GESIS), Michal Koščík (CESSDA, CSDA), Martin Vavra (CESSDA, CSDA), Janez Štebe (CESSDA, ADP), Lorna Ryan (ESS, CITY), Eva Payne (CESSDA, NSD), Audun G. Løvlie (CESSDA, NSD), Linn-Merethe Rød (CESSDA, NSD), Øyvind Straume (CESSDA, NSD), Marianne Høgetveit Myhren (CESSDA, NSD).

Date submitted: August 2019

Legal disclaimer: This does not constitute or should not be construed as legal advice and/or guidance

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654221.

www.seriss.eu @SERISS_EU

SERISS (Synergies for Europe's Research Infrastructures in the Social Sciences) aims to exploit synergies, foster collaboration and develop shared standards between Europe's social science infrastructures in order to better equip these infrastructures to play a major role in addressing Europe's grand societal challenges and ensure that European policymaking is built on a solid base of the highest-quality socio-economic evidence.

The four-year project (2015-19) is a collaboration between the three leading European Research Infrastructures in the social sciences – the European Social Survey (ESS ERIC), the Survey for Health Aging and Retirement in Europe (SHARE ERIC) and the Consortium of European Social Science Data Archives (CESSDA AS) – and organisations representing the Generations and Gender Programme (GGP), European Values Study (EVS) and the WageIndicator Survey.

Work focuses on three key areas: Addressing key challenges for cross-national data collection, breaking down barriers between social science infrastructures and embracing the future of the social sciences.

Please cite this deliverable as: Hagen, S., Bishop, E., Koščík, M., Vavra, M., Štebe, J., Ryan, L., Payne, E., Løvlie, A., Straume, Ø., Rød, LM., Høgetveit Myhren, M (2019) Report on legal and ethical framework and strategies related to access, use, re-use, dissemination and preservation of social media data. Deliverable 6.3 of the SERISS project funded under the *European Union's Horizon 2020 research and innovation programme* GA No: 654221. Available at: wwwww.seriss.eu/resources/deliverables www.seriss.eu/resources/deliverables

Contents

1 Introduction

In 2018, an estimated 2.65 billion people were using social media worldwide.[1] The large amounts of digital data that are being generated from the online social media networks have become a significant source of data to study human and social behaviour across research disciplines. Some of the potential benefits of using social media data for research are the ability to reach larger numbers of participants, reduce the costs of conducting research in large populations, and the opportunities for interaction across extended time periods.[2] However, the use of social media data also present some legal and ethical challenges as will be further discussed.

Work Package 6 of the SERISS project addresses the major legal and ethical challenges facing cross-national social science research which relies on access to large-scale data at an individual level. The focus is on social surveys and the use of new data types in a social survey context in particular, including biomarker, social media data and administrative data.

The main focus of Task 6.1 was the legal requirements and ethical challenges that may come about when data arising from electronic communications are obtained from social networks, customer databases and tracking devices. The task addresses the steps that need to be taken to meet these challenges in order to increase and improve the use of these data sources.[3] This report is the output of deliverable 6.3 *Report on legal and ethical framework and strategies related to access, use, re-use, dissemination and preservation of social media data*.

It is expected that research should safeguard principles of research ethics as well as legal conditions and in many ways legal conditions and research ethics overlap. Legal and ethical frameworks and strategies will be discussed throughout the research process, covering research design, data collection, publishing and sharing. Note that this is always dependent on the research topic, on which social media data will be collected, and on the methods of analysis that will be used.

1.1 Legal and Ethical Framework

Research ethics focuses on optimising the rights of the researcher to carry out research with the rights and the interests of those whose data is gathered for the research, in particular relating to issues of autonomy, confidentiality (security of data) and benefits of participation in research. Data protection law seeks to ensure the exercising of rights, with some exceptions, by individuals over information as it relates to them. This task considered the legality and ethics of using social media in research. In legal terms, this guidance is not legally binding. In terms of research ethics, the subject matters are complex and issues can often only be resolved on a case-by case-basis. Users are encouraged to refer to existing disciplinary codes of practice while being aware that the contents are subject to debate.

1.1.1 Legal concepts associated with the data

Data are a foundation, a tool and a product of research. In other words, data are a *resource* that can be used to receive, store and distribute information or knowledge. Every resource has

[1] https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[2] Taylor and Pagliari, "Mining social media data: How are research sponsors and researchers addressing the ethical challenges?", 2017: 2-3

[3] European Commission, Directorate-General for Research and Innovation, 2016: 45

its value. Therefore, it is quite common to speak about data in connection to words that relate to property (e.g. “my data” and “their data”), transactions (“buy data”, “share data”, “give data on”) or even property crime (“someone stole my data”, “the application took my data and gave it to company XY”). However, this analogy has its limits, and questions such as “who owns the data” often leads to hesitation. This is because the data can be associated with several cultural and legal concepts. If someone says “My data”, they might mean that:

1) The data describe me as a person (e.g. date of birth, marital status, health record), relate to my personal life (e.g. backup copies of my emails and messages) or to my performance
2) The data are in my possession (e.g. they are stored on my hard-drive)
3) The data are the results of my work (e.g. I made measurements and records in the course of research)
4) The data are the results of my intellectual activity (calculations, analysis) and creative choices (e.g. shortlist of the books and articles I find useful for my work)

The table below gives an overview of legal concepts that relate to the four perceptions of the data mentioned above:

Categories of data	**Corresponding legal concepts**
Data related to an individual person and their life	Privacy and Personal Data
Data stored on one’s hard-drive	Property and ownership
Data created by effort and undertaking	Origination of data and Sui generis database rights (IP)
Data created by intellect and creativity	Copyright (IP)

Tab.1. Categories of data and corresponding legal concepts

If researchers want to access or use certain data, they may be unsure as to whether such use constitutes a violation of someone's rights and how to request proper permission. In other cases, the researcher might perceive that their rights were violated by someone who used “their” data without appropriate permission or adequate recognition of their work. The legal concepts that must be taken into consideration are explained below.

1.1.2 Protection of personal data and privacy

Right to privacy can be explained as a right to exclude oneself from public activity or a right to exclude information about oneself from public space. The concepts of privacy in various jurisdictions usually consist of two aspects:

1) Freedom from unreasonable constraints in the construction of one’s identity. This is often interpreted as a “right to be let alone”, which was first used in a US case-law and,

2) control over (some) aspects of the identity one projects to the world.[4] One of the most influential concepts is the right to informational self-determination developed by the German Constitutional court, which means the right of the individual to decide what information about them should be communicated to others and under what circumstances[5]

The *right to privacy* or *right to informational self-determination* in the context of social media means that no one should disclose private information that an individual does not want to be disclosed. The mere fact that the information is voluntarily disclosed does not mean it can be processed without further restrictions. Further processing of data which carry the information will also affect one's privacy and therefore needs to be regulated (regulated does not necessarily mean prohibited). Online privacy is regulated by general legal tools for data protection. The current General Data Protection Regulation [Regulation (EU) 2016/679, further referred to as GDPR] regulates purposes for which the processing is allowed ("legal bases") and conditions that have to be respected in the course of processing such data.

The GDPR applies to the processing of personal data, which is defined as any information relating to an identified or identifiable natural person (data subject). The person who determines the scope and purpose of the processing (i.e. who makes the decision to process the data) is the controller of the data. The person who assists the controller by providing services, but is not interested in the ultimate purpose of processing is defined as a processor. Under the GDPR, controllers still bear the primary responsibility for compliance, although, processors also have direct compliance obligations (see Recital 74 and Article 24).

Processing of personal data must be met with one of the legal bases in Article 6 (1) in order for it to be lawful. The most common legal bases for the processing of personal data for research purposes are:

- *Consent* from the data subject for one or more specific purposes[6]
- Processing is necessary for the performance of a task carried out in the *public interest* or in the exercise of official authority vested in the controller[7]
- Processing is necessary for the purposes of the *legitimate interests* pursued by the controller, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject

There are cases where data can be processed even without the explicit consent of the data subject. Although research is not specifically mentioned as a legitimate interest in the GDPR, the recital no.157 identifies the benefits associated with personal data research, including the potential for new knowledge about "widespread medical conditions" and the "long-term correlation of a number of social conditions." However, the legitimate interest basis sets out a balancing test, where even if a controller has a legitimate interest in research, it may be "overridden" by the data subject's rights.

[4] Rouvroy, A., and Poullet, Y., "The right to informational self-determination and the value of self-development: Reassessing the importance of privacy for democracy", 2009: 46

[5] Westin, A.F., "Privacy and freedom", 1970

[6] See also part 3.1.2

[7] See also part 3.1.3

1.1.3 Data ownership and data as a tangible asset

Unlike information which is intangible and cannot be “owned”, data can exist in tangible form as a record (analogous or digital) contained in the storage medium. The storage medium itself is tangible and can be owned and used, sold and destroyed, like any other thing. The owner of the storage medium also owns the data contained in it and can exclude others from accessing the data. The mere ownership of the storage medium (record) however:

1) does not constitute the right to exclude others from having the same data on their own storage mediums,
2) has no relevance to the question whether these data can be further shared or copied.

In order to exclude others from copying or distributing data, the data have to be protected as an intangible asset - intellectual property.

1.1.4 Data as an intangible asset - intellectual property

The European legal framework does not recognize any exclusive right to the data. The originator of the data, i.e. the person who wrote, typed or recorded the data into the storage has no exclusive right to them. The EU law, however, recognizes intellectual property rights to *databases.* The database is defined as *“a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means”*[8]. The database can be protected by copyright if it constitutes the author's own intellectual creation “by reason of the selection or arrangement of their contents[9]”. The copyright does not necessarily belong to the person who created the individual elements of the database, but to the person who created the database by picking individual elements and arranged them in a creative form. Not every database is created by individual choices and creativity, but can still be valuable. A good example is databases that store the data gathered automatically by means of sensors or algorithms. The EU law recognizes a sui generis right of a database maker that protects the undertaken effort and investment to the creation of the database. A person (even legal entity) who makes a qualitatively and/or quantitatively substantial investment in either the obtaining, verification or presentation of the contents gets exclusive right to prevent extraction and/or re-utilization of the whole or of a substantial part of the contents of that database[10].

1.1.5 Contracts and terms of service as a means of information and data

Everything which is not forbidden is allowed. When working with the data the researcher has a liberty to use them in any way that does not violate the privacy or exclusive rights of others. The researcher however has to understand the concept of contract, which can enable them to use protected content, but also restrict them. If the data or database is protected by exclusive rights, the rightsholder may authorize the researcher to use the data by means of license contract/agreement. On the other hand, the terms of service of an online platform or social network may prohibit the researcher (and any other user) to use the data even if they are not protected by exclusive rights.

In certain cases, the researcher may use copyrighted or otherwise protected content even without the consent of a rightsholder. Member states are allowed to grant statutory copyright

[8] Art.1 par.2. of the Directive 96/9/EC on the legal protection of databases

[9] Art.3 par.1. of the Directive 96/9/EC on the legal protection of databases

[10] Art.7 par.1. of the Directive 96/9/EC on the legal protection of databases

exceptions that will enable extraction of data for the purposes or scientific research, as long as the source is indicated and to the extent justified by the non-commercial purpose to be achieved. This exception is not applicable for datasets that are not protected by sui generis rights (i.e. most databases created by US companies) that are protected only by the terms of service.

2 Social media data in social scientific research

2.1 Social media in research – What is new?

The availability of searchable traces of human communication on a large scale provides new opportunities for research[11], new ways to connect data and people, as well as a new and instantly interesting and recognisable subject of study that is constantly changing. Still the use of social media data also raises some new legal and ethical questions for researchers and data repositories, such as the difficulties with regard to ensuring anonymity of data subjects, the boundaries between 'public' and 'private', as well as questions of data ownership and intellectual property.[12]

Social media research includes any form of research using data obtained from social media sources. Research in this case can be classified into two broad types: using social media services and platforms as research tools and research on the activity and content of social media services and platforms themselves.[13] The source of social media data used in research today appears to be mainly platforms such as Twitter and Facebook, yet social media platforms like Snapchat, LinkedIn, and others, still contribute to the increased quantity and relevance of social media data.

Whilst data in the social media world consists of what humans care or think about, which also can be available through observations, interviews and surveys, at least one element differentiates the social media platforms from many other social arenas and that is the technological foundation; the organisation and storage of all interactions that happens on and through social media outlets on massive servers. In online communities human thoughts and activities take the form of data that can be scraped, downloaded, aggregated, and otherwise collected on a massive scale.[14] This represents a change in how we can gather, analyse and organise data of the social media population, and may lead to new data about the "human condition".[15]

The collection of data from social media platforms has provided large volumes of naturally occurring data with exceptional temporal and geographical granularity.[16] Even though the "capture" of such data may correspond to the data collection phase in the traditional research data lifecycle, the data are likely to have been created for a purpose other than research.

In addition to the legal and ethical issues mentioned above, a challenge concerning social media data may also include the self-selecting nature of social media users and inequalities in access to social media platforms and data.[17] Whether social media platform users are

[11] Kotsios, A., Magnani, M., Rossi, L., Shklovski, I., Vega, D., "An Analysis of the Consequences of the General Data Protection Regulation (GDPR) on Social Network Research", 2019

[12] Taylor and Pagliari, 2017

[13] Social Media Research Group, "Using social media for social research: An introduction", 2016: 6

[14] Mannheimer, S. and Hull, E.A., "Sharing Selves: Developing an Ethical Framework for Curating Social Media Data", 2017: 197

[15] One example is the myPersonality# project by Michal Kosinski and David Stillwell. This project has access to data from more than 4 million Facebook profiles. The data collected are a number of likes and the posts (updates) of the users. Additionally, one can link Twitter posts to the project. The project is not limited to linking and interpreting social media data; it automatically categorises the personal social media data into established psychological ideal types, or typologies, based on the individual's online footprint, and presents these and other findings to the participant.

[16] Edwards et al., "Digital social research, social media and the sociological imagination: Surrogacy, augmentation and re-orientation", 2013

[17] Taylor and Pagliari, 2017: 3-4

representative of a broader population or not, or even whether the subset the researcher gets access to is representative of the population on that platform, are questions to keep in mind. Further, the data may be unstructured, constantly changing and require new analytical techniques.[18] Their quality can also be difficult to assess since it is often user-generated, or combines self-reported and behavioural traces.[19] The rapid development of new social media services and technologies is another important element to consider in light of this, as well as the changes in services and target groups, and what new types of interactions are documented and stored for the services that the social media platforms develop.

2.2 Defining social media data

There are many different definitions of what social media are and are not. In "Using social media for social research: An Introduction" social media are understood to be web-based platforms that enable and facilitate users to generate and share content, allowing subsequent online-interactions with other users.[20]

Our working definition describes social media data as information created and/or provided by individual users (and/or groups) selected, organised and/or shared through – and collected from – online social spaces, such as blogs and microblogging sites (Twitter), social networking sites (Facebook) and content communities (Instagram).

Social media data is considered to be constituted by a type of agency in the generation, creation and/or sharing of the content based on some level/kind of intent. A central aspect is that it is a means of communication, based around an online service, where the content being communicated is produced by people using the social media services. That the content can be shared privately through messaging services, or publicly on social media "walls" and through "feeds" is also of importance. Still, this intent does not necessarily include awareness of how widespread the content is shared nor exactly who will be able to view the shared content.[21]

2.3 Current state of affairs

Researchers are currently using and analysing social media data with various methods. Whilst until 2012, 38% of, for example, Twitter research came from the field of computer science[22], recent years have seen the increasing use of social media data across research disciplines. In a study conducted by SAGE publishing[23] in 2016, among social scientists around the world, one can read that 33 percent of the respondents had been involved in big data research of some kind. Of the researchers that had not yet engaged in big data research, 49 percent (3057 respondents) said that they were either "definitely planning on doing so in the future" or "might do so". Several of the researchers responded that they were affiliated with interdisciplinary research centres or labs focusing on big data, mobile, online and web studies in some variation or other, but the vast majority were unsure whether their universities had any such centres or labs. As regards social media data, the study showed that 927 of a total of 9412 respondents had used social media data (from Facebook, Twitter and/or other services) in their most recent

[18] OECD, "Research Ethics and New Forms of Data for Social and Economic Research", 2016

[19] Davies, L., "Social media data in research: a review of the current landscape", 2019

[20] Social Media Research Group, "Using social media for social research: An introduction", 2016

[21] Obar, J., and Oeldorf-Hirsch, A., "The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services", 2018

[22] Davies, L., "Social media in research: a review of the current landscape", 2019

[23] Metzler, K., Kim, D. A., Allum, N., & Denman, A., "Who is doing computational social science? Trends in big data research", 2016

project involving ‘big data’. This gives a clear picture that social media data enters into and is used by researchers, and increasingly so.[24]

2.3.1 Projects registered with the Data Protection Services at NSD

In an effort to get an overview of the contemporary conditions of social media research in Europe, the task group looked into the database of projects registered with the Data Protection Services at NSD.[25] This is however one use case and cannot be generalized to the whole of Europe.

The assessment was only based on what documentation appeared in the Data Protection Services database, and as such the end result is not based on a certainty regarding reliability and validity of the findings, as details in documentation, written assessment and other variables are not, in a majority of the cases, conclusive. To make up for some of the level of uncertainty, it was endeavoured to discriminate strictly when assessing the projects. Therefore, the number of projects is lower than it could have been, especially considering that some projects tend not to notify the DPO when the data is “freely available” in a “public” sphere e.g. Facebook or Twitter, and that some projects that should be registered with the DPO never are due to a general lack of awareness. It would also be safe to assume that the general number of projects that are related to social media is significantly higher, as numerous students and researchers also commit to conduct projects anonymously.

Having mapped research activities on and/or using social media data, the statistics from NSD shows that of the total number of projects from 2007 to 2016, about 650 research projects were registered that with some level of certainty made use of social media data, explored the internet and/or social media through research, or intended to make use of other methods of collecting personal information through digital technologies. Figure 1 shows the increase in projects related to internet, blog, and social media research as notified to the DPO at NSD from 2007 to 2016.

[24] Davies, L. (2019) also notes that labs and research groups highlight its place in social science research with examples such as the Social Media Lab and Social Data Lab.

[25] The approach consisted of using a specially designed search engine of the database to extract any project that had variables stating a use of social media/internet/blog as a source of personal data, as well as searching the DPO written assessment for keywords like: Twitter, Facebook, and Social Media. Several other variables were extracted to be able to assess the likelihood of the project actually using or related to social media data in some form or the other, and if so in what way. The search yielded a large quantity of projects that were presently evaluated and assessed, and then coded according to four binary variables: using social media data for the research project, using social media to recruit to the research project, the topic of the project was social media, and the project linked social media data with other data sources (survey, administrative data, interview data, etc). These four variables are not mutually exclusive, so any one project may qualify for none, one or more of the variables.

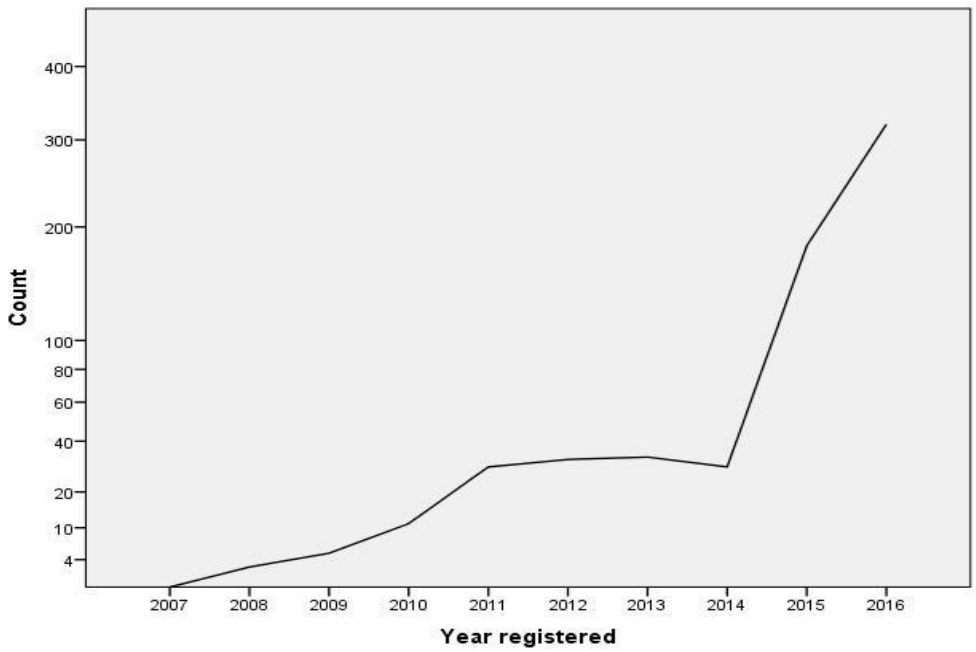

Fig.1. Projects related to internet, blog, and social media research as notified to the DPO at NSD, 2007-2016

The figure above shows a drastic increase from 2014 to 2016. The reason it spikes after 2014 is that the means of documenting and categorising projects at the DPO at NSD changed in 2014, adding a new category for internet, social media and blogs as sources for collecting personal information. Therefore, it would be reasonable to assume that the sudden increase should have been more evenly distributed over this period. There is no doubt, however, that social media research has increased with the advent of social media as the norm for communication and sharing information for a significant portion of the population.

One limitation when it comes to registered projects and getting a valid and reliable overview is the terminology used. Social media platforms have existed as far back as in the early days of MSN messenger, ICQ and mIRC amongst others, and the technology, and in some cases the data, have been around since at least the mid to late 1990s. Yet it was not until the advent of Twitter and Facebook that social media seemed to become of widespread interest for the social scientific research community.

Taking into account that information on research conducted on or with data from social media is sparse and not well documented over time, we can still get an indication of how it is and has been used. Figure 2 shows three different trends in relation to the use of social media data.

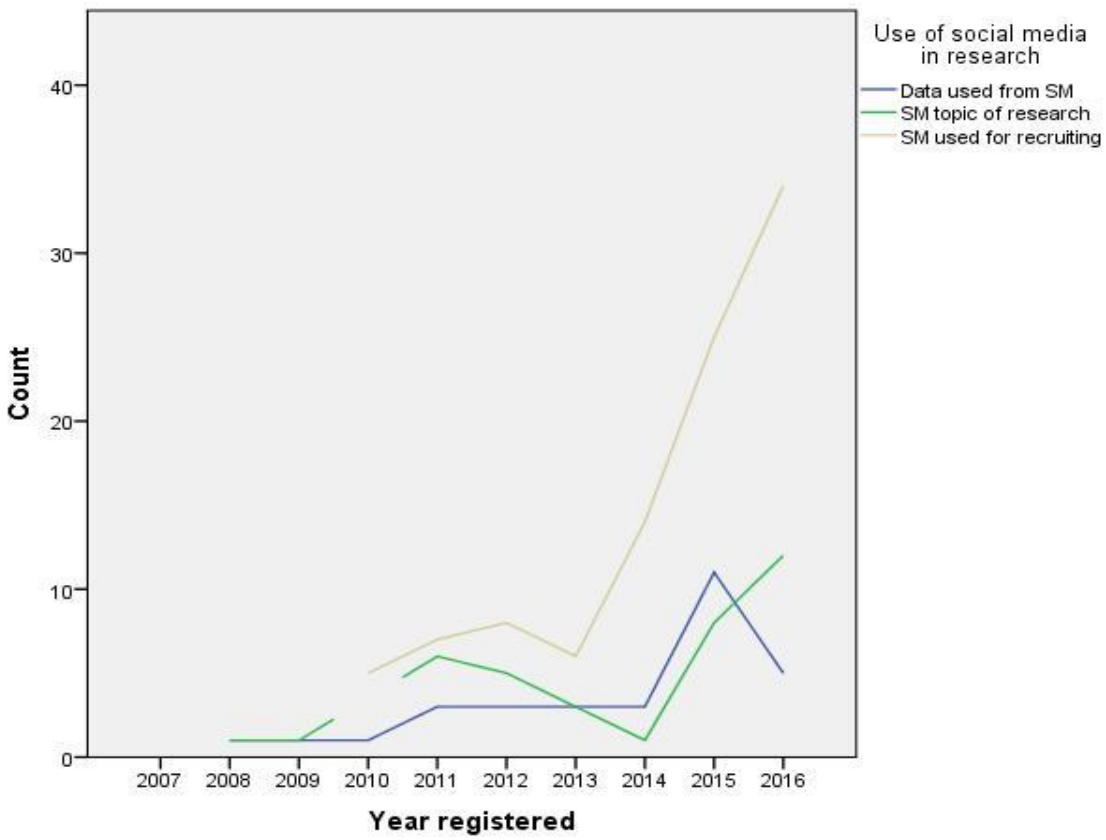

Fig.2. Projects using social media (SM) for various research purposes, 2007-2016

Isolating the various uses of social media services in research into three categories (to collect data, to recruit participants or where the phenomenon is the topic of the research itself) one observes that the most prevalent use of social media services in this period was as a recruitment platform. This emphasises the ethical and legal issues concerning how to best recruit, inform and obtain consent through social media channels in a legal and ethically sound manner, regardless of whether the method of collecting data is interview, survey, or some other method. As with fig.1, the increase during 2014 should be partly attributed to changes made in NSDs documentation options, although this increase started earlier for recruitment.

With about 650 projects over this 10-year period the total number of cases is fairly low when compared to the total number of research project notified to NSD during the period (which was 34826 projects). As noted this may to a certain extent be due to a lack of awareness of individuals' rights leading to underreporting.

2.3.2 Awareness of legal and ethical issues

In addition to reviews of the current literature, the task group had interviews/conversations with some social media researchers to get insight into awareness of legal and ethical sides of social scientific research in relation to social media and internet research. Two researchers based in Norway and one researcher based in Sweden provided their insight. The literature review and the researcher feedback provided the following insight into the field of social media research[26]:

[26] Please note that this work was carried out by the task group in 2016-2017 and that status could be changed since the time of writing.

A. Among the researchers, the impression was that what social media platform users publish and post was considered to be “fair game”, and that issues concerning consent and/or informing the users of the researchers’ use of their data was rarely considered thoroughly.
    1. A part of the justification of this was that informing (and asking for consent) was a too strict requirement, and it was often argued to be impossible to inform social media platform users.
    2. There was a general perception that posts and information shared “publicly” via social media services waives the individuals’ rights of privacy.
    3. Further, the term “anonymous” was often understood as only lacking direct identifiers, such as full name, social security number, telephone number, address and the like. Indirectly identifying variables seemed to be forgotten to a certain extent.
B. From researchers we learned that gaining access to user data through other means than personal observation and using searching/scraping methods could be expensive, if data must be bought from a social media platform provider. Further, personal observation and scraping methods are generally unreliable and inaccurate, particularly when using open source and publicly available automated methods for such data collection.
    1. Other aspects of the difficulties of gaining access to data were legal aspects such as copyright and the social media platforms’ terms of use.

Based on the inquiries made into this part of the project, researchers and articles in several journals only touched upon the ethics of research and legalities. While there was an ethical awareness in the research field, and awareness about data protection, the general attitude was that these topics were covered and treated with only a cursory glance, if covered at all. Regarding awareness of ethical issues, Williams, Burnap and Sloan[27] also make some similar points, noting that papers fail to include a mention of the ethics of conducting social media research.[28]

2.3.3 Existing guidelines and frameworks

There exist several codes of conduct, guidelines and frameworks when it comes to research ethics in internet research. As an example, The Association of Internet Researchers (AOIR), a member based academic association that promotes critical internet research across disciplines, have produced a guide for Internet research ethics to ensure commitment to serious, professional and ethical research on and about the Internet[29]. Further, ESOMAR, an international membership organisation for market research professionals, has constructed a

[27] Williams, M., Burnap, P., and Sloan, L., “Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users’ Views, Online Context and Algorithmic Estimation”, 2017

[28] As to awareness among the public, a study conducted by Evans, Ginnis and Bartlett from 2015 (Evans et al. 2015: 7-9), show a low level of awareness and trust of social media research among social media users. A quantitative survey of 1250 adults (16-75) in the UK, qualitative workshops and a statistical analysis as part of the survey were carried out. In the survey 38 % answered that they thought sharing social media data with third parties for the purposes of research currently happens under the terms and conditions they signed up to on social media sites. When asked to review how likely they were to approve a social media project the results showed that they were not very likely to. An important factor in the likelihood to approve social media projects were whether data was already public available prior to the research project, as well as the level of anonymity, who the project was for, and how personal the information being used was.

[29] https://aoir.org/ethics/

set of guidelines for online research[30] as well as for using data from social media[31] and a data protection checklist.[32] The OECD also released a policy paper on research ethics and new forms of data (i.e. social media data) in 2016[33] providing insights and guidelines, based on two broad groups of ethical principles concerning: inter-researcher relations and methodological standards; and respect, justice and ensuring good consequences, respectively. Another example is "A Guide to Internet Research Ethics" issued by The National Committee for Research Ethics in the Social Sciences and the Humanities (NESH).[34]

[30] https://www.esomar.org/what-we-do/code-guidelines/esomargrbn-online-research-guideline

[31] https://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ESOMAR-Guideline-on-Social-Media-Research.pdf

[32] https://www.esomar.org/what-we-do/code-guidelines/esomar-data-protection-checklist

[33] OECD Science, 2016

[34] The National Committee for Research Ethics in the Social Sciences and the Humanities (NESH), «A Guide to Internet Research Ethics», 2018

3 Research design and data collection

3.1 Legal and ethical issues

Social media data that is processed for research purposes is often personal data. The GDPR highlights personal data protection as an important ethical issue, with clear principles related to processing personal data. Legal and ethical issues related to the collection and use of social media data are thus often inextricably linked, and topics such as research approval, informed consent, potential risk to users, and the public/private nature of social media data, are important to discuss from both a legal and ethical standpoint. It is also important to consider the differences between the collection of data directly from social media users, and data collection where there is no direct contact between the researcher and the users whose social media data is being processed.

3.1.1 Research approval

Research involving human subjects often requires a research ethics review. Requirements for an ethics review may vary between countries, institutions and disciplines. Large scale studies involving scraping from social media platforms raise the question as to whether the gathered data are to be treated as text (no human input) or as human subjects' data. In general, there is an understanding that there has been human input at some point in the process of creation, and that the consequences for individuals and communities should be considered, even if there is no requirement for a formal research ethics review. If ethical review is required, the relevant body (e.g. ethical board, IRB, legal department) should be consulted before data collection begins, and if major changes are made to the project.

When a research project involves processing personal data, the data controller must meet certain requirements in the GDPR, including documenting the legal basis for processing. Depending on member state law, and the nature of the project, it may be necessary for the project to be assessed by the Data Protection Officer at the research institution, or an equivalent body. This assessment will often take into consideration ethical issues, and projects that involve high risk for social media users (data subjects) will require that a Data Protection Impact Assessment (DPIA) is carried out.

Also, when accessing social media data, permission by the owner/manager of the data (e.g. social media databases) may be required.

3.1.2 Informed consent: ethical consent and legal consent

Processing personal data must be met with one of the legal bases in Article 6 .1 in order for it to be lawful. The most relevant legal bases for processing personal data for research purposes are consent and a task carried out in the public interest.

In order for consent to be a valid legal basis for the processing of personal data it must be freely given, specific, informed and unambiguous.[35] It is important to note that for consent to

[35] http://www.privacy-regulation.eu/en/recital-32-GDPR.htm

be valid the data controller must be able to demonstrate that informed consent has been gained from users whose data will be processed. [3637]
The "informed" aspect of consent as a valid legal basis involves providing participants with information that meets requirements of form and content, as specified in the GDPR. Article 13 lists the categories of information with which data subjects must be provided when personal data is collected from the data subject. In order for this information to be accurate, the researcher must have considered legal and ethical issues relating to the processing activities, including the purpose(s) of processing, the security arrangements, the approximate duration of processing, the rights of data subjects and how these rights can be exercised.

The term "informed consent" is often understood as consent which meets legal conditions for consent and also adheres to principles of research ethics. However, it is possible to distinguish between consent as a legal basis (i.e. legal consent) and ethical consent. Ethical consent is in most cases a prerequisite for carrying out research where the researcher is in direct contact with participants. Therefore, when researchers are in direct contact with social media users as part of the project design (e.g. when linking survey data with social media data), it is necessary to gain ethical consent from the users. This is irrespective of whether consent is the legal basis for processing personal data or not. Note that when the researcher is in direct contact with users then information should always be provided.

When data is not collected directly from social media users, the researcher should consider whether legal and/or ethical consent can and should be gained, including the extent to which it is possible to get in contact with social media users and the type of social media data that will be processed. "Informed consent" as a both legal and ethical consent could be considered best practice in a project that will collect highly sensitive data that has been shared by a small number of social media users, where it is both possible and reasonable for the researcher to gain consent. Note that the obligation to provide information is independent of the legal basis for processing personal data.[3839]

3.1.3 A task carried out in the public interest

In some cases, it may be difficult to gain valid legal consent from social media users, due to the requirements for form, content and documentation in the GDPR and the way in which the Article 29 working party have operationalised consent. Gaining consent which is "informed" can be a challenge in complex research projects where it is difficult to ensure that participants understand the implications of their consent (e.g. remember what they have posted, know what they will post, or understand the implications of having their social media data analysed). In other projects, it may be difficult to achieve research purposes whilst also meeting documentation requirements for legal consent. It may also be difficult or impossible to get in contact with social media users and gain consent.

Article 6(1)(e) in the GDPR offers an alternative legal basis when "processing is necessary for the performance of a task carried out in the public interest". It is important to note that member state law may include special provisions for use of this legal basis, including that appropriate

[36] http://www.privacy-regulation.eu/en/article-7-conditions-for-consent-GDPR.htm

[37] Information on how to document consent can be found on the Information Commissioner's Office (ICO) website: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/consent/

[38] Article 14 lists the categories of information to be provided where personal data is not obtained from the data subject.

[39] Please see section 4.1.1 for a further discussion on informed consent and unconsented data.

measures are taken to safeguard the rights and freedoms of users. In cases where it is not possible to gain valid legal consent, gaining ethical consent is a significant means of protecting data subjects' interests.

3.1.4 No contact with social media users

In projects where researchers harvest data from social media, relating to a large number of users, it may not be possible or may involve a disproportionate amount of effort to contact individual users in order to give information and gain consent for the processing of their data.[40] In such projects it is important to remember that without information, the social media users will likely not be in a position to exercise their rights as data subjects.[41] Information plays an important ethical and legal role in research involving personal data, and a lack of information will increase the potential harm for social media users.

When social media data is to be processed on the basis of a "task in the public interest", and the researcher is to be exempted from the obligation to provide information, the societal benefit of the processing should outweigh the risks for social media users. It is therefore important for the researcher to identify potential risks, to take necessary measures in order to secure the collected data and, not least, to argue for the societal benefit of the research. This should be carried out before data collection begins and is often facilitated by creating a data management plan.

3.1.5 Sensitive data and vulnerable individuals

The 'scope' of data can be understood as its degree of sensitivity, the extent to which the data is personally identifiable, and the amount of data being processed, both relating to each user and in total. The extent to which text can be searched for using a social media search engine is also an aspect that will increase the scope of the collected data. As a general rule, the larger the scope, the higher the potential harm/risk for social media users.

The researcher should also consider the possibility that social media users may include vulnerable individuals, such as children, persons who are mentally impaired, persons who are unable to give consent etc. The researcher's inability to determine the mental capacity or verify the age of social media users should be taken into account when determining the degree of sensitivity and the potential risk for social media users.

It is important to note that the processing of special (sensitive) categories of personal data is prohibited unless the research project has a valid legal basis. Special categories of personal data may be processed for scientific research purposes under Article 9 (2)(j) of the GDPR, on the condition that appropriate safeguards for the rights and freedoms of data subjects are be implemented.[42] The researcher should therefore consider all aspects of the 'scope' of the data being collected, and the categories of social media users who data will be collected, in order to identify the potential risk/harm for the users. The planned safeguards should be sufficient to reduce the risk to an acceptable level.

[40] See article 14.5 b) in GDPR for exemption from obligation to provide information

[41] Rights in articles 15-22 of the GDPR. Exemptions can be made in certain circumstances, for example, when the right to erasure is likely to render impossible or seriously impair the achievement of the research objectives

[42] See Article 89 (1) in the GDPR

3.1.6 Public/private

Another aspect of social media data which should be considered from both an ethical and legal perspective before data collection begins is the extent to which data may be considered more public, or more private. The public/private distinction is based not only on the ways in which social media platforms function, including the privacy settings that are available, but also on the ways in which these platforms are used (as will be discussed in part 4.1.3).

An ethical assessment of the use of social media data should take into account the expectations of the users themselves. From a legal perspective, it is important to remember that social media data that relates to individual persons can have been shared/published by the person themselves, or shared/published on social media by others. Article 9.2 e) in the GDPR provides a legal basis for the processing of special (sensitive) categories of personal data when the data are “manifestly made public by the data subject”. In the context of social media research, the social media user must have published/shared information about themselves, and it must be clear that the intention was to make this information ‘public’. Processing social media data which the user has shared with the aim of reaching as many people as possible arguably carries less risk for these users than social media data which have been shared in a more closed or private forum (e.g. where a smaller group of users communicate). Moreover, when users have shared content which relates to others, there is a higher potential risk for these third persons who have not necessarily had a say in the sharing of these data. However, it is important to take into account the type of data in question, and the way in which these data will be used in the research project, when assessing the risk to data subjects of processing social media data based on the private/public nature of these data.

3.1.7 Measures to safeguard the rights and freedoms of data subjects

Due to these issues discussed above, and also as part of good research practice, a data management plan should be made at the planning stages of the project. The researcher should decide how the contact data shall be protected, stored and used. The following is a list of suggested measures to safeguard the rights and freedoms of data subjects:

- if it is not possible to provide information to individuals, make information about the project publicly available
- only collect data which is necessary for the purposes of the project and continuously delete unnecessary data
- reduce the extent to which the collected data is identifiable as soon as possible after data collection, including deleting/replacing unique online identifiers
- implement measures to secure the data at each stage of processing
- limit and control access to the collected data
- focus on groups of social media users during analysis rather than identifiable individuals

3.1.8 Intellectual property

The social media platform can own IP rights to the database from which the data are harvested. The platform can have terms and conditions that restrict such harvesting. The researcher needs to be aware that performing the data analysis via computer usually involves making copies of analysed content, which may be copyrighted. Even making temporary copies may be considered as the “use” of copyrighted material and a license from the rightsholder may be needed.

Social media data and databases are packed with the copyrighted content on several levels:

First level – individual contributions: The copyright protection does not have any restrictions as to the minimal required size of a work. Even a short tweet, Facebook post or profile update is protected as intellectual property. In this context it is worth mentioning the *Infopaq case (*C-5/08[43]) where the Court of Justice of European Union ruled that even the extracts comprising of 11 words come within the concept of reproduction of a copyrighted work as defined in the Art. 2 of the Infosoc directive[44]. Long term archiving of individual tweets constitutes the use of copyrighted work. Sharing a database which contains individual tweets can be considered as a communication to the public as defined under Article 3 of the Infosoc directive.

Second level – the "source database" i.e. the databases of Twitter, Facebook, etc. are intellectually protected as well under the EU Law. It is disputable whether such large databases of photos, contributions and statuses would acquire "copyright" protection as an intellectual creation. However, in case of dispute, they would most likely be awarded the "sui generis" right protection, because the investments undertaken by the social media companies would arguably meet the definition of "substantive investment" as defined by the Art. 7 of the Database directive. Mining these databases qualifies as an "extraction" under Art. 7 section 2 of the Database directive.

Note that data which has been collected directly from social media users may be eligible for protection under the sui generis database right, hence the researcher can claim Intellectual property to the dataset.

[43] Case C-5/08 ECLI:EU:C:2009:465

[44] This is notably different to US approach, see Moran, C., "How Much is Too Much-Copyright Protection of Short Portions of Text in the United States and European Union after Infopaq International A/S v. Danske Dagblades", 2011

4 Publishing and sharing data collected from social media

Publishing data as evidence is essential, and making it accessible is often a requirement or recommendation from funding agencies and publishers to meet mandates for more open science. Disseminating social media data - from publishing to archiving - faces challenges because the legal and ethical terrains are both dynamic and uncertain.

While GDPR has been in place for over a year, creating some more clarity on the legal front, there are few concluded cases and many unresolved questions.[45] Given these legal uncertainties and gaps (the scope of GDPR is large, but not international), it would be comforting if we could rely on agreed and robust ethical practice. Regrettably, the ethical terrain is no less challenging than the legal one. Halford[46] presents a number of "ethical disruptions" of social media. Social media data differ from traditional research data (e.g. surveys) in that they have almost never been generated for research purposes. As a result, the usual ethical protections (e.g. consent and anonymisation) have not taken place. Moreover, the processing of social media data often extends into disciplines, such as computer science, where there is often less familiarity with handling data from human subjects (beginning with differing definitions of what a human subject is). This cross-disciplinary work can make for path-breaking research, however, finding common frameworks across disciplines with different ethical regimes is demanding.

The issues below are not comprehensive, but rather attempt to address particular ethical and legal problems that arise when disseminating social media data, by publishing, sharing, or archiving.

4.1 Ethical issues

4.1.1 Informed consent and unconsented data

Informed consent is, and remains, the gold standard for conducting ethical research. It has been a respected ethical research practice for decades in many countries and when done properly, embodies respect for participants' autonomy and can produce informed and voluntary participation. However, when applied to social media research, two issues arise. First, much social media is done on a large scale, sometimes involving millions of data subjects. At this scale, the standard procedures involving consent forms backed up by additional information, are not likely to be feasible. The second concerns the meaning of consent, and what actions constitute consent. Facebook's Privacy Statement has been rated as more difficult than Kant's *Critique of Pure Reason*.[47] Consent for uses of data is part of most platforms' Terms of Service, but users typically do not read these, and may not understand the implications for data reuse even if they do.[48] This is the case even for the primary uses of the data, and the ethical ambiguity of such consent is only compounded when publishing and sharing data.

[45] http://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeeting&meetingId=15670

[46] Halford, S., "Ethical Disruptions of Social Media Data, in The Ethics of Online Research", 2018

[47] Litman-Navarro, K., "We Read 150 Privacy Policies. They Were an Incomprehensible Disaster", 2019

[48] Fiesler, C., and Proferes, N., "Participant perceptions of Twitter research ethics", 2018

That said, not all social media research is large-scale, and where technically feasible, consent is possible. In one example, Gray undertook a qualitative critical discourse analysis of misogynistic content found on Twitter and successfully obtained consent, using Twitter itself, to publish selected tweets.[49] Gray had to overcome numerous challenges such as reluctance of the relevant REC to permit using Twitter to obtain consent, in part due to fears that Gray would be at risk from tweeters of hate speech. In addition, contacting participants was problematic. Twitter does not prioritize discrete (narrowcast) communication, and only allows direct private messaging between accounts that mutually follow each other. Establishing a mutual following was not practical for a number of reasons: a requirement of ethical approval was use of an anonymized profile to contact participants, and users would be disinclined to follow an anonymous profile with very few followers and little content. This was further complicated by the fact that narrow time constraints meant building a public profile and contacting users to ask for private contact was impractical.

Gray relied on direct contact via tweeting to obtain consent. This consisted of replying to the specific tweets he wanted to feature in his research with a tweet containing a very brief introduction and request for consent. A highly problematic consequence of this situation is that it potentially allowed participants to contact each other. Ironically, Gray argues that given the highly antagonistic speech produced by some participants, this method potentially produced even greater risk of participant harm than publishing without consent, but was still seen as preferable to treating initial publication as consent.

Consent has also proved feasible, though not unproblematic, when social media data is used to augment conventional social surveys. Sloan et al. (2019)[50] conducted research linking Twitter accounts to three major studies in the UK: the 2015 British Social Attitudes, the July 2017 wave of the NatCen Panel, and the 10th wave of the Understanding Society Innovation Panel. A question asking respondents for permission to link their survey responses to their Twitter account was done in three stages, first asking if the respondent had a Twitter account, then for permission to link and for the Twitter username. A major challenge was to balance the requirement for the consent to be informed, yet not overwhelm respondents with details, especially if the respondent did not want to be bothered. Issues of archiving have not been resolved, but have been taken into account in the planning of the research and several options will be considered including secure facilities (onsite and remote), researcher accreditation, and limiting linkage to a data controller.

While these cases demonstrate examples in which consent has been feasible, it remains essential to address research without consent, a reality for much social media data. Again, it is useful to recall that not all is new with social media. Research without consent, including covert research, is a small, but legitimate, part of social science practice. It can be justified when no other approach will work, if the research will be addressing highly significant questions, where a clear public benefit is expected, and there should be no, or very little, risk to participants.[51]

[49] Bishop, EL., and Gray, D., "Ethical challenges of publishing and sharing social media research data", 2015

[50] Sloan, L., Jessop, C., Al Baghal, T., & Williams, M., "Linking Survey and Twitter Data: Informed Consent, Disclosure, Security, and Archiving", 2019

[51] https://esrc.ukri.org/funding/guidance-for-applicants/research-ethics/frequently-raised-questions/what-if-it-is-not-possible-to-obtain-informed-consent/

When applied specifically to social media, factors that affect the need for consent are how public or private is the original source, the extent of interaction with participants, topic sensitivity, and subject vulnerability[52]. The example below demonstrates the successful archiving of geo-tagged Twitter data, showing how these factors have been negotiated in practice.

Researchers archived a dataset of more than half a billion geotagged tweets collected in the US in two six-month periods in 2014 and 2015.[53] The purpose of this collection of all tweets sent in the US in those time periods was to allow studies of representativeness. The final data collection, a bundle of 21GB, consists of 53 files of tweets per state and county and per day: 48 text files with tweet IDs and aggregated number of hashtags, two files with the Python script used for data collection with the Twitter API and the Python script used for sorting geotags, and three help files. The archiving solution had to honour legal requirements and ethical standards, allow reproducibility, and enable the data to be easily reused to answer new questions. From a legal point of view, to comply with the Twitter Terms of Service, the files do not include the tweet text, user names, timestamps, geocodes or other metadata. Only the tweets IDs were archived. To reuse the data, the tweets can be rehydrated based on their IDs. The scripts that can be used to rehydrate the tweets from their IDs were also archived.

From an ethical point of view, even though only tweet IDs were archived, the researchers still had concerns about sharing data from people who had not explicitly consented, or were not necessarily aware that their Twitter data could be used for research. Therefore, access restrictions were put in place by archiving the dataset *with* access provided for research purposes upon request. Since the geolocation information may pose a particular threat to confidentiality, detailed geoinformation was not included in the datasets, but can still be accessed on Twitter itself once the tweets have been rehydrated. The archiving thus balanced the protection of privacy with a desire to make research reproducible.[54]

4.1.2 Shift from anonymisation to risk minimisation

Some traditional forms of anonymisation used for surveys can be adapted for social media. Derivatives or aggregates of data, e.g. counts of posts, followers, sentiment and other content analysis classifications, could to certain levels be considered anonymous. Masking of content and reduction of information of individual post are some of the methods suggested for anonymization. Structure of networks, if not containing identifying information, and if not collected in the known and narrow location or group, can also be considered unproblematic. More typically, however, there are features of social media data that preclude these approaches to anonymisation deployed for more traditional social science data.

As with consent, sometimes the problem is simply scale. Often, other obstacles are even greater, such as Twitter's requirement that content be published unaltered and with attribution, making it relatively simple to search such content and possibly identify individuals.[55] Equally, it may be undesirable from a research perspective to modify data, as doing so may destroy its

[52] McKee, H.A. and Porter, J.E., "The Ethics of Internet Research: A Rhetorical, Case-Based Process", 2009

[53] Pfeffer, J. and Morstatter, F., "Geotagged Twitter posts from the United States: A tweet collection to investigate representativeness", 2016

[54] Kinder-Kurlanda, K., Weller, K., Moltgen, W., Pfeffer, J., and Morstatter, F., "Archiving information from geotagged tweets to promote reproducibility and comparability in social media research", 2017

[55] Townsend & Wallace, "Social Media Research: A Guide to Ethics", 2016

value for research. More broadly, the debate about anonymisation of *any* data is shifting away from a binary framing (un/anonymised) to a recognition that disclosure risk cannot be eliminated, but rather it must be managed.[56] This reflects a sobering reality that the proliferation of data combined with growth in computing power increases disclosure risks, but in ways difficult to predict with any confidence. Publishing and sharing social media data can still benefit from anonymisation strategies, such as aggregation. But often additional measures will also be needed, such as systematic approaches to risk assessment[57] (e.g. disclosure risk reviews and Data Privacy Impact Assessments) and more reliance on data access controls, as the examples above of archived Twitter data demonstrate.[58]

4.1.3 Public or private or...

A number of researchers take the view that any public data has few, if any, restrictions on its use and reuse.[59] And one highly regarded journal has published an article based on the hacked Ashley Madison data[60]. The essential argument in the public/private debate for disseminating social media is twofold: 1) it assumes that two categories exist, and that they can be clearly differentiated as public and private, and 2) that for data in the public category, any forward (re)use can be justified, because, "the data is already public". However, there is a strong critique of both these arguments. Many ethical guidelines (AoIR) and scholars point out that social media (and other internet spaces) in fact cannot be clearly bifurcated into public and private. This may be based in technical features (if registration is required, but open to anyone, is this public or private?) or in user perceptions.

One consideration, specific to social media, is about social media users' expectation of privacy, which affects both publishing quotes of content, and archiving and disseminating data. As observed, not all social media users would expect the full public use of their data, depending on if the communication was meant for a narrow group or for wider audiences, perhaps even in a closed forum or Facebook circle of friends. More generally, users' views are mixed, making it impossible to claim unequivocally that "public means public"[61].

The second point is more ethically complex. If the data are public, then on what ethical grounds might we not permit further dissemination, either publishing or archiving? Here, the issue of "contextual integrity"[62] arises, saying that, context matters, for example, data posted on Facebook are not "the same" when transferred to an archive, due to the different context, with metadata to enhance findability, and accessible to a different community of users.

---

[56] Elliot, M., Mackey, E., O'Hara, K. and Tudor, C., "The Anonymisation Decision- making Framework", 2016

[57] Hemphill, L., Leonard, SH., and Hedstrom, M., "Developing a Social Media Archive at ICPSR. In Proceedings of Web Archiving and Digital Libraries", 2018; Van den Eynden et al. 2020 (forthcoming *Managing and Sharing Research Data, 2nd ed. Sage)*

[58] As regards data access controls, responsible persons such as "data stewards" may be appointed and can further define restricted data and permit control to make sure data access is limited and available only for legitimate use. Methods could include 'safe rooms' in secure locations or secure remote access, both of which include the application of The Five Safes, a set of principles providing guidance to data owners and researchers on how to handle the data by using: safe data, safe projects, safe people, safe settings, safe outputs. Code sharing can also be an alternative, both for statistical data and social media data.

[59] Kirkegaard, E. and Bjerrekær, J., "The OKCupid dataset: A very large public dataset of dating site user", 2016

[60] Griffin, J.M. et al. (2019) Personal infidelity and professional conduct in 4 settings
Proceedings of the National Academy of Sciences Aug 2019, 116 (33) 16268-16273; DOI: 10.1073/pnas.1905329116

[61] Evans, H., Ginnis,S., Bartlett, J., "A guide to embedding ethics in social media research", 2015

[62] Nissenbaum, H., "Privacy in Context: Technology, Policy, and the Integrity of Social Life", 2009

These points are more easily grounded in an example, the publishing of hate speech Tweets discussed above. It could have been argued that the content was public, indeed, the tweeters clearly were intending to broadcast hate speech widely. However, taking a "reflexive approach"[63] called for further considerations. Gray opted for seeking explicit consent to publish because of the sensitivity of the tweets. Moreover, respect for the views of tweeters (and indirectly, their autonomy) was also a factor: 80% of tweeters surveyed expected to be asked for consent to republish, even for research purposes[64].

4.1.4 Research integrity

Data repositories play a vital role in supporting research integrity by holding data and making them available to others for validation and reproduction, as well as providing new research opportunities. To do so, data must have clear "provenance", its sources and processing need to be known, identified, and documented. The attenuated relationship between data curators and social media data producers, who may not be 'researchers' per se, makes this challenging for a number of reasons. For some genres, often with commercial value, such as Twitter data, there are prohibitive legal restrictions on reproducing data, including providing data to support publications. For a comprehensive treatment of issues of preserving social media, see Thomson (2016)[65]. Compared to traditional formats for social science data, social media datasets may be too large and too dynamic for existing infrastructure, as well as having disclosure risks.[66]

Data archives have begun to address these issues.[67] This section will primarily address central developments within social science data archives and ethical issues in archiving social media data. A further review with examples of shared social media data follows in section 4.3.

The core ethical challenge in this domain is the tension between competing responsibilities of archives to enable reproducibility (an element of research integrity) while also respecting data subjects' rights. The current cases typically involve Twitter data. To comply with Twitter's restriction on republishing content, researchers and archivists are instead sharing only tweet IDs, augmented with documentation and code to enable the tweets to be rehydrated[68]. This procedure has been used for archiving Twitter data at GESIS and at the UK Data Service.

While this marks a major positive step in sharing social media, it does not resolve the issue of reproducibility. In addition to complying with Twitter's conditions, this process also respects the rights of Tweeters to remove their content from further reuse. One problem with this solution is that the datasets are almost certain to not be identical (because of deleted tweets), thus undermining the possibility for reproducibility. The problem is not small. One study shows that persistence rates for ID collections over four years range from 30-80%[69]. For now, no better solution has been found, so this is the current practice.

[63] Williams et al 2017

[64] Williams et al 2017

[65] See Thomson, S., "Preserving Social Media", 2016

[66] Hemphill, L., Hedstrom, M., Leonard, S., "How can we save social media data?", 2019

[67] Please see section 4.4. for results from a survey on social media data conducted among a sample of european data archives.

[68] Hemphill et al. 2019; Kinder-Kurlanda et al. 2017

[69] Zubiaga, A., "A longitudinal assessment of the persistence of twitter datasets", 2018

It is worth posing a further question that has received less attention than the reproducibility problem. First, could it be ethical to retain content, Tweets for example, even against the wishes of a tweeter? Arguments against usually rest on claims that tweeters own their content and respect for autonomy of individuals and their choices. Counter arguments point to some historical data sources that cannot be altered, except on grounds such as factual errors or defamation. The GDPR includes possible exemptions from the right to erasure and the right to object if the processing is necessary for archiving purposes in the public interest, scientific or historical research purposes, or statistical purposes.[70]

The responsibility for enabling data access raises a yet broader question: What is at stake when research crucial to major policy debates is based on social media data with very limited access? Raj Chetty at Stanford University has been granted access to Facebook data to study US inequality in unprecedented detail.[71] There seems little doubt that research by such a prominent scholar (MacArthur Genius award winner) on a visible and controversial topic will inform public debate and policy. Researchers bear responsibility for making data available, but archives also claim to share this duty. If so, do their responsibilities extend to engaging with the platforms (or researchers benefitting from privileged relationships) to attempt to negotiate access to these data?

Facebook and others have responded that their new academic collaboration, Social Science One, solves these problems. However, Bruns[72] argues that a more open solution is needed:

> *...what is necessary is an acknowledgment by the platforms that researchers have a legitimate need to share their data with each other[73], in a controlled and ethical way, and that the best way to do so is openly and transparently through the carefully controlled data repositories that already serve other academic fields, from genomics to econometrics. Rather than preventing the use of such safe, managed facilities for data sharing through their terms of service, platforms should work with university researchers to determine meaningful, workable approaches to sharing data that protect both the privacy of users and the integrity of the data.*

If nothing else, this issue is an important reminder that more is at stake than consent and disclosure risks, substantial as they are, but fundamental questions remain unresolved about access to data that underpin vital public policy debates.

4.2 Legal issues

4.2.1 IPR protection

Aggregation of social media data is strongly dependent on text and data mining of individual contributions (which may be protected as copyrighted works) aggregated from large databases (which are usually protected by both copyright and sui generis database rights). For the data repositories, copyright challenges encompass both the issues of copyright in

[70] See Article 17 (3)(d) and Article 21 (6) of the GDPR

[71] Scola, N., "Facebook's next project: American inequality", 2018

[72] Bruns, A., "Internet Policy Review", 2018

[73] Weller, K., and Kinder-Kurlanda, K., "A Manifesto of Data Sharing in Social Media Research", 2016

deposited materials and the need to manage any collaborative platforms which support the upload of content by researchers.

At the stage of archiving it is difficult for a repository to contact individual users and database holders to confirm that the data were collected with proper authorisation (license) in the previous phases of research. If proof of proper authorisation cannot be obtained, the archive may rely on the exceptions from copyright for long term archiving for research purposes. These statutory exceptions are allowed by EU law and many member states have stipulated such exceptions in their national laws.

Long-term archiving of large datasets harvested from social networks can surprisingly also be qualified as "extraction of sui generis database", since the extraction is defined as a "permanent or temporary transfer of all or a substantial part of the contents of a database to another medium by any means or in any form". Archiving of tweets and contributions can be made available to the public or shared with research institutions only at risk of being qualified as "re-utilization" under Art. 7 section 2 of the Database directive. "Re-utilization" has to be authorized by the sui-generis right owner. It is important to note that the statutory exemptions from the copyright and the sui generis database rights and copyright, usually only cover long-term archiving and do not allow research institutions to share copyrighted and protected content online.

4.2.2 Privacy considerations

Long-term archiving of personal data is also considered to be "processing of personal data", and all the principles of data processing have to be respected (i.e. lawfulness, transparency, purpose limitation, data minimisation, accuracy, storage limitation, integrity and confidentiality[74]). There is a myth however, that GDPR requests all long-term data to be pseudonymised and anonymised. These techniques are mentioned as possible instruments to protect the rights of the data subjects. Still, since social media companies may store data and metadata for long periods and much of this data is searchable, anonymization for secondary use can be challenging. If anonymization or pseudonymization means that the research data lose a significant value for further research, the archive may also consider keeping the data in "identifiable form" and adopt other safeguards to protect the privacy of research subjects (such as restrictions on who can access the datasets, approval of any subsequent use of the data by an ethical board etc.).

4.2.3 Social media platforms' terms and conditions

Parallel to privacy considerations and IPR is the question of platforms' terms and conditions. Publishing individual tweets on a social media platform does not mean that further publishing and sharing is allowed. Social media data is hosted by social media sites that have different privacy policies, terms of service, and developer agreements[75], which influence how social media data can be stored and shared. The terms of service is one important reason why currently, when detailed content of the posts and background metadata are considered important property, only twitter IDs for example may be stored in a data archive, and a code of script offered for 'rehydration'.

[74] For further explanations on how the GDPR principles relate to the archived research data see Koščík, "The Impact of General Data Protection Regulation on the grey literature." 2017.

[75] Day Thomson 2016

Kinder-Kurlanda et al.[76] note that '*Sharing legally and ethically also means to follow the changes and updates in terms of services and policies and to participate in negotiations about data sharing for the sake of reproducibility with platform providers*'[77]. In this regard, data repositories may work towards harmonised and equivalent policies regarding IPR and other obstacles for data sharing that follow from the private and commercial character of social network data. Further, government and other institutions may also provide agreements to enhance legal interoperability in licencing and data access[78]. Such approaches could reduce the need for individual bilateral agreements between (commercial) access providers and users.

4.3 Sharing of social media data

Although various research projects are currently using social media data, there are few datasets available for secondary analysis and replication. The amount of data available, e.g. through streaming APIs is limited, and if data is available then purchasing this data might be both expensive and entail restrictive terms and conditions. This situation goes against current requirements or recommendations for open access and the FAIR-principles (a set of guiding principles to make data Findable, Accessible, Interoperable, and Reusable) for reuse of research data[79].

Sharing social media data has two main advantages:

- *Reproducibility:* Even if sharing content from social media may be prohibited, sharing code and workflow that enables replication of data collection and analysis can increase reproducibility [80]
- *Re-usability:* Federating data from shared resources may enhance research opportunities[81]. Sharing data also promotes equal opportunities to access data, may increase quality of research, and lessen efforts in collecting, processing and cleaning[82].

Whilst archiving and sharing social media data can fulfil one or more of the requirements of FAIR data, the specific data management, curation and sharing regime depends on the type of data, on whether privacy concerns exist, on the purpose of use, on whether the content is sensitive, and on issues related to copyright and terms of use.

[76] 2017

[77] Kinder-Kurlanda et al. 2017

[78] 'Top-down harmonization through "hard" law, such as multilateral treaties or executive agreements, or national legislation or administrative regulation, can work in some contexts and can be extremely useful as a broad harmonization tool (Research Data Alliance, Legal Interoperability of Research Data: Principles and Implementation Guidelines, 2016)

[79] Final Report and Action Plan from the European Commission Expert Group *on FAIR* Data, "Turning FAIR into reality", 2018.

[80] Hemphill et al. 2018

[81] Hemphill et al. 2018

[82] Weller, K., Kinder-Kurlanda, K. "Uncovering the Challenges in Collection, Sharing and Documentation: The Hidden Data of Social Media Research?", 2015

The examples below, based on the efforts of Mannheimer et al.[83] (in relation to qualitative data) and Weller and Kinder-Kurlanda[84] (in relation to social media data), distinguish between a sample of modes for sharing data:

- Informal sharing, including sharing 'upon request'. This approach often lacks the possibility to search for data and sufficient documentation of the data.[85] Extended options include sharing on websites. Still, persistence of access to data may be uncertain since, for example, web pages could disappear.
- More persistence can be expected for access to thematic collections of datasets. One example is the CrisisLex, a repository of crisis-related social media data and tools[86]. A recent example of a systematically collected and curated thematic dataset is the "#metoo Digital Media Collection"[87], a digital project of the Schlesinger Library on the History of Women in America that (it has been announced) will be available for research in late 2019.
- A growing number of available social media datasets can be found at different general repositories such as Zenodo[88]. Sharing data through general repositories often rely on self-archiving principles that do not always offer curation and preservation independently of the data depositor. These datasets may also have unknown assessment of important properties or problems associated with secondary use. The OkCupid data release on the Open Science Framework, for example, caused a lot of concern because persons could be recognised.[89] Similarly, the "Tastes, Ties, and Time"[90] dataset published on Harvard's Dataverse has been removed due to privacy concerns. Mannheimer and Hull[91] also show the transparency problem of data deposited in a general repository as they may lack detailed documentation of the data.[92]
- Another example is sharing through conference websites such as the International Conference on Web and Social Media (ICWSM)[93], which allow the sharing of datasets along with papers.[94] One case from the last conference was Brena et al. 2019a[95], where the data referred to is available at the Harvard Dataverse[96]. The code is published on GitHub, and the data can be found through the data repository or through a linked article. Data is persistently identified, and thus citable and permanently

[83] Mannheimer, S., Pienta, A., Kirilova, D., Elman, C., & Wutich, A., "Qualitative Data Sharing: Data Repositories and Academic Libraries as Key Partners in Addressing Challenges", 2019

[84] Weller and Kinder-Kurlanda, "A Manifesto of Data Sharing in Social Media Research", 2016

[85] Weller and Kinder-Kurlanda, 2016

[86] http://crisislex.org/

[87] https://www.schlesinger-metooproject-radcliffe.org/collection

[88] https://zenodo.org/

[89] Hackett, R., "Researchers Caused an Uproar By Publishing Data From 70,000 OkCupid User", 2016

[90] https://dataverse.harvard.edu/dataverse/t3

[91] Gargiulo, F., Bindi, J., and Apolloni, A., "The topology of a discussion: the #occupy case", 2015a and Gargiulo, F., Bindi, J., & Apolloni, A., "The topology of a discussion: the #occupy case [dataset]", 2015b

[92] Mannheimer, S., and Hull, E.A, "Sharing selves. Developing an ethical framework for curating social media data", 2017

[93] See https://www.icwsm.org/2019/program/

[94] Weller and Kinder-Kurlanda, 2016

[95] Brena, G., Brambilla, M., Ceri, S., Di Giovanni, M., Pierri, F., Ramponi, G., "News Sharing Users Behaviour on Twitter: A Comprehensive Data Collection of News Articles and Social Interactions", 2019a

[96] Brena, G., Brambilla, M., Ceri, S., Di Giovanni, M., Pierri, F., Ramponi, G., "News Sharing User Behaviour on Twitter: A Comprehensive Data Collection of News Articles and Social Interactions", 2019b

accessible. Users can check and reuse the code which further enhances the potential for reuse.

- Domain and disciplinary repositories[97] will undertake the curation and long-term preservation of research data. These repositories may suggest different access regimes (in case data cannot be anonymised), inform on size and accepted formats, provide data in common formats, enhance metadata quality, and check for additional documentation to support the discovery and understandability of the data. As regards social media data, one example is a pilot project that archived a Twitter dataset related to the Federal Elections in Germany 2013 in the German Data Archive for the Social Sciences[9899]. The data includes the Tweet-ID and an ID identifying the candidates. The list can be used to retrieve the original tweets, which were posted between June and December 2013. Another example is the "*UK-EU referendum Twitter data*" at the UK Data Archive[100]. This data collection consists of Tweet IDs collected on the UK-EU referendum between September 2015 and August 2016, and was collected via the Twitter API.
- There are also cases of social media data related to research projects appearing in the self-archiving section of some data archives.[101] Some CLARIN ERIC national service providers are also active in collecting, annotating, and sharing social media content of particular language communities.[102]
- Some projects, which collect and archive social media content, involve national archives that try to apply long-established standards of appraisal, selection, preservation, and access, to the challenges of new social media.[103][104] As an example, in the failed attempt of the Library of Congress to preserve the whole twitter collection, the Australian TrISMA – Tracking Infrastructure for Social Media Analysis project[105] makes both current and historical data from Australian Twitter accounts available to accredited users.

4.4 Survey on social media data in European data archives

In an effort to get a further overview of the contemporary conditions of archived social media data, the task group also carried out a survey among CESSDA archives. The survey concentrated on archives connected to CESSDA (both members of CESSDA ERIC and partner archives), using the CESSDA contact sheet to reach the archives through email. Request for participation was sent to 28 data archives on 17th of June 2019. The following questions were asked:

1. *Does your archive store/archive any social media (e.g. Facebook, Twitter, YouTube, Instagram, Tumblr or Reddit) data? By that, we mean both social media content itself and various kinds of metadata (such as tweet IDs)*

[97] For the social sciences these for example include the CESSDA archives.

[98] Kaczmirek, L., Mayr, P., "German Bundestag Elections 2013: Twitter usage by electoral candidates", 2015

[99] Weller and Kinder-Kurlanda, 2016

[100] Cram, Laura and Llewellyn, Clare (2017) UK-EU referendum Twitter data. [Data Collection]. Colchester, Essex: UK Data Archive. 10.5255/UKDA-SN-852513

[101] Chukwuemeka , D. and Abdul, A., "The UK 2015 General Election, Twitter data [Data Collection], 2017

[102] Ljubešić, N., Erjavec, T. and Fišer, D., "Twitter corpus Janes-Tweet 1.0", 2017

[103] Oliveira, D., "Social Media as Records in the Context of the Olympic Public Authority (APO) and their Preservation: A Case Study", 2017

[104] Green, K., and Niven, K., "Archiving Social Media: Mesolithic Online Resources (Mesolithic Miscellany and Mesolithic Research Forum)", 2014

[105] https://trisma.org/

2. *If you do, can you tell us more about your social media data (e.g. from which social media does your archive have data, under which conditions are you able to share the data with users etc.). Moreover, we would be delighted if you could send us the links to your social media data.*
3. *If you do store social media data, have you come across any kinds of issues while storing/archiving these data? If you do not store such data, please, tell us what kinds of problems you think might occur when you would be storing/archiving such data and also why your archive does not contain such data. Is there any particular reason (e.g. researchers are not interested in sharing them, no place to store them, ethical issues etc.)?*

In one month, 17 responses were obtained (61 % response rate). A large majority of the respondents answered that their institution does not archive/store social media data at the moment. Two important exceptions are GESIS – Leibniz Institute for the Social Sciences and the UK Data Archive. GESIS currently holds approximately 10 data collections of social media that have been used in research contexts, including Twitter activities by political actors[106], geo-coded Twitter data[107], and specific subsets of Wikipedia. In addition, some documentation and scripts to facilitate rehydration of tweet IDs have been archived. The UK Data Archive has about 20 collections of Twitter data that have been deposited by researchers.[108] UKDA typically asks to deposit the tweet ID and timestamp, and not the actual content of tweet. Hashtags or topics are also included in some cases. Moreover, a qualitative data set containing messages from Suunta E-Guidance Service is archived in the Finnish Social Science Data Archive (FSD).[109] Also, the dataset "Political Jokes in Post-socialist Slovenia, 2017", which consists of Facebook posts on public Facebook pages, is stored at the Slovenian Social Science Data Archives (ADP).[110]

As regards expectations and plans for the future, some of the respondents mentioned plans for future work with social media data in their archive, and see this as an important area. However, some examples of pilot projects are already being carried out or under preparation:

- AUSSDA (The Austrian Social Science Data Archive) is planning on concentrating more on big data generally and collaborates with a partner in political science/communication science on a pilot project in social media data archiving.
- SoDaNet (Greek research infrastructure for the social sciences) is participating in a project with partners from other fields of social sciences and humanities, where the focus is on social media data, among other topics.
- GESIS is collaborating with researchers to produce a metadata standard for social media data. Further, GESIS is collecting a combined survey, social media and web tracking data to develop procedures for documenting and sharing these linked data. GESIS will also work on piloting a system to create an integrated workflow from data collection (via API) to the production of an "archive-ready" data set.

[106] Stier, S., "Elite actors in the U.S. political Twittersphere", 2016

[107] Geotagged Twitter posts from the United States: A tweet collection to investigate representativeness, http://dx.doi.org/10.7802/1166

[108] Devine, F., Savage, M., "BBC Great British Class Survey, 2011-2013 [data collection]", 2015

[109] Gretschel, A.(Finnish Youth Research Society) and Finnish Youth Research Society, "Career and Study Counselling Provided in Finnish Suunta E-Guidance Service 2013 [dataset]", 2016

[110] Mlačnik, P. and Stanković, P., "Political Jokes in Post-socialist Slovenia", 2017

It seems like partners from the research sector or other fields closely related to data archiving, are an important factor in the development of pilot projects in the data archives.

On the other hand, many respondents from the individual data archives indicated some challenges connected with (potential) social media data archiving at their institution. The results can be summarized as follows:

A. Legal issues were mentioned most frequently as an obstacle for archiving social media data. This was sometimes specified with:
    - Complexity in obtaining consent for data sharing, and with identifying the data ownership.
    - Platforms' terms of service which limit the possibilities for archiving. Moreover, ToS are frequently changing, and the platforms are usually large companies that could be hard to reach for the individual archives - "*Negotiating with big commercial social media platforms for any kind of agreements enabling data archiving ought to be done in CESSDA level*".
    - GDPR was directly mentioned by only one archive, but "privacy issues" (sensitive data, difficulties in getting consent, complicated anonymization) were included in more responses. Still, it seems that problems related to personal information are not perceived as much as a limiting factor as copyright issues and platforms` ToS.

B. Technical issues related to social media data storage were mentioned too. They included:
    - The complex structure of social media data which is challenging for archives which are used to more "traditional data formats".
    - A missing metadata standard for social media data.
    - Complicated anonymization (both technically and legally)
    - This can be further complicated by the fact that researchers (and potential users of archived data) are not fully familiar with methodologies and procedures for dealing with such kinds of data.

C. Ethical issues were mentioned, maybe surprisingly, less frequently compared to the legal issues. And even when mentioned, it was quite general – in the sense "we know there are ethical issues as well," without further specification.

Definitions

Term	Abbr.	Explanation
Anonymous data		Data that cannot identify individuals in the data set, neither directly through name or social security number, indirectly through background variables, nor through a list of names or through an encryption formula and code/scrambling key.[111]
Consent		Consent of the data subject (legal consent) means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her.[112]
Data access		The activity by which a researcher is given access to data.[113]
Data controller		'Controller' means the natural or legal person, public authority, agency or other body which alone or jointly with others, determines the purposes and means of the processing of personal data.[114]
Data curation		The activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purposes and available for discovery and re-use. A more formal definition is: Digital curation is all about maintaining and adding value to a trusted body of digital information for future and current use: specifically, the active management and appraisal of data over the entire life cycle.[115]
Direct identifier		A person will be directly identifiable through name, social security number or other uniquely personal characteristics.
Data management plan	DMP	A DMP is a formal document that provides a framework for how to handle the data material during and after the research project.[116]
Data processor		Processor means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller.[117]
FAIR principles	FAIR	A set of guiding principles to make data Findable, Accessible, Interoperable, and Reusable.[118]

[111] http://www.nsd.uib.no/personvernombud/en/help/vocabulary.html

[112] GDPR Article 4

[113] OECD 2016: 41

[114] GDPR Article 4

[115] OECD 2016: 41

[116] https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/1.-Plan/Benefits-of-data-management

[117] GDPR Article 4

[118] https://www.force11.org/group/fairgroup/fairprinciples

General Data Protection Regulation	GDPR	The GDPR (Regulation (EU) 2016/679) is a regulation by which the European Parliament, the European Council and the European Commission intend to strengthen and unify data protection for individuals within the European Union (EU). It came into force in May 2018.
Indirect identifiers		A person will be indirectly identifiable if it is possible to recognize the person through background information such as place of residence or institutional affiliation, combined with data on age, gender, occupation, diagnosis, etc.[119]
Long-term preservation		In data management, data preservation is the process of maintaining access to data so that it can still be found, understood and used in the future.[120]
Metadata		Provides information on data and the processes of producing and using data. Metadata are data which are needed for proper reproduction and use of the data[121]
Open access		Open access can be defined as the practice of providing online access to scientific information that is free of charge to the reader. In the context of research and development, open access typically focuses on access to 'scientific information' or 'research results'.[122]
Personal data		Any information relating to an identified or identifiable person. A person may be identified by name, images/video, email, IP-address or a number referring to a list of names, through photo/video of recognizable faces, or through a combination of background information.[123]
Research data		May be defined as information relevant to, or of interest to, researchers either as inputs into or outputs from research. They are research materials resulting from primary data collection or generation, or derived from existing sources intended to be analysed in the research project.[124]
Synergies for Europe's Research Infrastructures in the Social Sciences	SERISS	A Horizon 2020 project focused around three key themes – key challenges facing cross-national data collection, breaking down barriers between research infrastructures, and embracing the future of social sciences. It addresses issues relating to survey design and data collection, data management and curation from a collaborative, cross-national perspective. The project will better equip Europe's social science data infrastructures to play a major role in addressing the key societal challenges facing Europe today and help

[119] http://www.nsd.uib.no/personvernombud/en/help/vocabulary.html

[120] https://www2.le.ac.uk/services/research-data/keep-data/lterm-pres

[121] OECD 2016: 42

[122] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm

[123] http://www.nsd.uib.no/personvernombud/en/help/vocabulary.html

[124] OECD 2016: 43

		ensure that national and European policymaking is built on a solid base of the highest-quality socio-economic evidence. See www.seriss.eu.
Social media		Internet based platforms and technologies that permit users' interaction and/or facilitate the creation and exchange of user generated content. Examples include online forums, social networks (e.g. Facebook), video/photo sharing (e.g. YouTube).[125]
Social media data		Social media data refers to the information (photos, comments, etc.) that users generate or share while engaged in or with social media. It often includes personally identifiable data.[126]
Social media research		Social media research encompasses all research where social media data, the information that users generate or share within social media platforms, is used for research purposes either by itself or together with information from other sources. Examples include monitoring or crawling social media platforms and ethnographic research like observing online social behaviour.[127]
Trusted digital repository	TDR	A trusted digital repository is one whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future.[128]

[125] ESOMAR Guideline on Social Media Research (2011: 1-2)

[126] ESOMAR Guideline on Social Media Research (2011: 1-2)

[127] ESOMAR Guideline on Social Media Research (2011: 2)

[128] https://www.oclc.org/research/activities/trustedrep.html

References

Al Baghal, T., Sloan, L., Jessop, C., Williams, M. L., & Burnap, P. (2019). Linking Twitter and Survey Data: The Impact of Survey Mode and Demographics on Consent Rates Across Three UK Studies. *Social Science Computer Review*. https://doi.org/10.1177/0894439319828011

Association of Internet Researchers (2019) Ethics. Retrieved from https://aoir.org/ethics/

Beard-Knowland, T. and Ginnis, S. (2019) Practical Ethics in Social Media Research: The greatest good for the greatest number? Retrieved from https://www.schlesinger-metooproject-radcliffe.org/collection

Bishop, EL. and Gray, D. (2017) Ethical challenges of publishing and sharing social media research data. In Woodfield K (ed) The Ethics of Online Research. Bingley, UK: Emerald Publishing, pp.159–187.

Brena, G., Brambilla, M., Ceri, S., Di Giovanni, M., Pierri, F., Ramponi, G. (2019a) News Sharing Users Behaviour on Twitter: A Comprehensive Data Collection of News Articles and Social Interactions. Retrieved from https://aaai.org/ojs/index.php/ICWSM/article/view/3256/3124.

Brena, G., Brambilla, M., Ceri, S., Di Giovanni, M., Pierri, F., Ramponi, G. (2019b) News Sharing User Behaviour on Twitter: A Comprehensive Data Collection of News Articles and Social Interactions. Retrieved from https://doi.org/10.7910/DVN/5XRZLH

Bruns, A. (2018) Internet Policy Review. Retrieved from https://policyreview.info/articles/news/facebook-shuts-gate-after-horse-has-bolted-and-hurts-real-research-process/786

CESSDA Training Working Group (2017-2018) CESSDA Data Management Expert Guide. Bergen, Norway:
CESSDA ERIC. Retrieved from https://www.cessda.eu/DMGuide

Chukwuemeka, D. and Abdul, A. (2017) The UK 2015 General Election, Twitter data. [Data Collection]. Colchester, Essex: UK Data Archive. Retrieved from 10.5255/UKDA-SN-852772

Davies, L. (2019) Social media data in research: a review of the current landscape. Digital Humanities MA/MSc student, UCL. Retrieved from: https://ocean.sagepub.com/blog/social-media-data-in-research-a-review-of-the-current-landscape?priorityCode=9P0091A&utm_source=Adestra&utm_medium=email&utm_content=9P0091A&utm_campaign=not+tracked&utm_term

Devine, F., Savage, M. (2015) BBC Great British Class Survey, 2011-2013. [data collection]. UK Data Service. SN: 7616, http://doi.org/10.5255/UKDA-SN-7616-1

Digital Preservation Coalition 2016 and Sara Day Thomson (2016) Preserving Social Media. DOI: http://dx.doi.org/10.7207/twr16-01

Edwards, A.et al. 2013. Digital social research, social media and the sociological imagination: Surrogacy, augmentation and re-orientation. International Journal of Social Research Methodology 16(3), pp. 245-260. (10.1080/13645579.2013.774185)

Elliot, M., Mackey, E., O'Hara, K. and Tudor, C. (2016) *The Anonymisation Decision- making Framework*, UK Anonymisation Network, Manchester. Retrieved from https://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making- Framework.pdf.

ESRC (2019) What if it is not possible to obtain informed consent? Retrieved from https://esrc.ukri.org/funding/guidance-for-applicants/research-ethics/frequently-raised-questions/what-if-it-is-not-possible-to-obtain-informed-consent/

ESOMAR (2011) Esomar Guideline on Social Media Research. Retrieved from https://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ESOMAR-Guideline-on-Social-Media-Research.pdf

ESOMAR (2019) Esomar Guideline on Social Media Research. Code & guidelines. Retrieved from https://www.esomar.org/what-we-do/code-guidelines

ESOMAR (2019) ESOMAR/GRBN Online Research Guideline. Retrieved from https://www.esomar.org/what-we-do/code-guidelines/esomargrbn-online-research-guideline

ESOMAR (2019) ESOMAR Data Protection Checklist. Retrieved from https://www.esomar.org/what-we-do/code-guidelines/esomar-data-protection-checklist

European Commission, Directorate-General for Research and Innovation (2016) ANNEX 1 (part A) Research and Innovation action. NUMBER — 654221 — SERISS Ref. Ares(2016)3276367, 08/07/2016.

European Commission, Research & Innovation (2018) Final Report and Action Plan from the European Commission Expert Group on FAIR Data. Turning FAIR into reality. Retrieved from https://ec.europa.eu/info/sites/.../turning_fair_into_reality_1.pdf

European Commission, Research & Innovation (2018) Open access & Data management. Retrieved from http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm

European Court of Justice, Case C‑5/08 ,Judgment of the Court (Fourth Chamber) of 16 July 2009. Infopaq International A/S v Danske Dagblades Forening. CLI:EU:C:2009:465. Retrieved from http://curia.europa.eu/juris/document/document.jsf?text=&docid=72482&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=3681045

Evans, H., Ginnis, S., Bartlett, J., (2015) in Halford; A guide to embedding ethics in social media research.

Fiesler C, Proferes N. (2018) Participant perceptions of Twitter research ethics. Social Media + Society. Retrieved from: https://doi.org/10.1177/2056305118763366

Force11 (2018) FAIR Data Principles. Retrieved from https://www.force11.org/group/fairgroup/fairprinciples

Gargiulo, F., Bindi, J., & Apolloni, A. (2015a) The topology of a discussion: the #occupy case. PLOS ONE 10(9): e0137191. http://dx.doi.org/10.1371/journal.pone.0137191

Gargiulo, F., Bindi, J., & Apolloni, A. (2015b) The topology of a discussion: the #occupy case [dataset]. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.q1h04

Green, K., and Niven, K., (2014) Archiving Social Media: Mesolithic Online Resources (Mesolithic Miscellany and Mesolithic Research Forum). Retrieved from https://historicengland.org.uk/content/docs/research/social-media-case-study-archiving-social-mediapdf/

Gretschel, A. (Finnish Youth Research Society) and Finnish Youth Research Society: Career and Study Counselling Provided in Finnish Suunta E-Guidance Service 2013 [dataset]. Version 1.0 (2016-11-18). Finnish Social Science Data Archive [distributor]. http://urn.fi/urn:nbn:fi:fsd:T-FSD3087

Hackett, R., (2016) Researchers Caused an Uproar By Publishing Data From 70,000 OkCupid User. Fortune, May 18, 2016. Retrieved from http://fortune.com/2016/05/18/okcupid-data-research/

Hagen, S., Straume, Ø., Bishop, E., Vavra, M., Koščík, M., Štebe, J., Ryan, L., Rød, L.M., Payne, E., Somby, A., L'Hours, H., Emery, T., Cizek, T., Krejci J., Høgetveit Myhren, M., Jessop, C. (2019)

Guidelines on the use of social media data in survey research. Deliverable 6.2 of the SERISS project funded under the European Union's Horizon 2020 research and innovation programme GA No: 654221. Available at: wwwww.seriss.eu/resources/deliverables www.seriss.eu/resources/deliverables

Halford, Susan (2018) Ethical Disruptions of Social Media Data, in The Ethics of Online Research (Woodfield, K.ed.) Emerald.

Hemphill, L., Leonard, SH., and Hedstrom, M., (2018) Developing a Social Media Archive at ICPSR. In Proceedings of Web Archiving and Digital Libraries (WADL'18). ACM,New York,NY,USA. DOI: https://doi.org/10.475/123_4

Hemphill, L., Hedstrom, M., Leonard, S. (2019) How can we save social media data? Retrieved from http://hdl.handle.net/2027.42/149013

Hootsuite Inc. (2019) 130+ Social Media Statistics that Matter to Marketers in 2019. Retrieved from: https://blog.hootsuite.com/social-media-statistics-for-social-media-managers/

Information Commissioner's Office (2018) Consent. Retrieved from https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/consent/

Kaczmirek, L., Mayr, P. (2015) German Bundestag Elections 2013: Twitter usage by electoral candidates. GESIS Data Archive, Cologne. ZA5973 Data file Version 1.0.0, doi:10.4232/1.12319

Kinder-Kurlanda K, Weller K, Zenk-Möltgen W., et al. (2017) Archiving information from geotagged tweets to promote reproducibility and comparability in social media research. Big Data & Society. Epub ahead of print 01 November 2018. DOI: 10.1177/2053951717736336

Kirkegaard, E. O. W. & Bjerre Klær, J. (2016) The OKCupid dataset: A very large public dataset of dating site users. Open Differential Psychology. 10.26775/ODP.2016.11.03.

Kotsios, A., Magnani, M., Rossi, L., Shklovski, I., Vega, D. (2019) An Analysis of the Consequences of the General Data Protection Regulation (GDPR) on Social Network Research, arXiv:1903.03196v1

Koščík, M. (2017) The Impact of General Data Protection Regulation on the grey literature." *Grey Journal (TGJ)* 13 (2017): 42-45.

Litman-Navarro, K. (2019) We Read 150 Privacy Policies. They Were an Incomprehensible Disaster. Retreived from https://www.nytimes.com/interactive/2019/06/12/opinion/facebook-google-privacy-policies.html

Ljubešić, N., Erjavec, T. and Fišer, D. (2017) Twitter corpus Janes-Tweet 1.0. Slovenian language resource repository CLARIN.SI, Retreived from http://hdl.handle.net/11356/1142

Mannheimer, S., and Hull, E.A (2017) Sharing Selves: Developing an Ethical Framework for Curating Social Media Data, International Journal of Digital Curation, 2017, Vol. 12, Iss. 2, 196–209.

Mannheimer, S., and Hull, E.A. (2017) Privacy tool is one initiative to help assess the issues. Retrieved from https://privacytools.seas.harvard.edu/.

Mannheimer, S., Pienta, A., Kirilova, D., Elman, C., & Wutich, A. (2019) Qualitative Data Sharing: Data Repositories and Academic Libraries as Key Partners in Addressing Challenges. American Behavioral Scientist, 63(5), 643–664. https://doi.org/10.1177/0002764218784991

McKee, H.A, and Porter, J.E. (2009) The Ethics of Internet Research: A Rhetorical, Case-Based Process. Peter Lang, New York.

Metzler, K., Kim, D. A., Allum, N., & Denman, A. (2016). *Who is doing computational*

social science? Trends in big data research (White paper). London, UK: SAGE Publishing. doi: 10.4135/wp160926.

Mlačnik, P. and Stanković, P., (2017) Political Jokes in Post-socialist Slovenia. Retrieved from https://www.adp.fdv.uni-lj.si/opisi/humor17/

Moran, C. (2011) How Much is Too Much-Copyright Protection of Short Portions of Text in the United States and European Union after Infopaq International A/S v. Danske Dagblades. Washington Journal of Law, Technology and Arts, Vol.6, issue 3, winter 2011.

NESH (2018) A Guide to Internet Research Ethics. Retrieved from https://www.etikkom.no/en/ethical-guidelines-for-research/ethical-guidelines-for-internet-research/

Nissenbaum, H. (2009) Privacy in Context: Technology, Policy, and the Integrity of Social Life. Stanford University Press

NSD Data Protection Services (2018) Information and consent. Retrieved from http://www.nsd.uib.no/personvernombud/en/help/information_consent/index.html

NSD Data Protection Services (2018) Vocabulary. Retrieved from http://www.nsd.uib.no/personvernombud/en/help/vocabulary.html

OECD (2016) Research Ethics and New Forms of Data for Social and Economic Research, *OECD Science, Technology and Industry Policy Papers*, No. 34, OECD Publishing, Paris, https://doi.org/10.1787/5jln7vnpxs32-en

Obar, Jonathan A. & Anne Oeldorf-Hirsch (2018) The biggest lie on the Internet: ignoring the privacy policies and terms of service policies of social networking services, Information, Communication & Society, DOI: 10.1080/1369118X.2018.1486870

Oliveira, C., D. (2017) Social Media as Records in the Context of the Olympic Public Authority (APO) and their Preservation: A Case Study. Retrieved from http://www.pokarh-mb.si/uploaded/datoteke/Radenci/radenci_2017/05_oliveira_2017.pdf

Pfeffer, J. and Morstatter, F. (2016) Geotagged Twitter posts from the United States: A tweet collection to investigate representativeness. [datasets online] GESIS. http://dx.doi.org/10.7802/1166

Playford, C.J., Vernon, G., Connelly, R., and Gray, A. JG (2013) Administrative social science data: The challenge of reproducible research Big Data & Society. Retrieved from: https://doi.org/10.1177/2053951716684143

Research Data Alliance (2016) Legal Interoperability of Research Data: Principles and Implementation Guidelines.

RLG - OCLC (2002) Trusted Digital Repositories: Attributes and Responsibilities. Retrieved from https://www.oclc.org/research/activities/trustedrep.html

Rouvroy, A., and Poullet, Y. (2009) The right to informational self-determination and the value of self-development: Reassessing the importance of privacy for democracy. Reinventing data protection? Springer, Dordrecht, 45-76.

Scola, N. (2018) Facebook's next project: American inequality. Retreived from https://www.politico.com/story/2018/02/19/facebook-inequality-stanford-417093

Sloan, L., Jessop, C., Al Baghal, T., & Williams, M. (2019). Linking Survey and Twitter Data: Informed Consent, Disclosure, Security, and Archiving. Journal of Empirical Research on Human Research Ethics. DOI: https://doi.org/10.1177/1556264619853447

Social Media Research Group (2016) Using social media for social research: An introduction. Government social research.

Statista (2019) Number of social media users worldwide from 2010 to 2021 (in billions). Retrieved from https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

Stier, S. (2016) Elite actors in the U.S. political Twittersphere. Retrieved from http://dx.doi.org/10.7802/1178

Taylor, J., and Pagliari, C. (2017) Mining social media data: How are research sponsors and researchers addressing the ethical challenges? Research Ethics 1–39. DOI: 10.1177/1747016117738559

The Privazy Plan (2018) Article 7, GDPR, Conditions for consent. Retrieved from http://www.privacy-regulation.eu/en/article-7-conditions-for-consent-GDPR.htm

The Privazy Plan (2018) Recital 32, EU GDPR. Retrieved from http://www.privacy-regulation.eu/en/recital-32-GDPR.htm

Townsend, L & Wallace, C (2016) Social Media Research: A Guide to Ethics. University of Aberdeen. Retrieved from www.gla.ac.uk/media/media_487729_en.pdf.

University of Leicester (2018) Long-term preservation. Retrieved from https://www2.le.ac.uk/services/research-data/keep-data/lterm-pres

Visual Capitalist (2019) What Happens in an Internet Minute in 2019? Retrieved from https://www.visualcapitalist.com/what-happens-in-an-internet-minute-in-2019/

Weller, K., Kinder-Kurlanda, K. (2015) Uncovering the Challenges in Collection, Sharing and Documentation: The Hidden Data of Social Media Research?. International AAAI Conference on Web and Social Media, North America. Available at: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10657

Weller, K., and Kinder-Kurlanda, K. (2016) A Manifesto of Data Sharing in Social Media Research. DOI: http://dx.doi.org/10.1145/2908131.2908172

Weller K., and Kinder-Kurlanda K. (2017) To Share or Not to Share? Ethical Challenges in Sharing Social Media-Based Research Data. In: Zimmer M and Kinder-Kurlanda K (eds) Internet Research Ethics for the Social Age: New Challenges, Cases, and Contexts. New York, NY: Peter Lang, pp.115–129.

Westin, A. F. (1970) Privacy and freedom. Atheneum: New York.

Williams, ML, Burnap P. and Sloan L. (2017) Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. Sociology 51(6): 1149–1168.

Zimmer, M. (2010) But the data is already public: on the ethics of research in Facebook. Ethics and Information Technology 12(4): 313–325.

Zubiaga A. (2018) A longitudinal assessment of the persistence of twitter datasets. J Assoc Inf Sci Technol. Retrieved from: http://doi.wiley.com/10.1002/asi.24026

Appendix A: Short guide on legal and ethical issues for the researcher to consider when using social media for research

This task considered the legal and ethics of using social media in research. In legal terms, this Guidance is not legally binding. In terms of research ethics, the subject matter are complex and often issues can only be resolved on a case-by case-basis. Users are encouraged to refer to existing disciplinary codes of practice while being aware that the contents are subject to debate.

A: Data collection direct: research activities directly gathering data from users
B: Data collection indirect: research scraping/harvesting from social media platforms
C: Publishing: publishing data collected from social media
D: Sharing: sharing data collected from social media

Ethical issues

	A. Data collection direct	
	Questions	**Guidance**
Research Ethics approval	Do I need to obtain research ethics approval?	Research involving human subjects often need to be subject to research ethics review (to be obtained before the study commences or if the project changes significantly in the course of research). Requirements for an ethics review vary between countries, institutions and disciplines. Consider the risk of harm to the participant as a result of participation, or of the use of information that identifies them, being included in the research.
Informed consent	Do I need to obtain informed consent from the social media users? What should the informed consent cover? How will the data be accessed? What do I have to think about when linking survey results to social media data?	If information is collected directly from individuals the research will be understood as involving humans and so often requires informed consent. Note that informed consent in research ethics terms differs from consent as a legal basis in the General Data Protection Regulation (GDPR). Depending on the nature of the project and the data collection activities of the researcher, social media users should know what information will be collected, for what purpose, the security arrangements, whether s/he can access and have the right to delete the data gathered. The right not to take part must also be considered. Permission by the owner/manager of the data (e.g. social media databases) may be required. Researchers are advised to check conditions of copyright and the Terms of Service (ToS). Informed consent should be obtained if survey data are to be linked to social media data.
Privacy and confidentiality	Do I need to anonymise the social media data collected?	This will depend on the nature of the research. How privacy and confidentiality is addressed relates to the data security arrangements in place and should be detailed in a data management plan (DMP). The standard techniques relating to safeguarding relate back to traditional methods discussions. Whether social media data are considered 'public' or 'private' by users should be taken into account. This will be especially important in respect of topics that are considered 'sensitive'. The technical accessibility to the data site and how the users treat the information on the site should be considered.

	How can I ensure that assurances of confidentiality and anonymity are meaningful?	If assurances as to the confidentiality and anonymity of data are provided to social media users, this must be realistic and meaningful. Justification of why the research data will not be anonymised/ pseudonymised, if relevant, should be provided if seeking informed consent and confidentiality assurances should take account of the challenges of maintaining confidentiality.
Vulnerability	Are the social media users vulnerable individuals or groups?	In research that involves direct contact with social media users, the general guidance is that some attempt to establish whether the users are vulnerable is warranted. Particular attention must be paid to vulnerable categories of individuals such as children, patients, minorities, people unable to give consent, etc. The inability to determine the mental capacity and to verify age of social media users should be considered and safeguards of their interests specified.
Security	What are the possible risks to the social media user?	The risk of harm to the participants as a result of participation or of the inclusion of information that identifies them being included in the research should be considered. The GDPR focuses on risks to the rights, freedoms and legitimate interests of data subjects. Possible risks can also include security risks when processing personal data, e.g. when it comes to how the data will be stored and who will have access to the data. Depending on the study, textual extracts and social media usernames and handles are searchable text. Assessment as to whether the technique of data collection or method of analysis could result in the identification of individuals should be undertaken before the research commences.
	B. Data collection indirect	
	Questions	**Guidance**
Research Ethics approval	Do I need to obtain research ethics approval?	Research involving human subjects often need to be subject to research ethics review (to be obtained before the study commences or if the projects changes significantly in the course of research). Large scale studies involving harvesting from social media platforms raise the question as to whether the data gathered are to be treated as text (no human input) or as human subjects' data. In general, if there has been human input at some point in the process of creation, then the consequences for individuals and communities should be considered, even if there is no requirement for a formal research ethics review.
	When do I need to obtain ethical approval?	The need for research ethics review will be related to the nature of the research and whether it is categorised as 'low', 'medium' or 'high risk' in terms of the topic under investigation and the possible consequences of the research on the individual, communities and the academic disciplines and/or sectors associated with the researchers.
Informed consent	Do I need to obtain informed consent from the social media users?	It will depend on the research question and nature of data collection and analysis. E.g. if linking survey data to Twitter accounts, informed consent will be necessary. While appreciating that, in general, the data are produced by humans, a pragmatic approach requires that the ability to seek informed consent in large scale data sets should be weighed against its feasibility, the expected benefits of the study and data security measures in place.

	How will the data be accessed?	Large scale social media research data that are identified as public and are subject to platform Terms of Service (ToS). This may mean that the requirement for obtaining informed consent is waived. Consider whether the data are to be treated as in the public domain, for use without restriction, or treated as data from human subjects with corresponding rights (to informed consent et cetera). Whether the users generating social media content consider their posts as public or as private will be one of the factors determining whether informed consent for use of the content is required. Permission by the owner/manager of the data (e.g. social media databases) may be required. Researchers are advised to check conditions of copyright and the platform ToS.
Privacy and confidentiality	Are privacy and confidentiality important? Can privacy be assured by my method of extraction of data?	This depends on whether the social media data are considered public or private. Different social media data may pose risks of different disclosure. Using Tweets as an example, outliers should be considered, e.g. abnormally high number of followers. Cross referencing may also result in identification. Combinations of data may result in identification of individuals. Data security should be considered. See principles for maintaining security.
Vulnerability	Are issues of vulnerability relevant?	In research employing scraping/harvesting data verification is generally not possible, but the researcher should directly address the nature of such issues. Particular attention must be paid to vulnerable categories of individuals such as children, patients, minorities, people unable to give consent, etc. In research ethics terms, attention to the possible consequences for such groups is part of the process of considering the consequences of the research activity. The inability to determine the mental capacity and to verify age of social media users should be explicitly considered and safeguards of their interests specified.
Security	What are the possible risks to the data subject?	The risk of harm to the participants as a result of participation or of the inclusion of information that identifies them being included in the research should be considered. Assessment as to whether the technique of data collection or method of analysis could result in the identification of individuals should be undertaken before the research commences. A DMP should be prepared, covering matters such as how the data will be stored; who will have access to the data; arrangements for anonymization or pseudonymisation of the datasets; this plan should be in compliance with legal requirements.
	C. Publishing	
	Questions	**Guidance**
Research Ethics Approval	Are my publishing plans clearly described in my Ethical Review application?	Publishing intentions should be made clear in the ethics application. It may be possible to provide the name of a social media researcher to your REC if the committee does not have such expertise.

Informed consent	If using consent, do I need obtained consent for publishing (as well as for research)? If consent is not possible, is publication ever permitted?	Yes. It is good practice to have separate consent for research, dissemination, and data sharing. Yes. Key factors to assess are privacy, topic sensitivity, and subject vulnerability. Higher values increase the probability that consent is needed.
Private/public	My data are public; can I now publish without restrictions? How public or private is the source (forum, etc.) from which I collected the data?	Possibly, yes. But ethical considerations remain, even for "public data". It is necessary to assess the "publicness" of the setting. Requirements to register, presence of a moderator, password protection, etc. all suggest some intent toward private communication. Open access, institutional accounts, and broadcast messages all suggest more public intentions.
Vulnerability	Can I publish content from vulnerable participants (e.g. children, elderly)?	Best practice is to seek opt-in consent for publication. This is especially true when publication increases disclosure risk (e.g. content of Tweet is findable and reveals user ID).
Sensitive data or topics	Are the data about sensitive topics (e.g. health, religion, political views, sexuality etc.)? If yes, have I considered if the potential benefits of this research offset the additional risks?	Best practice is to seek opt-in consent for publication. This is especially true when publication increases disclosure risk (e.g. content of Tweet is findable and reveals user ID).
Access restrictions	My publisher insists that I deposit my data, but my data have disclosure risks. What are my options?	Most publishers permit an alternative, a "data access statement" explaining restrictions. Some repositories provide gradations of control that may enable regulated access to data.
If publication/ sharing not possible	What should I do if the platform [Twitter, Facebook, etc.] does not allow me to publish any data? My data simply cannot be published.	Check any Terms and Conditions. Consider how you will handle replication requests.
	D. Sharing	
	Questions	**Guidance**
Research Ethics Approval	Are my data sharing plans clearly described in my Ethical Review application?	Data sharing intentions should be made clear in the ethics application. It may be possible to provide the name of a social media researcher to your REC if the committee does not have such expertise.
Informed consent	If using consent, do I need consent for data sharing? Is broad consent an accepted ethical practice when data may be reused for purposes other than the primary one, or even for unknown purposes? If consent is not possible, is data sharing ever permitted?	Yes. It is good practice to have separate consent for research, dissemination, and data sharing. Broad consent is still used in domains such as biobanking. GDPR (Recital 33) allows for consent even when detailed purposes are not known, subject to "recognized ethical standards". Yes. In addition to topic sensitivity, subject vulnerability and privacy, other factors to consider are: is the research impossible with consent, could the research be done with another method, and are there compensating benefits from doing the research?
Private/public	My data are public; can I now archive and share what I want?	Possibly, but ethical duties may prevent sharing, e.g., if a participant has consented, but the researcher believes she does not fully understand the risks of data sharing. (See legal section for other restrictions)

	What additional factors do I have to consider before sharing data?	When sharing data, it is necessary to consider all factors in the consent section above, and further, to consider possible risks that sharing can make data more visible, and can increase possibilities for linkage and disclosure.
Vulnerability	Can I share or archive content from vulnerable participants (children, elderly, disability)?	Best practice is to seek opt-in consent for sharing. It is necessary to consider all factors in the consent section above and further, to consider possible risks that sharing can make data more visible, and can increase possibilities for linkage and disclosure.
Sensitive data or topics	Can I share or archive data about sensitive topics?	Best practice is to seek opt-in consent for sharing. It is necessary to consider all factors in the consent section above and further, to consider possible risks that sharing can make data more visible, and can increase possibilities for linkage and disclosure.
Access restrictions	I cannot make my data open. Can I still share or archive any data?	Some repositories provide gradations of control that may enable regulated access to data.
If publication/ sharing not possible	What should I do if the platform [Twitter, Facebook, etc.] does not allow me to share any data? My data cannot be archived. Do I need to do anything more?	Check any Terms and Conditions. There are research exceptions. Twitter permits sharing of Tweet IDs. Yes, create a metadata record of your research and data.

Legal issues

	A. Data collection direct	
	Questions	**Guidance**
Terms of service	Do TOS address the issues of research use? Are ToS global or are they localised?	This depends on the platform in use. E.g. Twitter has different rules for US and non-US users. Other services have different conditions for different countries incorporated into one document.
IPR protection	Are the data collected protected by IPR? Can statutory exception be used to harvest data without permission?	The individual records can be protected by copyright. The whole or parts of a database can be protected by sui generis database rights. Some data are not protected and can be harvested at will. EU law has certain statutory exceptions for non-commercial research purposes.
GDPR/privacy	Does the harvesting of data qualify as “processing personal data”? What measures should be taken to protect the privacy?	In many circumstances yes. There are some exceptions in GDPR that permit harvesting without consent. If data contains personal information, make sure to be compliant with the privacy rules. DPIA, good practices, technological measures, pseudonymisation and/or anonymization.
	B. Data collection indirect	
	Questions	**Guidance**
Terms of service	Do TOS address the issues of research use? Are ToS global or are they localised? Will the data be used in different jurisdictions?	This depends on the platform in use. E.g. Twitter has different rules for US and non-US users. Other services have different conditions for different countries incorporated into one document. This report covers rules of using data within EU/EEA. Please note that jurisdictions outside EU/EEA may have different rules for using the data.

IPR protection	Does the compilation of the data create a new form of IP?	Compiling, validating or structuring the data can give rise to a completely new IPR owned by the consortium or its individual members. Adding content to the database usually requires consent from the IPR holder.
GDPR/Privacy	Can the data analysis be qualified as processing personal data?	In many circumstances yes. Again, there are exceptions. If data contains personal information, make sure to be compliant with the privacy rules.
	C. Publishing	
	Questions	**Guidance**
Terms of service	Do I have permission to publish data?	The various platforms differ in how they use their data and how their users' data can be used by third parties. Researchers are advised to check the social media platforms` ToS.
IPR	Have all copyright issues been cleared? Distinguishing between source data and published data. How can I ensure correct citation of data sources?	Consider issues of co-authorship and joint ownership of rights. If the new IPR was created in the phase of research, the decision to publish must not violate the rights of any co-author. The published data are usually a small part of all the data gathered. The fragmental pieces of information are often not protected by IPR. Consider how to cite data sources properly in order to avoid academic misconduct.
GDPR/privacy	Do I have permission to publish personal data?	If participants consent to their data being shared then anonymization may not be required.
	D. Sharing	
	Questions	**Guidance**
Terms of service	Do I have permission to share data with third parties?	The various platforms differ in how they use their data and how their users' data can be used by third parties. There are research exceptions. Researchers are advised to check the social media platforms` ToS.
IPR	Who owns the IPR to the data? Have all copyright issues been cleared? Do any exceptions allow archiving and sharing if the content is considered having cultural heritage value? How and when should I cite data? When is the use of CC licences recommended? Which are the statutory exceptions to archive and share the data even if IPR issues are not clarified?	Researchers are advised to check conditions of copyright. Note that even by sharing one's own dataset the rights of third parties may be violated if the dataset contains IPR of third persons. Agreements and permissions about archiving licenced content can be reached. Data should be considered legitimate, citable products of research. Data should be cited every time any recognizable part of the data is copied. CC may be used for databases. It is recommended in cases when you are certain that the IPR issues in your database is resolved and there is no unauthorized content involved. Copies of works or other subject-matter made in compliance with Art. 3 § 1 of the DSM directive may be retained for the purposes of scientific research, including for the verification of results.
GDPR/privacy	Do I have permission to share data personal data?	If participants consent to their data being shared then anonymization may not be required. It is necessary to consider all factors in the consent obtained, and further, to consider the possible risks of sharing.