SOCRATES: Social Media Research Assessment Template for Ethical Scholarship

Jan Rau Felix Münch Mani Asli

Abstract

The aim of this document is to provide practice-oriented, concise, and helpful guidance to pursue legal and ethical research with digital trace data, particularly online communication and online media data. The guide asks researchers questions which they should answer before, during and after the research process and combines these questions with information that helps to address them. The guide covers ethical and legal challenges including harm avoidance towards and informed consent by research participants, amplification of harmful content, and researchers' safety on the one hand and legal challenges including data protection, terms of services of social media platforms, as well as copyright on the other hand. While the ethics section of the guide addresses questions that are relevant for an international research community, the legal section will be most helpful for researchers based in Germany due to national legislation boundaries.

Contents

If you use this guide in your research, please cite it like this:

`Jan Rau, Felix Münch, Mani Asli (2021): Social Media Research Assessment Template for Ethical Scholarship (SOCRATES): Your politely asking data ethics guide. (Social) Media Observatory.`

Disclaimer: This ethics guide is a living document, compiled according to our best knowledge and intent. We consider it already very useful. However, due to its new and fast evolving subject matter, it might include false or misleading information. We cannot accept any liability for the recommendations laid out here, please double check with a legal and/or ethics expert, if in doubt.

---

1. Introduction

With this document, we want to provide practice-oriented, concise, and helpful guidance to pursue legal and ethical research with passive digital trace online communication and online media data – or basically the type of research the (Social) Media Observatory builds and maintains applications and resources for.

In the following, we present questions that, in our humble opinion, should be answered before, during, and after any research based on online data. The checklist follows a risk based approach: **greater risks to fundamental rights of participants** (all persons affected by the research) **and greater risks of harm to participants, researchers, and society must be accompanied by greater countermeasures to mitigate these risks** – or, if this is not possible, to refrain from following through with a project at all. There are some intended redundancies between the questions, but users are welcome to copy from answers in previous sections if deemed appropriate.

The guide can be used as soon as you want or need to hold yourself and your research accountable against ethical and legal standards. It can be completed and its answers provided to your ethics board; it can be published on your project website, be stored in your archive for personal documentation and/or to comply with legal obligations, or you can ask others to work through it for joint research projects and/or data sharing (the latter, of course, in combination with a respective data processing or data sharing agreement). As it is CC-licensed, you can append to it, cut it or change it according to your specific needs, as

long as you license it similarly and acknowledge the original document and its authors. As it will be version-controlled in an open-source fashion, we invite you to contribute improvements to our version, or 'fork' your own version, if you consider our decisions on improvements questionable in the future.

In addition to this document and based on its guidance, we provide a GDPR documentation template that gives an overview of all the relevant details of your data processing and can be used to fulfil the GDPR requirements for data processing documentation. We recommend taking an initial look at this template to familiarize yourself with the set of details that need to be addressed.

Finally, we explicitly emphasize that this ethics guide operates in an emerging field with debated ethical questions and **substantial** legal uncertainties (from a jurisprudence and case law perspective, relevant legislation like the GDPR is still very new). This guide represents our approach to operate in this space of ambiguity and we hope it can be also helpful to you. However, given that we cannot know how these issues will develop, **we cannot accept any liability** for the recommendations set out here. That's what our lawyers said.

> **Our stance**: In line with the Association of Internet Researchers (AoIR) and rooted in the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979) our understanding of ethical research is characterised by the goals of protecting the fundamental rights of human dignity, autonomy, protection, and safety, maximizing benefits and minimizing harm, or, in the most recent accepted phrasing, respect for persons, justice, and beneficence. As these goals sometimes conflict, careful guidance is necessary to ensure that our research is as beneficial as possible while protecting participants and their rights, and preventing harm towards participants, researchers and society throughout the research process. We are aware that we are operating in a field of ethical pluralism, where varying national and cultural backgrounds might offer differing perspectives on the discussed issues. Acknowledging this, we do not claim general validity of our guide, instead we simply offer our perspective (Internet Research Ethics 3.0, Section 2.3).
>
> **Ethics by design**: This ethics guide draws on and recommends the ethic by design principles as outlined by Leidner and Plachouras (2017): Proactive not reactive: by planning to do things in an ethical way we avoid having to react remedially to non-ethical situations in an unplanned way more often; Ethical as the default setting: by committing to pursuing originally ethical paths, we create alignment within organizations towards a more streamlined set of options that comply with common values; Ethics embedded into the process: a process that firmly includes ethics at all stages and levels is less likely to cause unintended harm; End-to-end ethics: ethics cannot

be confined to one stage; it must be an all-encompassing property of a process from basic research to product design to dissemination or delivery, i.e. the full life-cycle of a technology; Visibility and transparency: a process that is published can be scrutinized, criticized and ultimately improved by a caring community; Respect for user values: whatever values a research institute, university or company may hold is one thing, yet being user-centric means also taking into account the values of the user (of a component, product) and the subjects participating in experiments (ratings, data annotations).

2. Project and data processing description

This section aims at providing some general context about the research project and is based on similar (or sometimes literally the same) questions by the Information Commissioner's Office by the UK government.

The Project

1. Explain broadly the background of your project, what it aims to achieve and what type of **data processing** this involves. This should include information like a) potential project partners and/or sponsors b) research aims c) a list of data sources d) types of data e) methods proposed for data collection and analysis (please detail potential data triangulation), f) plans for dissemination of research findings and/or data. You may find it helpful to refer or link to other documents, such as a project proposal or the GDPR documentation template.

   Processing data: Processing of data is any operation such as the collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure, or destruction of personal data. (see European Commission: What constitutes data processing?)

The purposes of the data processing

2. What do you want to achieve? What are the benefits of the data processing – for you, your stakeholders, and society at large?

The research subjects

3. Who are your research subjects? Are public figures, public speakers, children, or other *vulnerable groups* included?

   Research subject: A person who is the subject of study and whose personal information is used in that research. This information may

be gathered directly from the individual or obtained indirectly. This explicitly includes all theoretically identifiable individuals whose data are processed during your research, even if they are not the main subject of your research (for example, Twitter followers of a person of public interest). (Source: Oxford University Research support Glossary, Human participants)

Public figures and public speakers: Different degrees of prominence and publicity of your research subjects have implications on what might be considered legal and ethical research. Public figures (e.g. politicians or journalists) or public speakers (research subjects prominently participating in public discourse) have different expectations towards the publicity of their data and their data has higher relevance for public discourse and opinion formation than it might be the case for members of the general public. However, even if the processing of such data is thus justifiable to a greater extent, public prominence must not be understood as a blank check for data processing.

Vulnerable research subject: The potentially greater vulnerability of certain subject groups requires special attention by the researcher, e.g. minors, persons with disability, minorities, etc. (see Internet Research: Ethical Guidelines 3.0, Section 3.2.5 Involved Subjects)

Additional information

4. Are there prior concerns over this type of data processing or security concerns? Is it novel in any way? Are there any current issues of public concern that you should consider and address?

   Misuse of research: Your research might involve methods, technologies, data or generate knowledge that has the potential to harm individuals or society if used for unethical ends. While the risk of misuse of research can never be eliminated, it can be reduced by identifying risks and implementing precautions. We want to encourage you to ask yourself whether you can see such a potential for misuse for your own research. If so, it might be appropriate to not only conduct an ethical assessment as provided by this document, but also to conduct a risk assessment for your research project.

3. Ethical requirements

This section provides an overview of ethical requirements and responsibilities that we consider relevant while conducting research with passive digital trace online communication and online media data. These ethical considerations will sometimes go beyond the actual legal requirements as outlined in part three of

this document. However, they might provide helpful guidance for some of the legal consideration processes needed later on. What should become clear in this section is that just because research is legal, it might not necessarily be ethical (and vice versa, as we will see later). And if it is not, it should not be pursued.

3.1 Ethics: Processing personal data

There are two major ethical concerns with the issue of processing personal data: avoiding foreseeable harm towards research subjects and ensuring (informed) consent by the research subjects. The following questions,based on work by the Oxford Central University Research Ethics Committee and the Association of Internet Researchers, help to address these issues appropriately:

Harm avoidance

5. How could your research subjects potentially experience harm as a result of your data processing? If your research subjects include vulnerable groups, please specify the potential harms for these specific groups. Please refer back to question 1) and details like data sources, data types, data collection and analysis (especially privacy invasive approaches such as data triangulation) as well as data dissemination and archiving.
6. How do you intend to minimize such risk? Please elaborate on your measures, e.g. a strict application of data minimization, anonymisation, security and safety rules and/or a non-dissemination and/or a (partial) deletion of the data. If your research subjects include vulnerable groups, please elaborate how you intend to address a potentially higher risk for these groups (if you intend to fill out the GDPR section below, you are welcome to provide a short summary and refer to the more detailed questions and answers below, and/or to refer to the GDPR documentation template).
7. If you cannot rule out the possibility that your research subjects may experience harm as a result of your data processing, do you think your research goal still justifies the processing of the data? If yes, why do you think so? Your response should include references to your answers to the questions 2,3, 5 and 6.

   Personal data: Personal data is information relating to an individual who can be identified directly or indirectly by this information, in particular by reference to an identifier. If such data is publicly available (e.g. in a public Facebook post), it does still count as personal data. (Source: Data Protection and Research, Section C2; Additional reading: https://gdpr.eu/article-4-definitions/)

   Harm avoidance: Avoiding harm to the research subjects is a primary ethical imperative. When processing data from subjects, the researcher(s) must carefully assess how the data they are collecting and processing, the results insights they are gaining, and the dissemination of this data and insights might negatively affect the

research subjects. This is especially important if vulnerable groups are involved, e.g. LGBTQ individuals who do not want to be publicly disclosed. While the need to protect such vulnerable groups seems intuitive, it is important to highlight that the responsibility of the researcher towards their research subjects must not depend on the researcher's political proximity with or distance to the research subjects. Thus, the responsibility to protect research subjects extends to research subjects who might be perceived as "bad actors" (e.g. participants who make racist or misogynist statements). Also for these research subjects, the potential negative consequences of disclosure must be assessed carefully.

Research with children: As the Oxford Research Support points out, children are human beings whose ability to give free and informed consent is in question, while at the same time they are considered as particularly vulnerable. Research with or about children is therefore subject to particularly strict requirements regarding consent and harm avoidance. One strategy to address these concerns might be to implement appropriate precautions to avoid collecting data from children at all, and/or to immediately delete such data if found in the data collection. The required age to use a platform as stated in the ToS as well as general knowledge regarding its usage by children might inform such precautions.

Data minimization and security: The GDPR section of this document provides a detailed overview of useful data minimization and security measures and might be helpfulwhether or not you are obliged to comply with the GDPR. Please read the relevant sections below for more information.

Data dissemination: When considering data dissemination, you should take into account whether the research subjects involved can be considered public figures, members of the general public, and/or members of a vulnerable group. We recommend a brief look at the flowchart developed by Williams, Burnap and Sloan (2017) and/or the text version of it provided by the Oxford Central University Research Ethics Committee (p.7f).

Data storage beyond the completion of the research: Personal data should not be stored indefinitely and there should be a date of erasure as well as processes ensuring this erasure once the scientific purpose permits. However, it is important to notice that it is good scientific practice, and often required by your research institution, to store research data for a certain amount of time (e.g. 3, 5 or 10 years) beyond the publication of the research (Data Protection and Research, Annex B and Forschungsdatenschutz, Section 7). This measure ensures that the research can be scrutinized and therefore might justify the storage for a certain amount of time. In the context

of open science, long term archiving of research data is encouraged, if it is compatible with the rights of the participants. Additional information about the process and the challenges connected with long term archiving can be found in these publications (Jensen et al 2019 and especially Weller 2019) or at the Website of GESIS.

Informed consent

8. Is the data you want to process publicly available (i.e. no registration needed to view the data)? If not, what steps are necessary to access the data (e.g. registration, entering groups or channels, answering questions to enter groups or channels, admission by a moderator, etc.)?
9. What do you think are the (reasonable) expectations of the research subjects regarding their data being processed by researcher(s)? Do you think they might anticipate such processing?
10. If the data is not public, do you have any indications that the research subjects might consent to you processing the data?
11. If the data is not public and you have no or only weak indications that the research subjects consent with the data being processed, do you think your research goal still justifies the processing of the data? If yes, why do you think so?
12. Do you intend to provide general transparency information about your data processing? If yes, does the information indicate:
    - a) the categories of personal data to be processed?
    - b) the source of the personal data, and whether it came from public sources?
    - c) the purposes for which personal data/special category data will be processed?
    - d) the people or organisations the personal data/special category data will be shared with?
    - e) the legal basis for the processing of the personal data/special category data?
    - f) any international transfers of the personal data/special category data?
    - g) when the personal data/special category data will be erased?
    - h) the rights of the persons involved in your research which they have under the GDPR?

Informed consent: Oxford University Research Support describes informed consent as one of the founding principles of research ethics, which aims to ensure that human participants can enter research freely (voluntarily) and with full information about what it means for them to take part, as well as giving their consent before participating in the research. However, as highlighted in Internet Research Ethics 3.0 by the Association of Internet Researchers, this kind of consent is clearly impractical in the case of big data projects. While the

question of consent remains controversial, there is agreement that the following factors should be considered when deciding whether or not explicit consent is needed for processing data of research subjects:

- Research purpose (the degree of public interest in and public benefit of the research. For some types of research and their research goals, processing data against the will of the research subjects lies in the nature of the research field, which may be the case for research on disinformation or cyberbullying, for example).
- Publicly available data vs. not publicly available data (technical layers of accessibility. E.g. the researcher needs to register or get a moderator's approval before viewing data).
- Public figures (e.g. politicians or journalists) vs. public speakers vs. general public (e.g. ordinary citizens).
- Expectations of the research subjects: They might not perceive their online activity as "public" or might not expect more attentio n beyond their immediate online community, even though their posts are technically publicly accessible).
- Type of processing (e.g. analysis vs. dissemination).

Transparent processing: If you are processing personal data of individuals, being transparent about this data processing strengthens the individuals rights and their possibilities to object to the data processing. However, in research with passive digital trace data about online communication, such a requirement would make this kind of research mostly unfeasible. The GDPR as an example (which usually requires that participants are informed about data processing for research purposes), reflects this challenge with an exemption of the obligation to inform if it implies a disproportionate effort. Still, we believe that some form of transparency about the research and the data processing should be established, as it can help to achieve a certain level of scrutiny towards the public and potentially involved research subjects. This could be achieved, for example, by publishing respective information on your website (before the start of the data collection) and by open-sourcing your methods.

3.2 Ethics: Amplification

Unreflective dissemination of research data and insights within the context of sensitive research fields, for example mental health or extremism research, can lead to the amplification of harmful content or ideology. Researcher(s) investigating such a field must therefore carefully reflect on whether and how data and insights can and should be shared.

13. If your research is centered around sensitive issues, such as mental health (e.g. self-harm or eating disorders), abuse, misogyny, disinformation, or

extremism research, how do you address the potential problem of amplifying harmful content?

Amplification: The concept of amplification of harmful content and ideology has gained increasing attention in the context of media reporting on extremist ideology. In this case, it describes the process of unreflective coverage of extremist ideology and activity that can lead to an unintentional amplification of this kind of ideology. Similar discussions of harmful unintentional side effects of media coverage on sensitive issues center around suicides or mass shootings. While these debates are mostly focused on media coverage, research in general faces similar challenges: An unreflective dissemination of the research and its results could similarly lead to unintentional harmful consequences. For example, a disclosure of extremist communities and accounts could draw further people to them. It is therefore crucial for this kind of research to carefully reflect on whether and how findings and data can and should be shared. Here, a careful balancing between harm avoidance and good science principles like open science needs to be pursued. Potential solutions can, for example, include the introduction of additional safeguards, e.g., sharing data only upon request with trustworthy institutions and individuals.

3.3 Ethics: Researchers' safety

14. Have you informed yourself about potential risks for your own safety related to your research and are you confident that sufficient counter-measures are in place to minimize such risks? (Note: Here, a simple "yes" will be sufficient in most cases, but we want to highlight the importance of this question. Please see below for some resources and feel free to ask us for more.)

    Researcher safety: As highlighted in Internet Research Ethics 3.0, there is a growing need to protect the safety of researchers. Phenomena such as "Gamergate" and similar events highlight comparatively new risks and levels of risk posed to researchers whose work – and/or simply their public identity (e.g., ethnicity, minority identity, sexual identity, political activism, etc.) – triggers strong, often ideological reactions: these include personal insults, death threats, "doxing" (publishing private information about the researcher, thereby fanning the flames of further hate speech, threat, etc.) and so on. Similarly, research on violent online and offline political extremists, including jihadists, risks direct threats and retaliation if researchers' identities are exposed. To counter such risks, appropriate measures starting with personal data safety and crisis communication strategies should be in place. Finally, the mere reviewing and curating, e.g., videos of beheadings and other forms of violence, abuse or self-harm, can have serious consequences for researchers' own psychological health

and well-being that in turn require – ethically, if not always legally – therapeutic counter-measures as part of the research project and process. (Source: Internet Research Ethics 3.0, Section 3.1.3)

Additional resources: https://ssd.eff.org/en/playlist/academic-researcher https://datasociety.net/library/best-practices-for-conducting-risky-research

4. Legal requirements

The legal requirements for your research will differ substantially depending on the legal framework you are operating under. In the following, we will lay out the relevant legal aspects for the European Union and in particular for Germany. Even if you are based in another country, this might give you an indication of the relevant issues you need to consider and how they might be addressed. However, **please make sure to check your national legislation**, especially whether it holds additional rights and/or duties in the field of data processing for academic research.

Once again, it is important to highlight that after working through the questions below you might find certain aspects of your research to be illegal even if you found them ethically acceptable earlier and the other way around. As we underlined already: Make sure that your research is both legal as well as ethical.

4.1 Legal: GDPR and the German Federal Data Protection Act

For any type of data processing within the European Union, the *General Data Protection Regulation (GDPR)* combined with the national specifications of the GDPR's opening clauses (in our case: the *German Federal Data Protection Act, BDSG)* provide the minimum legal framework for processing personal data. There might be additional legal requirements, for instance legal provisions applicable to your university or to the type of data collection you are conducting and/or tools you are using. Please check (again) whether you are aware of any additional legal frameworks.

Among other things, the GDPR requires a fair, lawful, and transparent processing of all *personal data* and imposes data minimization and accountability obligations on the "data controller" as well as the "data processors". This will be you as the researcher(s) and potential third parties you might involve.

Data protection principles under the GDPR

Personal data must:

- be processed lawfully, fairly and in a transparent manner;
- be collected only for specified, explicit and legitimate purposes, and not be further processed in any manner incompatible with

those;
- be adequate, relevant and limited to what is necessary in relation to the purposes for which it is processed;
- be accurate and, where necessary, kept up-to-date; not be kept as identifiable data for longer than necessary for the purposes concerned;
- be processed securely.

To make sure that the researcher(s) comply with all legal obligations and requirements, the following questions, some of which are based on similar (or in part identical) questions by Oxford University Research support, should be answered:

Scope of GDPR

15. Does your project involve the processing of personal data?
16. Does your project involve the processing of special categories of personal data?

    Processing data: Processing of data is any operation such as the collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction of personal data. (see European Commission: What constitutes data processing?)

    Personal data: Personal data is information that relates to an individual who can be identified from that information, whether directly or indirectly, and in particular by reference to an identifier. Please note that this definition is intentionally broad and that it specifically covers information which "only" relates to an identifiable individual. If you are unsure about whether your data does count as personal data, have a look at this elaboration provided by gdpr.eu.

    Special category personal data: Special categories of personal data include any personal data consisting of the following information of a person: racial or ethnic origin; political opinions; religious or philosophical beliefs; trade union membership; genetic data; biometric data to uniquely identify a person; health; sex life and sexuality. You should be aware that the derivability of such information on the basis of the data classifies the data as special category data, too.
    (Source: Data Protection and Research, Section C2; Additional reading: https://gdpr.eu/article-4-definitions/)

    Publicly available data: You might wonder whether data that relates to an individual but is publicly available (e.g. data posted publicly on a social media platform) counts as personal or special category data, even if the research subject has published it voluntarily. The answer is: Yes, it does. As long as the data relates to an individual and an individual can be identified from this information,

this data must be considered as personal respectively special category data.

Lawful and fair processing

For processing personal or special category data under the GDPR, you need a legal basis to do so. Please have a look at the infobox below for more information about what kind of legal basis might be suitable for your research.

17. Do you use **Art 6 (1) f) GDPR in connection with Art 85 GDPR, § 27 BDSG** to conduct your research (see information box below for information about the legal basis of your research)? **If no, simply answer no.** If yes, please elaborate why it is an appropriate legal basis for your research and how you intend to comply with the respective legal obligations. To do so, please specify:
    - a) Why is the data processing necessary to achieve your research goal?
    - b) Does your research interest substantially outweigh the interests of your research subjects? You are welcome to build your arguments on your answers to the previous questions. Please also consider the additional information about fair processing and participant rights.
    - c) Do you intend to anonymize your data? If not, why is this case and how will you proceed instead?
    - d) If you intend to publish the data, do you have consent to do so or can you convincingly argue that this is indispensable for presenting research findings on events in contemporary history?
18. Do you use **a different legal basis** than Art 6 (1) f in connection withArt 85 GDPR, § 27 BDSG for data processing in pursuing your research? **If no, simply answer no.** If yes, please detail how you intend to comply with the respective requirements. As an example: Informed consent might free you from some obligations (e.g. anonymization), but might impose additional ones on you (e.g. that research subjects have the right to revoke their consent and you must be able to delete their data at any point in time). Make sure to list all relevant details regarding your legal basis and how you comply with it.

    Legal basis: To ensure and document the lawfulness of the processing of personal data, there is a wide range of potential legal bases from which you can choose the most appropriate one for your specific research project. Within the scope of this paper, however, we focus on a specific derogation provided by the GDPR in Art. 6 (1) f in conjunction with Art. 85, which covers certain types of data processing for research purposes. This basis embraces the majority of research involving digital trace data, particularly online communication and online media data.

    Legal basis for processing personal data: When processing

data for journalistic purposes or for the purpose of academic, artistic, or literary expression, and if you can ensure to only process personal data that falls into the category of normal and not into the category of special data, you might be able to base your data processing on Art. 6 (1) f in connection with Art. 85 GDPR. Be aware that a distinction between personal data and special category can often be challenging as the data might not directly include special category information, but can be taken as a basis to infer such special category information (e.g., group associations via follower networks). In that case, the data needs to be treated as special category data. Examples of personal data that is not special category data might be account meta information like name, date of creation or follower number.

Legal basis for processing special category data: When processing (special category) data for journalistic purposes, or the purpose of academic, artistic or literary expression, the GDPR tasks the Member States in Art. 85 (2) with providing exemptions or derogations from a variety of GDPR provisions. This template therefore incorporates the Federal Data Protection Act (BDSG) as the German implementation of this task. Specifically, it takes Art 6 (1) f in connection with Art 85 GDPR, § 27 BDSG as a legal basis for the data processing, as this basis seems to be the most appropriate basis for most of our work. It allows us to conduct the data processing without explicit consent from the research subjects.

Data processing for purposes of scientific or historical research and for statistical purposes: To use this legal basis for your data processing, you need to

- outline how the data processing is necessary to achieve your research goal. Here you need to provide information on why there is no less intrusive research approach to achieve your research goal.
- outline whether your legitimate interest in conducting research based on personal data from research subjects outweighs the right to data protection of the research subjects/persons affected. You might use the elaborations and arguments from your ethical considerations above.
- Paragraph 27 BDSG states that special category data must be anonymized if it is processed based on this legal basis (pseudonymous data is NOT anonymized data in the GDPR sense). In case you want to process the special category in a non-anonymized way, you have to outline why it is not possible to pursue your research interest with anonymized data. In this case, and if your research interests can be considered more important as the potential level of risks of the individuals affected, you are allowed to proceed with pseudonymous

data and separately stored identifiers to link the pseudonyms with the actual persons. (Note: This step focuses on special category data as personal data does not necessarily need to be anonymized; for information about anonymization and pseudonymisation look at these elaborations provided by the Oxford University Research Support).

- You are only allowed to publish personal data if the data subject has consented, or if this is indispensable for the presentation of research results on events in contemporary history. The latter can, for example, apply to a person of public interest, e.g. a politician or other forms of public speakers. The higher you deem the public interest in your findings, especially their relevance for public discourse and opinion formation, the more likely your research interest tops the rights of the individual.
- Art. 6 (1) f in conjunction with Art. 85 GDPR, § 27 BDSG continues to apply as long as technical and organisational measures are in place to provide appropriate safeguards for the rights of research participants, as described below (Source: BDSG, Art 27).

'Fair' processing requires researchers to consider more generally how their use of personal data affects the interests of the individuals to whom it relates. Where your use may cause detriment to an individual, you need to consider whether or not that detriment is justifiable. Fairness is also naturally linked to the transparency of the processing and the ability of the individual to object. Here, the answers to the ethical questions in this document might help you in your argumentation. (Source: Data Protection and Research, Section F1)

Participants rights: The GDPR grants individuals affected by the data processing new or improved rights regarding their personal data, including the right to access the data, the right to object to processing, the right to request that the data be deleted (the 'right to be forgotten'), the right to request that the processing of the data be restricted and the right to request the rectification of inaccurate or incomplete data. However, these rights are in any event not absolute. Where personal data is processed solely for research purposes, these rights do not apply to the extent that they would prevent or seriously impair the achievement of those purposes. (Source: Data Protection and Research, Section H)

Data dissemination: If you are processing personal data or special category data, you are only allowed to publish personal data if the data subject has consented or if this is indispensable for the presentation of research results on events in contemporary history.

The latter can, for example, apply for a person of public interest, such as a politician or other public speaker. The higher the public interest in your findings, especially based on their relevance for public discourse and opinion formation, the more likely your research interest tops the rights of the individual.

Data Minimization

19. Are the items of personal data/special category data to be collected at the minimum level that is necessary to achieve the research objectives, while still being suitable for the research purposes?
20. Has the possibility of using anonymized or pseudonymized data been considered? (See Data Protection & Research C2.) Explain how/why the data will (not) be anonymised or pseudonymised. (Note: You are obliged to anonymize if possible and you need to justify if you do not intend to do so.)
21. Will access to the participants' personal data/special category data be restricted to authorized persons? Will there be different access layers within the research team? Will these layers be clearly documented?
22. Will participant data be kept in the form of fully identifiable data for a fixed period of time?
23. Is there a clear rationale for how long the data will be kept as fully identifiable data?
24. Will organizational arrangements be made for the secure disposal and or destruction of personal data/special category data when it is no longer required?

Data Minimization: This data protection principle aims to prevent the collection of unnecessary personal data. Given the sensitivities associated with personal data, it follows that no organisation should store personal data that it does not require. However, this data protection principle also imposes an obligation to ensure that such data is suitable for the researchers' purposes. The GDPR emphasises that the principle of minimisation applies to all aspects of processing, and not just the amount of data collected. It is therefore important for researchers to consider their obligations under this principle in relation to each aspect of work that involves the processing of personal data. (Source: Data Protection and Research, Section F3)

Date of erasure of personal data: The GDPR requires that data is not kept as identifiable personal data for longer than is necessary in relation to the purposes for which it is processed. However, to account for good science principles, personal data processed solely for research purposes may be stored for longer periods, provided there are appropriate safeguards, such as pseudonymisation and/or encryption. As two examples: The Leibniz-Institute for Media Research | Hans-Bredow-Institute (HBI) and the University of Cologne state a 10 year period as good scientific practice for the storage of

research data. (Source: Data Protection and Research, Annex B; Grundsätze zur Sicherung guter wissenschaftlicher Praxis, Section 1; Forschungsdatenschutz, Section 7)

Security

25. Will personal data/special category data be collected, transmitted, and stored securely? For example, will it be encrypted in transit and/or storage and how is access control implemented, also and in particular on remote or mobile hardware?
26. Is the level of security to be provided appropriate to the risks posed by the processing? If you are undertaking high risk research (e.g. particularly sensitive data, particularly intrusive methods, particularly vulnerable research subjects) please elaborate on how your security measures account for such an increased level of risk.

Accountability

27. Will you provide records of all processing activities, including information about the data, the data source, the involved research subjects, the legal basis for the processing, the nature of the data, the scope of the data, a list of all processing activities and their timeline, as well as technical and organizational for data minimization and security? (Note: We propose that this requirement by the GDPR will be met by filling out the linked data flow information chart.)

Other Safeguards

28. Are all individuals involved in processing aware of the importance of data protection and how it can be achieved? (Note: This could be achieved, for example, via regular awareness training and onboarding protocols.)
29. Is there a procdeure in place to ensure the security of the processing via regular monitoring, assessment and evaluation of the effectiveness of the technical and organizational measures?
30. Do you have a data protection officer to oversee your data processing? (Note: If your data processing is not particularly intrusive, you might not need a data protection officer. However, in cases of intrusive data processing, it might be an important additional safety layer to secure a legal and ethical process).
31. If the data is to be shared with another organisation, will there be a written agreement with the other organisation setting out respective roles and responsibilities, and how individuals may exercise their rights in respect of their data?
32. Do you plan to conduct a data protection impact assessment?
33. If the data is to be shared with another organisation specifically for processing purposes under your authority as a data controller, will there be a data processing agreement setting out respective roles and responsibilities, and how individuals can exercise their rights in respect of their data? (This

will be usually provided by the other organization, e.g. a hosting provider etc.)

34. If you transfer data outside of the EU (e.g. to a third partner or when storing it on data storage services such as Dropbox) have you ensured that you comply with the GDPR requirements when doing so? This is not a simple question, since the European Court of Justice has invalidated the EU-US Privacy Shield in July 2020.

    Data controller and data processor: A data controller is the person who (either alone or jointly with other persons) determines the purposes and manner in which any personal data is or will be processed. Essentially, you are in control of how you use the data – the data processor is the one who does the processing of personal data on behalf of the data controller. (An example for a data processor involved in your research might be a cloud computing platform you use to process your data.) (Source: Data Protection and Research, Section E)

    Data protection impact assessment: A type of processing likely to result in a high risk to the rights and freedoms of individuals might require an assessment of the impact of the intended processing operations on the protection of personal data. This can be the case, for example, for processing a large scale of special category data or for the systematic monitoring of a publicly accessible area on a large scale.

    Transfers of personal data to a country or territory outside the EEA: Such a transfer may take place if one of the following conditions are complied with:

    - **Transfer on the basis of an adequacy decision.** The European Commission considers the data protection laws in that country or territory ensure an adequate level of protection for data subjects. To date (16.03.2021), only the following have passed the test: Andorra, Argentina, Canada (for commercial organisations), Faroe Islands, Guernsey, Israel, Isle of Man, Japan, Jersey, New Zealand, Switzerland and Uruguay. It should be noted that the EU US Privacy Shield Framework has been ruled unlawful by the European Court of Justice and does not provide an adequate framework for data sharing anymore.
    - **Transfer subject to appropriate safeguards.** Transfers may occur if the controller and processor have provided appropriate safeguards, and on condition that enforceable data subject rights and effective legal remedies for data subjects are available. This includes (most commonly) the use of standard contractual clauses which have been approved by the European Commission.

Cloud service providers: Researchers need to bear in mind that using international cloud-based services, e.g., Dropbox, OneDrive, Google Drive/Cloud, or Amazon Web Services, may involve a transfer of personal data outside the EEA. The risks will be greater where the personal data involved is confidential or sensitive. You can counter or reduce these risks if the service in question offers the opportunity to store the data within the EAA (which, however, might still not be sufficient to protect it from state authorities), if you technically secure the data with strong encryption, or if you use an alternative service that cannot be forced by foreign state authorities to not comply with GDPR. (Source: Data Protection and Research, Section G)

4.2 Legal: Terms and Conditions of Social Media Platforms or Social Networking Apps

35. Do you expect a breach of terms and conditions of a service you are researching? If yes:

- a) Are there other ways of conducting this research to avoid the breach?
- b) Do you plan to use a developer account or other data sources by the platform that require agreement to special terms? (Note: The platform might deactivate or delete your account if you breach the ToS).
- c) Is your data collection embedded in a non-commercial and non-profit research project and is the data collection necessary to pursue your research?
- d) Does your data collection focus on certain parts of the social media platform rather than the platform as a whole?

Web Scraping and Terms and Conditions: The Terms and Conditions (T&C) or Terms of Service (ToS) of Social Media Platforms or Social Networking Apps often contain phrases that limit or forbid the possibility of automatic data collection without explicit permission. However, as Golla and v. Schönfeld point out, such a limitation might be a unilateral and, in the absence of legal binding force, non-binding condition, as long as the data is publicly available and can therefore be collected without agreeing to the ToS. While web scraping of publicly available data for scientific research purposes is indeed connected to various legal uncertainties, it can generally be considered legal as long as it is limited to certain parts of the platform (as opposed to the platform as a whole) and as long as it is necessary to pursue the research. (Reminder: This section reflects the German legal situation and might differ in other legislations.)

4.3 Legal: Copyright

36. Is the material you wish to use protected by copyright?
37. If so, what is your legal basis for processing the data and why is it appropriate for your data processing?

Copyright: When analysing social media data, you should be aware that different types of content including text, image and video data might be protected by copyright. This copyright protection remains regardless of the type of use, such as publishing the content at a publicly accessible place (e.g. a social media platform). If content is protected by copyright, you need a legal basis to proceed with data processing. As it is usually impossible to get the permission of each content creator for the data processing in the context of digital trace online communication and online media data, it is important to understand what other legal bases your national legislation offers you to continue with your data processing. Looking at Germany, Section 60c and 60d of the German Act on Copyright and Related Rights define the following rules for scientific, non-commercial research:

Small-scale work (e.g. social media posts): Section 60c paragraph 1 and paragraph 3 allow the reproduction of small-scale works for the purposes of non-commercial scientific research to the necessary extent. Posts on social media sites can usually be considered as such (Golla and Müller, 2020). Content beyond small-scale work (e.g. news articles) for your own research: Section 60c paragraph 2 allows the reproduction of up to 75% of a content piece for your own research, however, this data must not be shared (this does not apply to automated text analysis, see below).

Content beyond small-scale work (e.g. news articles) to be shared with others: Section 60c, paragraph 1 allows the reproduction of up to 15% a content piece and its sharing with a limited set of people for their own scientific research and for individual third parties, insofar as this serves to verify the quality of scientific research.

Text and data mining (automated text analysis): Section 60d allows the automated and systematic reproduction of copyrighted text material to create a text corpus for text and data mining purposes. You are allowed to share the corpus to a defined group of people for joint research, this might include members of your organisation in general or external researchers for a specific joint research effort. This text corpus must be deleted or transfered to an archive after finishing the research project. As the end of a research project sometimes might be difficult to assess, it might be helpful to draw up a project plan with clear research objectives before the actual start of the research (Ahlberg/ Götting/Hagemeier 2019: § 60d Rn. 20). Another indicator might be the administrative end of a research

project, for example due to the expiry of the funding period for externally funded projects (Spindler/Schuster/Anton 2019: § 60d, Rn. 15-17). (Source: Urheberrecht in der Wissenschaft, p.21-23)

Additional note and disclaimer: Assessing whether internet content is actually protected by copyright or not can become an extremely challenging and tedious process. The assessment is dependent on a variety of different components which might not all be apparent for every single content piece in the process of the data collection. To conduct this assessment for a large data collection might be completely infeasible. As soon as you are processing content data (e.g. text, image and video), you should therefore think about implementing your research process to be compliant with copyright law by default. This will make your data collection scalable, protect you in borderline cases and might save you from some nasty legal trouble. Having said that, we need to underline that the listed legal paragraphs usually provide a sufficient legal basis for the respective type of content processing. However, there might always be exceptions or additional legal requirements for your specific content. Please check (again) whether you are aware of such a potential exception or additional requirement.

GDPR documentation template

This GDPR documentation template is intended to provide an overview of all the relevant details of your data processing and in addition can be used to fulfil the GDPR requirement for data processing documentation.

DISCLAIMER/LIMITATIONS

We hope that this guide has been, is, and will be helpful to you and many others. However, we are neither lawyers (although we discussed this template thoroughly with law scholars during its preparation) nor have we studied ethics and philosophy beyond our practical needs, nor are we familiar with all jurisdictions and cultural contexts. In addition to these personal limitations, this Ethics Guide operates in a still comparably new field with unanswered ethical questions and substantial legal uncertainties. We think it is unlikely, but it may well be, that you still get into legal difficulties, or that some ethic commission has a different perspective than we do, or that an online platform owner shoots down your project, even after you have worked through this guide and are completely sure that your research is legally and ethically sound.

Therefore – similar to software licensed under the MIT license – this guide is provided “as is”, without warranty of any kind, either expressed or implied, including, but not limited to the warranties of fitness for a particular purpose

and noninfringement. In no event shall the authors or copyright holders be liable for any claim, damages or other liability, whether in an action of contract, tort or otherwise, arising from, out of or in connection with this guide or the use or other dealings in the guide.

List of external resources

Disclaimer regarding citation: As highlighted repeatedly, this document relies heavily on already existing guidelines and other resources, especially provided by the Association of Internet Researchers and Oxford University. As this document aims at simplifying the life of researchers and to ensure clarity and readability, we have not always followed exact academic citation rules when using, building on, or changing text components from these external resources. However, we ensured that the resources are always highlighted as the sources of the respective document parts. In case the authors and providers of the external resources do not feel adequately referenced in our document, we politely ask to contact us so that we rework the respective document parts.

Ahlberg, H., Götting, H.-P. (2018). BeckOK Urheberrecht. München.

Bundesministerium für Bildung und Forschung (BMBF) Referat Ethik und Recht (2019). Urheberrecht in der Wissenschaft Ein Überblick für Forschung, Lehre und Bibliotheken.
https://www.bmbf.de/upload_filestore/pub/Handreichung_UrhWissG.pdf (accessed: 15.13.2021)

Bundesministerium für Justiz und Verbraucherschutz. Act on Copyright and Related Rights (Urheberrechtsgesetz – UrhG). https://www.gesetze-im-internet.de/englisch_urhg/englisch_urhg.html (accessed: 15.13.2021)

Bundesministerium für Justiz und Verbraucherschutz. Federal Data Protection Act (BDSG). https://www.gesetze-im-internet.de/englisch_bdsg/

Central University Research Ethics Committee (CUREC). Internet based research (IBR). Best Practice Guidance 06_Version 6.2. https://researchsupport.admin.ox.ac.uk/files/bpg06internet-basedresearchpdf (accessed: 08.10.2020)

European Commission. Adequacy decisions- How the EU determines if a non-EU country has an adequate level of data protection. https://ec.europa.eu/info/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en

European Commission. What constitutes data processing? https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-constitutes-data-processing_en (accessed: 28.10.2020)

franzke, a., Bechmann, A., Zimmer, M., Ess, C. and the Association of Internet

Researchers (2020). Internet Research: Ethical Guidelines 3.0. https://aoir.org/reports/ethics3.pdf (accessed: 10.02.2020)

First Draft (2019). Responsible Reporting in an Age of Information Disorder. https://firstdraftnews.org/wp-content/uploads/2019/10/Responsible_Reporting_Digital_AW-1.pdf

GDPR.EU. General Data Protection Regulation. https://gdpr.eu/tag/gdpr/

GDPR.EU. What is considered personal data under the EU GDPR? https://gdpr.eu/eu-gdpr-personal-data/

GESIS - Leibniz Institute for the Social Sciences. Data Archiving. https://www.gesis.org/en/services/archiving-and-sharing/sharing-data/data-archiving (accessed: 08.10.2020)

Golla, S. (2019). Kratzen und Schürfen im Datenmilieu – Web Scraping in sozialen Netzwerken zu wissenschaftlichen Forschungszwecken. Kommunikation & Recht, Heft 1. https://baecker.jura.uni-mainz.de/files/2019/01/KUR_01_19_Beitrag_Golla_Schoenfeld.pdf

http://methods.sagepub.com/book/researching-power-elites-and-leadership (accessed: 08.10.2020)

Golla, S. & Müller, D. (2020). Web Scraping Social Media: Pitfalls of Copyright and Data Protection Law. PRIF Blog. https://blog.prif.org/2020/04/15/web-scraping-social-media-pitfalls-of-copyright-and-data-protection-law/ (accessed: 15.03.2021)

Gould, M. S., & Lake, A. M. (2013). The contagion of suicidal behavior. In *Forum on Global Violence Prevention*. https://www.ncbi.nlm.nih.gov/books/NBK207262/ (accessed: 08.10.2020)

Leibniz-Institute for Media Research | Hans-Bredow-Institute (HBI). Grundsätze zur Sicherung guter wissenschaftlicher Praxis. https://leibniz-hbi.de/de/grundsaetze-gute-wissenschaftliche-praxis (accessed: 04.05.2021)

Jensen, U., Netscher, S., & Weller, K. (Eds.). (2019). Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten: Grundlagen und praktische Lösungen für den Umgang mit quantitativen Forschungsdaten. Opladen; Berlin; Toronto: Verlag Barbara Budrich. doi:10.2307/j.ctvbkk1p8

Information Commissior's Office by the UK government. Sample DPIA template. https://ico.org.uk/media/for-organisations/documents/2553993/dpia-template.docx (accessed: 08.10.2020)

Leidner, J. L., & Plachouras, V. (2017, April). Ethical by design: Ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (pp. 30-40).

Marwick, A., Blackwell, L., & Lo, K. (2016). Best Practices for Conducting Risky Research and Protecting Yourself from Online Harassment

(Data & Society Guide). New York: Data & Society Research Institute. https://datasociety.net/library/best-practices-for-conducting-risky-research (accessed: 08.10.2020)

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research. Washington, DC.

Spindler, G., Schuster, F. (2019). Recht der elektronischen Medien. München.

Surveillance Self-Defense. Academic researcher? Learn the best ways to minimize harm in the conduct of your research. https://ssd.eff.org/en/playlist/academic-researcher (accessed: 08.10.2020)

Research Support University of Oxford. Data protection checklist. https://researchsupport.admin.ox.ac.uk/policy/data/checklist (accessed: 08.10.2020)

Research Support University of Oxford. Research ethics glossary. https://researchsupport.admin.ox.ac.uk/governance/ethics/faqs-glossary/glossary (accessed: 08.10.2020)

Research Support University of Oxford. Informed consent. https://researchsupport.admin.ox.ac.uk/governance/ethics/resources/consent (accessed: 08.10.2020)

Towers, S., Gomez-Lievano, A., Khan, M., Mubayi, A., & Castillo-Chavez, C. (2015). Contagion in mass killings and school shootings. *PLoS one*, *10*(7), e0117259. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0117259 (accessed: 08.10.2020)

Weller, K. (2019). Big Data & New Data: Ein Ausblick auf die Herausforderungen im Umgang mit Social-Media-Inhalten als neue Art von Forschungsdaten. In Weller K., Jensen U., & Netscher S. (Eds.), *Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten: Grundlagen und praktische Lösungen für den Umgang mit quantitativen Forschungsdaten* (pp. 193-210). Opladen; Berlin; Toronto: Verlag Barbara Budrich. doi:10.2307/j.ctvbkk1p8.14

Universität zu Köln. Forschungsdatenmanagement. Latest version change: 26.05.2020.
https://verwaltung.uni-koeln.de/stabsstelle02.3/content/forschungsdatenschutz/index_ger.html (accessed: 15.03.2021)

Williams, M. L., Burnap, P., & 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.

https://journals.sagepub.com/doi/full/10.1177/0038038517708140 (accessed: 15.03.2021)