Guidelines for researchers sharing large data sets.

Molly Drummond, Ed de Quincey and Elizabeth Poole[1]

Keele University, February 2023

1. Introduction

A summary of the guide, its scope, who it is for, and how it was made.

Open research aims to enable researchers to share and access data and findings, incorporate multi- and cross- disciplinary approaches, and collaborate across academic and non-academic institutions worldwide. These practices require and rely on an accessible, stable infrastructure of knowledge sharing resources as well as a coherent set of accepted and shared approaches. Researchers are continually developing these practices, in negotiation with existing resource infrastructure, legal and institutional codes, and the ongoing emergence of new and overlooked ethical considerations.

This guide aims to summarise existing practice for researchers sharing and accessing large datasets, particularly for social media data containing sensitive information. It draws from multidisciplinary approaches to open research as many of these have been developed from existing research sharing and archival practices, as well as the work and considerations of other disciplines working with sensitive data. This guide is intended as a buildable summary of current best practice, containing additional considerations for researchers navigating new and developing legal and institutional frameworks contributing to the formation of research practice (for more detailed guidance please see resources list).

To develop this guidance, we conducted an initial scoping study into the existing general practices of researchers sharing large social media data sets on research sharing repositories. From there, we conducted six interviews with researchers and academics who had shared large datasets, who have developed research sharing infrastructures such as repositories, and whose work engages with the ethical and academic considerations of social media research and data generation. Through these understandings of both general practice and specific cases we developed a set of recommended best practice for current research. However, in identifying existing limitations to open research practice, we have also considered further developments, that will help to inform and navigate future research practices.

[1] The authors are not legal experts and the guidance should not be seen as legal advice. Researchers should contact their University's legal services and the social media platform's terms and conditions. We would like to acknowledge Dr Eva Giraud and Dr John Richardson's contribution to developing this project.

2. Checklist

A list of considerations for researchers to cover before sharing a data set

(This document will further expand on these points)

- **I.** Check platform's terms and conditions and other legal requirements (GDPR, privacy, you may need to refer to ii here)
- **II.** Check institutional support
- **III.** Consider sensitivity of data and make ethical decisions, consider FAIR and other relevant principles (see section 3)
- **IV.** Choose repository
- **V.** Label files, share search and criteria, a description of content/project and, if possible, information about how to hydrate data
- **VI.** Share any requirements for the data's reuse and ethical considerations for sharing
- **VII.** Reconsider sharing practices and update files if needed at regular intervals

3. Current research practices

The main substance of the guidance: an explanation of the existing resources available and used by researchers to share data, with recommendations for best practice.

Researchers are using existing infrastructure for data sharing, such as online repositories, archives, and databanks, to store and make their data available (as well as associated tools like software and analysis models), and do so according to current ethical practices, including consent practices, terms and conditions; and academic practices including development histories, associated papers, and outputs used[2] (*see* The Turing Way Community 2021). Subsequent sections will discuss ethical and legal codes in more depth, including how these have generated different practices across countries and disciplines, and how researchers are navigating existing and emerging ethical considerations and practices for sharing data. This section will discuss general current research practices, how they overlap, and their limitations.

Current open research practices for sharing large data sets generally adhere to 3 main codes:

- Collegial/Academic: These practices are concerned with the motivations to share data, and what researchers do currently to make their data accessible.
- Legal/Contractual: These practices describe the adherence to the terms and conditions of infrastructures of knowledge and data sharing (e.g. repositories and social media platforms), and the laws concerning digital data privacy and protection. In this guide we will focus on UK practices, but will also cover laws and terms applying to other countries.
- Case-by-case: These practices are examples of how to navigate the limitations of sharing and accessing shared data, mainly focusing on some issues that researchers have faced when navigating ethical and practical problems with data sharing.

Repositories and data storage

Researchers often use an open repository to share and store their data. A repository, such as Zenodo, or ReShare, can be used to search for data and projects dependent on discipline, topic, type

[2] Outputs are the different kinds of publications that can come out of research, which includes journal and conference papers, but also dataset versions, analysis models, reading lists, coding schedules, and software recommendations (like instructions on how to use hydrators).

of data, as well as particular researchers. Researchers will use these online sites to collate all of the data, models, and methods of analysis for their project – including, often, publications of their findings. These files are available for download; often they are open-access, but sometimes researchers request that you ask permission. **You should always cite the dataset that you use to develop your research, just as you would with any publication.** If you are using a repository to share your data, it is useful to label your files and provide a brief description of their contents as well as the project they were used for. If you are collecting data over a long period of time, many repositories allow for multiple versions of the same dataset to be uploaded, so you can keep updating your research or work together with other researchers to collate similar data.

Your university or institution might have its own repository that you can use and can provide direct support with uploading, categorising, and storing your data long-term. However, you might want to use a more public platform, such as Zenodo, to increase the visibility of your data. Either way, a published dataset will come with a DOI – a unique identifier – that you can use to connect multiple platforms to your dataset, so you won't be limited to a single platform.

Repositories can be used to store, share, and access a range of data types. This guide focuses on social media data, particularly Twitter datasets. The next section focuses on gathering this kind of data specifically, and sharing it on a repository, while later sections discuss data sharing approaches and ethical considerations more generally.

Things to consider:

- How user-friendly is the repository?
- How much control do I have over my data once it is shared?
- Do I need to share *all* of my dataset? Should I provide a test or sample set instead?
- Do I have a DOI for the data?

Sharing Twitter Data

Researchers have used Twitter data to conduct large-scale social research projects, for example on sentiment analysis, the spread of public information, or the building of popular political movements (Huang et al. 2022; Banda et al. 2022; Tekumalla et al. 2022 *see also* Bak-Coleman et al. 2022).

To collect tweets, researchers go through the publicly available API provided by Twitter. This can be used to collect tweets from particular users, containing particular words and phrases (for example hashtags), or between particular dates, although there are set limitations (for example, how many tweets you can collect per day, or how far back you can go to access older tweets). The API is also used to access tweets already collected as a dataset for past or ongoing research.

Datasets of Tweets are often collected and shared in online repositories, such as Zenodo, GitHub, and ReShare. If they are for a specific study, researchers usually include a code that can be used in similar studies, or developed for new ones; these data sets are often called replication packages.

The datasets contain "unhydrated" Tweets, which means that the data only contains the Tweets IDs (a code unique to each tweet). To access the tweet, the researcher would have to use a hydrator software which would then provide additional information, such as the content, date and time, location, and whether or not it is a reply or a retweet.

While technically this information can be used to identify the tweet, sharing data in this way adheres to the terms and conditions of Twitter and general research sharing ethics because those accessing and using the data also have to adhere to these terms. Additionally, if the researcher was to use this data for their own work, they would have to make sure it was anonymized. The researchers sharing data packages on repositories often explicitly request that their dataset and any additional software is cited, as datasets have their own citation and publication history, but it is generally accepted that their work will be cited by researchers using it for further, or "downstream" research.

Things to consider:

- What kinds of data do I want to gather?
- How long do I want to gather data for?
- How will I store the data?
  - What repository will I use to make the data available?
  - Will I need to provide multiple versions of the dataset?

4. Legal considerations

There are two main things to consider when you are planning your research project: firstly, what are the terms of the social media platform you are going to be drawing data from and, secondly, what are the terms of your institution on sharing this kind of data?

The terms and conditions of a social media platform can often be updated, so it is good to check them frequently. The same goes for the API you are using as Twitter, for example, has one API to use whereas other platforms have many – or none! An API is the safest, legal route for you to access data and is supported by the platform, but some researchers argue that this means the platform has too much control over access to the data, as the platform can remove access to the API without warning. Before laws protecting online and personal data, the use of bots and scrapers was widespread for legitimate as well as potentially harmful means – from ticket scalping to investigating online bias. Since they have widely been made illegal, their use has been deterred which can be limiting to forms of research that want to bypass a platform's potential to intervene or oversee data collection. However, due to the advances made to protect online data, as well as social media research developing ethical approaches, researchers are developing innovative approaches to data collection when an API can't be used. Depending on what approach is available to you, it is crucial to conduct research ethically, as the rules of social media platforms are not always in their users' best interests – section 5 (below) discusses this in more detail.

You need to find out what your institution's stance is on social media data collection and sharing. They may have a particular mandate to share project data publicly, and may offer support to do so. For example, asking about what support is available may direct you to specific training opportunities on data management and sharing, ethical and technical expertise, and – potentially – funding. As most of your data sharing will come after the main project has ended and the findings have been published, this is definitely an important question to ask. On the other hand, if something goes wrong or if you are facing problems with gathering and sharing data, it is important to find out if your institution can support you and where to go. This is particularly crucial to find out *before* you start gathering data, as you may need to publish, store, or share it quickly and should know in advance where to go and who to ask.

Things to consider

- What do the terms and conditions (e.g. Twitter's) allow me to do? Is this the same for other platforms (e.g. Instagram? Tik Tok?)
- Have the terms changed since I started the project?

- What does my institution say about sharing data? What support can/should they offer to my project? Will they support me if something goes wrong?

5. Ethical considerations

Researchers of different disciplines have to navigate the ethics of sharing the specific kinds of data that they're working with – this applies to sensitive data of various kinds, including bio-data and online data. Current research practices have developed in consideration of ongoing engagement with ethical questions arising from big data research. Accessibility is a large factor in these considerations. Researchers have discussed how the gathering and sharing of big data sets for academic use requires a careful navigation of problems of privacy, sensitivity, and abuse that come from the large-scale use of publicly-available personal data.

In social media research, scholars have pointed out that social media platform users are mostly not creating content with academic researchers in mind. Although it may be impractical or unreasonable to expect researchers to gain informed consent for each platform user individually, there must be built-in safeguards for social media research that limit the possibilities for abuse of such a publicly available form of data. This is of particular concern with social media data as it is dynamic, hence there may be tweets in your datasets that have been deleted on Twitter. To address this, researchers have developed ongoing general practices that are in accord with social science ethical practice more generally – ways of anonymizing and protecting the data of potential participants. It should be noted, however, that it is extremely difficult to totally anonymise Twitter data and in some cases, subjects have asked not to be anonymized, also for ethical reasons so this might also be a consideration.

In addition, researchers have argued that data from particular communities can be weaponized to exacerbate or create certain socio-economic vulnerabilities, and have developed critical theories, methodologies, and research practice to guide research seeking not to replicate, or even to challenge, such practices. Finally, researchers have also offered situated critiques of uneven accessibility to this kind of data; they argue that the legal and contractual limitations that favour social media platforms, as well as the uneven distribution of resources and knowledge infrastructures, mean that forms of research and researchers themselves have more or easier access to this data than other researchers and research institutions.

As a researcher, your own discipline might have some well-established ethical questions and perspectives that can be adopted (or adapted) for a social media project. However, it may be that looking to cross-disciplinary approaches can help answer questions you have, or offer questions you haven't considered. Look to spaces like The Turing Way (*see below*), a multidisciplinary guide to data research to support you, particularly in the planning stages of your project. Additionally, work in other areas (such as Critical Archival Studies; *see below*) has contributed to the development of large-scale social media data projects and is a good place to start. It is important to consider the FAIR and TRUST principles in each stage of your data management plan, as these are widely used by researchers, librarians, archivists, and repositories to inform best practice. Ethical clearance should always be sought from your own institution before engaging in this type of research. This guide summarises what is currently considered best practice for researchers looking to share and publish their data. However, it is important for you to establish your own ethical stance; in areas such as social media research, practices are still emerging and being debated, and there is room to critique and further develop the gathering, management, and sharing of research data.

Things to consider:

- Who is my research for?
- Who is my research about?
- Why am I sharing my data? How can I make it FAIR (findable, accessible, interoperable, and reusable)?
- Are my discipline's approaches ethical enough? Should I consider another approach?
- Should my data be available for unlimited reuse? How will I know how my data is being used?

6. Useful resources

- **Data sharing websites**

Zenodo, https://zenodo.org/

ReShare, https://reshare.ukdataservice.ac.uk/

Archive It, https://archive-it.org/

Symplectic Elements, https://www.symplectic.co.uk/

ResearchGate, https://www.researchgate.net/

Github, https://github.com/

- **Useful tools**

Hugging Face, https://huggingface.co/

Documenting the Now, http://www.docnow.io/

- **Sharing principles**

Dublin Core, https://www.dublincore.org/

FAIR https://www.go-fair.org/fair-principles/

TRUST principles https://www.nature.com/articles/s41597-020-0486-7

- **Further information and other guidelines**

AOIR, https://aoir.org/reports/ethics3.pdf

Oxford Internet Institute, https://www.oii.ox.ac.uk/

https://www.schlesinger-metooproject-radcliffe.org/ethics-bibliography

UCL guidelines https://www.ucl.ac.uk/data-protection/sites/data-protection/files/using-twitter-research-v1.0.pdf

- **Copyright types**

https://creativecommons.org/

https://www.gov.uk/guidance/artists-resale-right

DMP Online

JISC, https://www.jisc.ac.uk/