By Dr Libby Bishop and Simon Parker

Big data - Ethical best practice

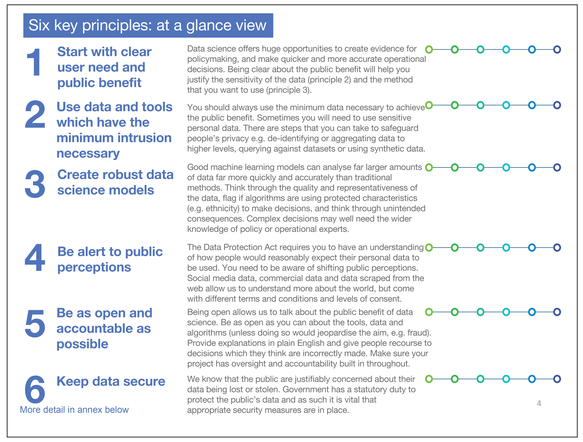
These data often differ from traditional research data (e.g., surveys) in that they have not been generated specifically for research.

- The data collection was not subject to any formal ethical review processes.
- Protections applied when data are collected (e.g., informed consent) and processed (e.g., de-identification), will not have been implemented.
- Using the data for research may substantially differ from the original purpose for which it was collected (e.g., data to improve direct health care used later for research), and this was not anticipated when data were generated.
- Data are less often held as discrete collections and may be linked with diverse sources, indeed a benefit of big data lies in the capacity to link many data sources.

Ethical data resources

UK Cabinet Office Data Science Ethical Framework

This is a general guide for all data scientists. The target audience is researchers in government, but the guide is useful for any data scientist and is relevant across disciplines and genres of data. There are detailed questions and excellent examples.



The OECD report

Is the data personal or sensitive (under Data Protection laws)?

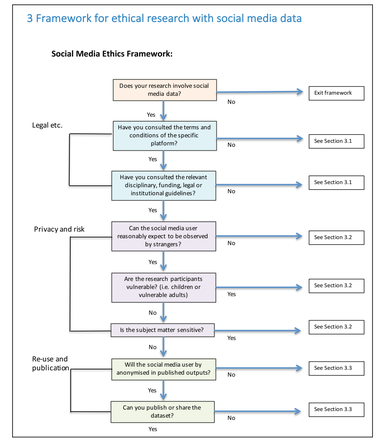
Might the research discriminate against groups, not just individuals?

In what settings is the information gathered, and what uses are expected in those settings?

What are data subjects' reasonable expectations concerning the research project's re-contextualisation of their information?

Social Media Ethical Framework

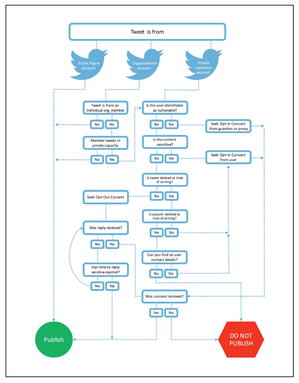
The Social Data Science Lab at the University of Cardiff brings together social, computer, political, health, statistical and mathematical scientists to study study diverse dimensions of New and Emerging Forms of Data in social and policy contexts. Their extensive and practical research has informed this guide for doing ethical research with social media.



Sharing complex data

Publishing social media

Because social media platforms vary greatly in their terms of use, as well as expectations of users (e.g., about privacy), each must be considered separately when publishing data. The Social Data Science Lab has produced an easy-to-use flowchart for making publishing decisions about aggregated and textual data from Twitter.



Privacy Heuristic from the Organisation for Economic Co-operation and Development (OECD)

This guide is the output from an OECD Expert Group on ethics for new and novel forms of data. It is structured around clear questions and offers an international perspective.

What to do when you cannot share the data

In some cases, the data provider's terms of service, or sensitivity of the data, or both, make it impossible to publish or share data. Researchers can still take steps to be transparent about their data and methods. They should create a metadata record in a repository and publish a data access statement, such as this one, explaining why the data cannot be shared and how other researchers can obtain it, where possible.

> Data availability: The data is subject to Twitter terms and conditions which prevent us from redistributing any data collected from the 1% API beyond the research team. However, we are able to supply two limited datasets as Supporting Information files: For age: user ID and derived age in years. For occupation: user ID, derived occupation, SOC2010 cosd and NS-SEC group. Whilst we cannot provide the text from which we have derived our proxy demographics, according to the T&Cs of Twitter we are able to supply user IDs so that others can retrieve the same data as us from the Twitter API.

UK Data Service

ADRN

Luke Sloan ,Jeffrey Morgan, Pete Burnap, Matthew Williams Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data; PLOS One. March 2, 2015 https://doi.org/10.1371/journal.pone.0115545.

Williams, M., 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 (forthcoming).