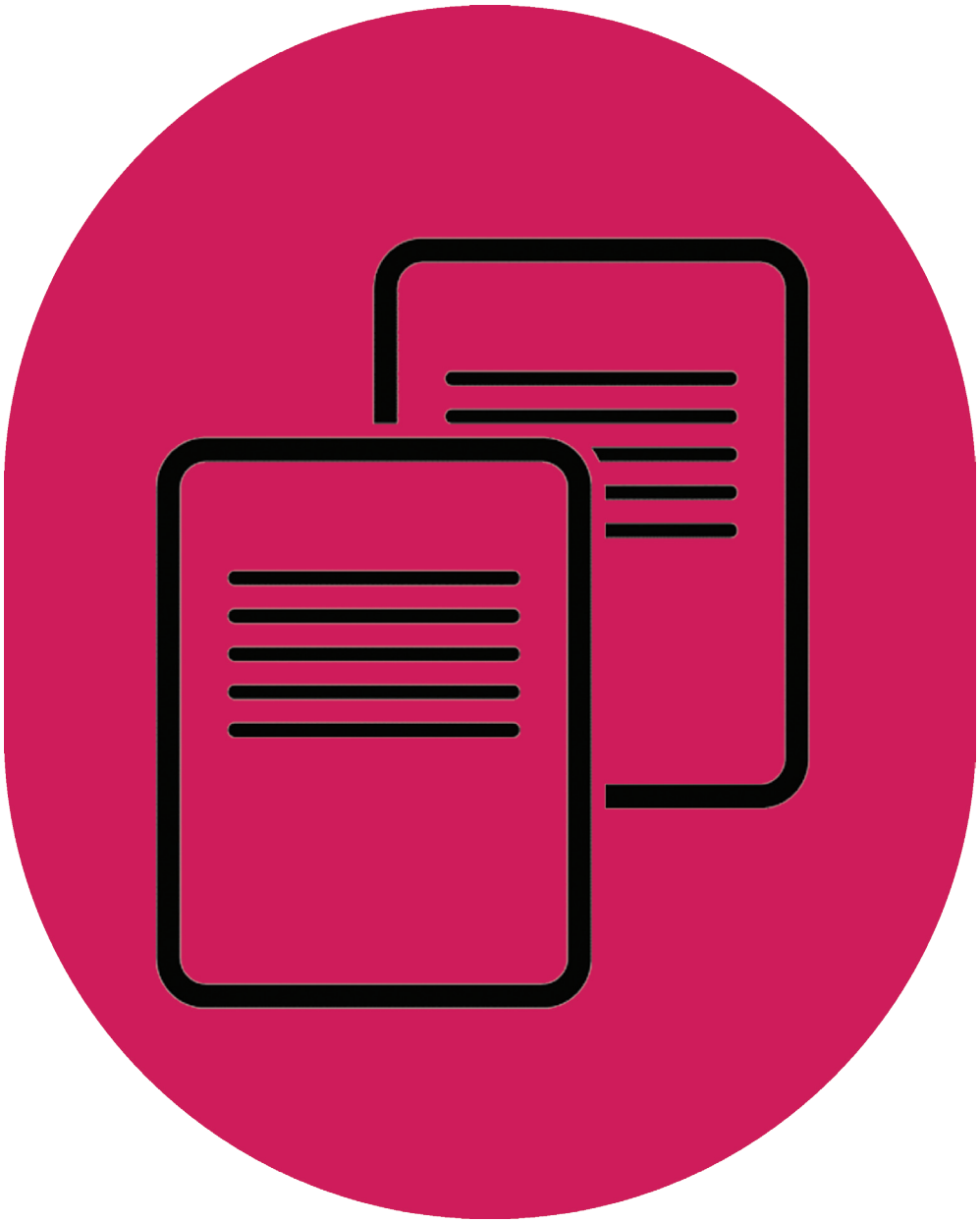
ukdataservice.ac.uk



Big data and data sharing: Ethical issues

UK Data Service

Author: UK Data Service
Updated: February 2017
Version: 1

Contents

This a brief introduction to ethical issues arising in social research with big data. It is not comprehensive, instead, it emphasises ethical issues that are most germane to data curation and data sharing.

The ethical challenges raised by research that uses new and novel data can seem daunting; the risks are both real and substantial. However, the opportunities are also great, and with a growing collection of guidance and examples, it is possible to pursue many such opportunities in an ethical manner.

What are research ethics?

Ethics refers to standards of right and wrong that prescribe what we ought to do, typically guided by duties, rights, costs and benefits. In research ethics, these relationships are among researchers, participants, and the public. Many guides exist, such as the 2016 ESRC's Framework for Research Ethics. There are also more general codes, such as the 1978 Belmont Report, which identifies the core principles of respect for persons, beneficence and justice in human subjects research, and the more general European Convention on Human Rights, or ECHR, ratified in 1953.

What makes research ethics for social research with big data different?

An OECD report (2013) on new data has identified the forms of big data most commonly used for social research as administrative data, records of commercial transactions, social media and other internet data, geospatial data and image data.

These data differ from traditional research data (e.g., surveys) in that they have not been generated specifically by researchers for research purposes. As a result, the usual ethical protections that are applied at several points in the research data life cycle have not taken place.

- The data collection was not subject to any formal ethical review process, e.g., research ethics committees or institutional review boards.
- 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, indeed the value of big data lies in the capacity to accumulate, and pool and link many data sources.

The relationship between data curators and data producers is often indirect and variable. A recent OECD (2016) report argues this relationship is often weaker or non-existent with big data, limiting the capacity of repositories to carry out key activities to safely manage personal or sensitive data.

What are the principal ethical issues in social research with big data?

Privacy

Privacy is recognised as a human right under numerous declarations and treaties. In the UK, the ECHR has been implemented through the Human Rights Act 1998, with protection of personal data provided by the Data Protection Act 1998. The privacy of research subjects can be protected by a combination of approaches: limiting what data are collected; altering data to be less disclosive; and regulating access to data. But big data can challenge these existing procedures:

- The definitions of "private" and "privacy" are ambiguous or contested in many big data research contexts.
- Are social media spaces public or private? Some, such as Twitter seem more public by default, whereas Facebook is more private.
- Many users believe, and act as if, the setting is more private than it is, at least as specified in the user agreements of many social media platforms. Is compliance with formal agreements sufficient in such cases?
- Some approaches to ethical research depend on being able to unambiguously distinguish public and private users or usages. However, data costs and analytical complexity are driving closer collaborations between public and private organisations, blurring these distinctions.
- There is debate as to whether data science should be classified as human subjects research at all, and hence exempted from concerns—such as privacy—that are grounded in human rights.

Informed consent

The ethical issue of consent arises because in big data analytics, very little may be known about intended future uses of data when it is collected. With such uncertainty, neither benefits nor risks can be meaningfully understood. Thus, it is unlikely that consent obtained at the point of data collection (one-off) would meet a strict definition of "informed consent". For example, procedures exist for "broad" and "generic" consent to share genomic data, but are criticised on the grounds that such consent cannot be meaningful in light of risks of unknown future genetic technologies. In 2002, O'Neill noted how this limitation of consent is not new, but the use of data for such different purposes, and the scale of possible harms make it more problematic with big data.

Even if such conceptual issues are minimised (or ignored), practical challenges remain.

- Obtaining informed consent may be impossible or prohibitively costly due to factors such as scale, or the inability to privately contact data subjects.
- The validity of consent obtained by agreement to terms and conditions is debateable, especially when agreement is mandatory to access a service.

De-identification

Unfortunately, there exist no robust, unanimously internationally agreed definitions for the terms *de-identification*, *anonymisation*, and *pseudonymisation*. Generally, a dataset is said to be de-identified if elements that might immediately identify a person or organisation have been removed or masked. In part because a number of relevant laws, such as data protection legislation, define different treatments for identifiable and nonidentifiable data, much has rested on being able to make this distinction. Despite this legal situation, recognition is growing that such distinctions are becoming less tenable.

- Identifiability is increasingly being seen as a continuum, not binary.
- Disclosure risks increase with dimensionality (i.e., number of variables), linkage of multiple data sources, and the power of data analytics.
- Disclosure risks can be mitigated, but not eliminated.
- De-identification remains a vital tool to lower disclosure risk, as part of a broader approach to ensuring safe use of data.

Inequality – digital divide

While the benefits of scale in many domains are clear (e.g., medical care), some see risks in the accumulation of data at a new scale with power that entails, whether data is held in public or private institutions. For reasons of scale and complexity, a relatively small number of entities have the infrastructures and skills to acquire, hold, process and benefit from big data.

- While the question of who owns data is a legal one, the consequences of inequality pose ethical questions.
- Who can access data? In principle, any researcher can access Twitter via its API, but the costs and skills needed do present access barriers.
- Who governs data access? Increasingly, data with disclosure risks can be safely curated, with access enabled through governance mechanisms, such as committees. Is such access genuinely equally open? How is this documented?

Research integrity

Data repositories play a vital role in supporting research integrity by holding data and making them available to others for both validation, replication, 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 data producers, who may not be 'researchers' per se, makes this challenging for a number of reasons:.

- Much data not collected for research, such as administrative data, has different standards (e.g., quality, metadata) to research data.
- For some genres, often with commercial value, such as Twitter data, there are legal restrictions on reproducing data, including providing data to support publications. For a comprehensive treatment of issues of preserving social media, see Thomson 2016.
- Data repositories face challenges in upholding their commitments to standards of

transparency and reproducibility when working with groups of data producers who do not routinely generate data for social research.

Emerging issues

- Alternatives to individual informed consent, e.g., "**social consent"** are being tested whereby sufficient protections are in place to ethically permit data use without individual informed consent.
- There is growing recognition of the need to respect the source and provenance of data—and more broadly its "**contextual integrity"—**when deciding what, if any, reuse is permissible.
- Most research ethics are based on the assumption that the entity at risk is an individual, hence de-identification offers protection. If harms can be inflicted, for example, denial of health care, based on **group membership** with no need for individual identification, then the protection of de-identification is no longer adequate.
- If it is no longer possible to neatly divide public and private, then some suggest assessing data use based on outcomes, and permitted uses with "**public benefit**" or in the "public interest". However, definitions are often vague, and such benefits accrue long after the decision about data use has been made. How can data users be held accountable for delivering the promised public well-being?

Example – Informed consent to publish Tweets

Background to the research question

In 2015 Dan Gray, at the University of Cardiff, used Twitter to study misogynist speech. He encountered numerous legal and ethical challenges with consent and anonymisation when considering how to fairly represent research participants. He collected some 60,000 Tweets in 2015 by filtering on keywords of hateful speech and needed to be able to publish selected quotations of Tweets to support his arguments.

Issues, constraints and decisions

- Twitter's Terms and Conditions prohibit modifying content, meaning that tweets could not be anonymised.
- Gray had to decide if the Tweets could be considered public, and moreover, would their public status be sufficient to justify publishing without consent.
- Survey analysis done at the Social Data Science Lab at Cardiff, where Gray was connected, showed that Tweeters did not want their content used, even for research, if they were identifiable.
- If he did decide to seek consent, there was no way to do so as private communication to the Tweeter. This would have been possible only if the Tweeters were following him, and they were not.
- Mutual following was not possible as a way of contacting Tweeters because the Research Ethics Committee required that he use an anonymised profile.

Outcomes

- He opted to contact by direct Tweet, though this risked allowing tweeters to find him, and also to contact other tweeters of hateful discourse.
- "Consent by Tweet" severely constrained his ability to explain risks and benefits of the research.
- Consent was successfully obtained for a number of tweets, enabling sharing of selected unanonymised tweets in publications.
- Gray was able to draw upon the UK's COSMOS Risk Assessment for guidance, but points out that its rigorous attention to harm and privacy can become a barrier, shielding hateful discourse from critical scrutiny.

Legal disclaimer

- Ethical practice has to develop, in part because the law nearly always lags what is possible.
- There will always be acts that are legal, not ethical, so law not enough.
- There are complex legal questions for big data also (copyright) and researchers should get legal advice through their research support offices.

Below are just a selection of the resources and references that have informed the content on these pages.

Resources

Guides and checklists for ethical issues in big data social research

- A recent OECD report includes a Privacy Heuristic in Appendix 4 with key questions to consider when beginning research, such as: what are data subjects' expectations about how there information might be used?
- The UK Cabinet Office has produced guidance with detailed case examples for UK research by government, but is relevant to more general research as well.
- The UK Data Service is committed to developing tools and procedures to safely handle data, even when that data has disclosure risks. The Service currently uses a framework called the 5 Safes for selected data, and is adapting the framework for more general application, such as for big data.

Anonymisation

- UK Anonymisation Network. Anonymisation Decision-making Framework. http://ukanon.net/ukan-resources/ukan-decision-making-framework/
- ONS *Disclosure control guidance for microdata produced from social surveys* http://www.ons.gov.uk/methodology/methodologytopicsandstatisticalconcepts/disclosurecontrol/policyforsocialsurveymicrodata

Genre specific

- Tweet publication decision flowchart. http://socialdatalab.net/wp-content/uploads/2016/08/EthicsSM-SRA-Workshop.pdf

References

Evans, H., Ginnis, S. Bartlett, J. (2015) Social Ethics a guide to embedding ethics in social media research. IPSOS Mori. https://www.ipsos-mori.com/Assets/Docs/Publications/im-demos-social-ethics-in-social-media-research-summary.pdf

Gray, D. Talking About Women: Misogyny on Twitter, Master of Science Dissertation, Cardiff University, September 2015.

Information Commission Office. (2014) Big data and data protection. https://ico.org.uk/media/for-organisations/documents/1541/big-data-and-data-protection.pdf

Markham, A. and Buchanan, E. (2012) Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0) https://aoir.org/reports/ethics2.pdf.

Metcalf, Jacob, Emily F. Keller, and danah boyd. 2016. "Perspectives on Big Data, Ethics, and Society." *Council for Big Data, Ethics, and Society*. Accessed December 16, 2016. http://bdes.datasociety.net/council-output/perspectives-on-big-data-ethics-and-society/

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

OECD (2013) "New Data for Understanding the Human Condition", Global Science Forum Report. http://www.oecd.org/sti/sci-tech/new-data-for-understanding-the-human-condition.pdf

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. http://dx.doi.org/10.1787/5jln7vnpxs32-en

O'Neill, O. (2002) *Autonomy and Trust in Bioethics*. Cambridge: Cambridge University Press. http://www.mlegal.fmed.edu.uy/archivos/be/bcc1/trust_and_autonomy.pdf

Richards, N. and King, J. (2014) Big Data Ethics. *Wake Forest Law Review* (49).

Schneier, B. 2015. *Data and Goliath*. New York: W.W. Norton and Co.

Social Data Science Lab (2016). Lab Online Guide to Social Media Research Ethics. Retrieved from http://socialdatalab.net/ethics-resources.

Thomson, S. D. , Preserving Social Media, DPC Tech Watch Report 16-01 (2016). http://dpconline.org/publications/technology-watch-reports

UK Cabinet Office-Data Science Ethical Framework (2016) https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/524298/Data_science_ethics_framework_v1.0_for_publication__1_.pdf

Weller, K. and Kinder-Kurlanda, K. A Manifesto for Data Sharing in Social Media Research. ACM. http://dl.acm.org/citation.cfm?id=2908172&CFID=686568339&CFTOKEN=73278098

Zwitter, A. (2014) Big data ethics. *Big Data and Society*. http://journals.sagepub.com/doi/abs/10.1177/2053951714559253

About Libby Bishop

Libby Bishop (Ph.D.) is Producer Relations Manager at the UK Data Service and based at the UK Archive (University of Essex). She specialises in ethics of data reuse: consent, confidentiality, anonymisation and secure access to data. As a member of the Advisory Panel, she helped to revise the Economic and Social Research Council's Framework for Research Ethics. She is also a member of the University of Essex Research Ethics Committee. Her recent work has focused on big data, ethics and data sharing. She is a member of the Big Data, Ethics, and Society Network and has a forthcoming chapter, "Ethical challenges of sharing social media research data" in *The ethics of Internet-mediated Research and Using Social Media for Social Research* (ed. K. Woodfield).

20 February 2017

T +44 (0) 1206 872143
E help@ukdataservice.ac.uk
W ukdataservice.ac.uk

The UK Data Service delivers quality social and economic data resources for researchers, teachers and policymakers.

UK Data Service

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL