[Collecting and Using Social Media for Statistics and Statistical Research Policy]

Contents

Policy name	Collecting and Using Social Media for Statistics and Statistical Research
Date policy was introduced	August 2018
This policy has been authorised by (SRO)	Sarah Henry
Policy owner	Methods Data Research
Other contacts	Data as a Service
Scope of the policy	All data in ONS
Next review date	July 2020
Release Version	Version 1.0
Status	Approved

1. Introduction

This policy sets out the practices and procedures that Office for National Statistics (ONS) staff will follow when collecting or using data obtained from social media platforms to produce statistics and conduct statistical research, including exploratory research, which serves the public good.

The policy outlines key ethical considerations of using social media data and provides practical guidance to ensure that we use social media data ethically, in line with the National Statistician's Data Ethics Advisory Committee (NSDEC) principles[1], and consistently across ONS.

For the purpose of this policy *social media* is defined as any internet-based application that supports user-generated content and social networking through connecting individual and group profiles[2]. A more detailed definition of social media, and other terms in *italics* are available in the ONS data glossary.

2. Why ONS uses social media data

Use of alternative data sources is a key element of ONS's strategy for delivering statistics, analysis and advice which helps Britain make better decisions[3]. The Independent Review of UK Economic Statistics[4] also recommended ONS pursued the evaluation of alternative data sources and data science techniques for better statistical outputs.

Social media data are potentially valuable in helping understand social and economic features of the UK. The collection of social media data may have advantages over traditional forms of data collection like surveys, such as reduced respondent burden and improved timeliness of statistical outputs. The variety of social media; blogs, discussion forums, videos, images and other shared content, virtual worlds and public comments on digital media, is indicative of the complexities of online social interactions and can offer invaluable insights on population and economic statistics.

Most social media networks provide tools such as *Application Program Interfaces* (APIs) that facilitate the access of datasets for use within the remit of the accompanying terms and conditions. While terms and conditions can vary considerably between platforms they are generally explicit in what uses are permissible.

Examples of potential use within Official Statistics include developing real-time indicators and understanding public sentiment and social dynamics to produce aggregate statistics. Throughout the policy we will use the following examples to help illustrate the ethical and

---

[1]www.statisticsauthority.gov.uk/national-statistician/national-statisticians-data-ethics-advisory-committee/
[2]Obar, J.A. and Wildman, S. (2015). *Social media definition and the governance challenge: An introduction to the special issue.* Telecommunications Policy, 39(9), 745-750.
[3] https://www.statisticsauthority.gov.uk/gsspolicy/better-statistics-better-decisions/
[4] Independent Review of UK Economic Statistics by Professor Sir Charles Bean

methodological issues which must be considered when using social media data for the production of statistics and research:

- Analysis of sentiment within Tweet content to estimate well-being statistics;
- Geo-located social media as the proxy for population mobility and migration patterns; and
- Analysis of public opinion around topics or events e.g. use of data science in official statistics

3. Scope

This policy outlines procedures for the ethical collection, use, analysis and curation of data obtained from social media platforms to be followed by all ONS staff when using social media data for statistics and statistical research.

This includes data obtained through formalised and well-documented routes, such as APIs or direct acquisition from the data owner, and through secondary methods such as web-scraping – for which the ONS web-scraping policy also applies. The scope of this policy does not include any use of social media data for non-statistical purposes, statistical purposes such as operational research and management of an organisation's web presence on social media platforms.

4. Objectives

The purpose of this policy is to ensure that social media data are used responsibly for statistics analysis and advice within ONS. This includes ensuring that social media data are used:
  i.   to meet specific user needs that serve the public good;
  ii. legally, complying with all relevant legislation and upholding the associated terms and conditions of service;
  iii. ethically, in line with the independent National Statistician's Data Ethics Advisory Committee (NSDEC) ethical principles[5];
  iv. consistently following advice from the ONS Data Governance Committee (DGC) and the Commercial Data Team in ONS Data as a Service (DaaS) division; and
  v.   using the highest professional standards for statistics, social research and data science in government.

5. Principles

In using social media data, ONS will seek to maximise the benefit while minimising any risks and negative consequences associated with the production of statistics and research.

---

[5] https://www.statisticsauthority.gov.uk/about-the-authority/committees/nsdec/

The following principles will help ensure social media data are used fairly, ethically and lawfully.

1. Data are used for producing statistics, or statistical research that has clear public benefit that outweighs any associated risks and ethical implications of research on the data subjects.

2. ONS uses the most appropriate data to realise the potential benefits of statistical research.

3. Data are used lawfully,
   - abiding by all applicable legislation;
   - monitoring and proactively adapting to the evolving legal situation;
   - following best practice for Data Protection Impact Assessment; and
   - using appropriate data security and disclosure control measures to ensure the anonymity of individuals is protected in all processes and outputs.

4. Data are used ethically and fairly, including
   - assessing potential ethical concerns, risks to individual privacy and other harm, especially for minors and other vulnerable groups;
   - considering the public acceptability of the data use; engaging the public on sensitive uses of their data; and
   - following advice from National Statistician's Data Ethics Advisory Committee and the ONS Data Governance Committee

5. Statistics, analysis and advice based on social media data are produced using scientific principles; following professional best practice and guidance.

6. Statistics, analysis and advice based on social media data are disseminated transparently with appropriate disclosure control measures

6. Process for using social media data in ONS

Figure 1 shows a summary of the process and decision framework that should be followed if considering using social media data within ONS. The purpose of the process is to:

- provide researchers with a self-assessed checklist when planning new work using social media;

- ensure legal and ethical considerations are embedded in our use of social media data; and

- prompt researchers to seek further advice where necessary, for example when proposing a new or novel use of social media or when the proposed use includes personal data or content of a sensitive nature.

Detailed practices and guidance for each step is provided in section 7.

A light touch approach can be taken for very early scoping / discovery work to evaluate the potential in a future project. At a minimum, staff should inform the DaaS team and ONS Legal Services of their intended use of social media and complete an ethical self-assessment. Sample data should contain no more than 10,000 records and be treated with an appropriate level of data protection and security.

1. Need

Is there a **clearly defined requirement** for statistics, analysis or advice that has **explainable benefit** to users and the public?

NO

2. Source selection

Is the proposed data **the best option** to meet this need? **Have you:**
- Justified the proportionality of the data use?
- Considered accessibility, cost and limitations of the data (using recognised quality frameworks – see Annex A)?
- Consulted DaaS team and the Web Data Group on data held and previous use cases?

NO

YES

3. Legal framework

Does the proposed use fall **within the relevant legal framework**? **Have you:**
- Reviewed relevant legal frameworks; consulted Legal Services on new sources or uses of social media data (and before seeking advice from NSDEC and DGC)?
- Checked the terms and conditions of the network; sought advice from DaaS if you are unsure, e.g. new sources or uses?
- Considered preparing a Data Protection Impact Assessment?

NO

4. Ethics and risks

Are **risks for individuals** and **public acceptability** minimised and approved by DGC and NSDEC? **Have you:**
- Completed an ethics self-assessment?
- Checked for public engagement on sensitive projects / issues?
- Received approval from DGC and NSDEC for all new sources, novel uses and risks that have not previously been considered?

NO

YES

5. Scientific methods

Can **professional best practice** be **implemented**? **Have you:**
- Met the quality required for the need and methods?
- Used the highest level of aggregation possible for this project?
- Put appropriate data protection and security in place?
- Reviewed ethics and consulted DGC, NSDEC if risks emerged?
- Recorded the data on the information asset register?

NO

6. Disseminate

Are the **outcomes cleared** for **publication**? **Have you:**
- Consulted ONS Communications and Media Relations?
- Applied disclosure control?
- Invited peer review?
- Highlighted limitations of the data, methods and results?

NO

DO NOT PROCEED

7. Practices

7.1 Defining the need and public benefit

All uses of social media data must be related to a specific and well-defined need for statistics, analysis or advice that serves the public good. Examples of public benefit are provided in Annex A).

ONS will:
- Clearly outline how the data relate to a defined need and user/statistical requirements;
- clearly articulate the expected public benefit(s) of using the data; and
- included this information in work proposals / plans, when seeking advice from NSDEC and DGC, and when communicating this project and its outcomes.

Figure 1 - Case study examples of needs and benefits

Example 1: Analysis of sentiment within Tweet content to estimate well-being statistics

Better Statistics

To provide estimates of regional aggregate personal well-being at fine-grained temporal level (monthly); to provide timely information on changes in well-being following changes in policy and other public events.

Better Decisions

Measuring the well-being of the nation allows government, and other bodies such as health organisations, to assess the need for investment in well-being related services and monitor the impact of policy and environmental changes. Having more timely information on well-being would allow faster response to changes in well-being, and using geographic information gives a better picture of where changes in services may be required. If alternative sources of data on well-being proved accurate there is potential for savings and reduced respondent burden through scaling back primary data collection.

Example 2: Geo-located social media as the proxy for population mobility and migration patterns (e.g. Twitter, Flickr)

Better Statistics

To provide population mobility estimates that can be linked to official migration flows to calibrate for hard-to-count groups such as internal migration of students and young men - more likley to be missing in surveys / census.

Better Decisions

Knowing where the population is concentrated at different times of the year, or even days of the week, can help planning of resources including access to health and transport services. Using geo-location information from social media to estimate aggregate changes in population concentrations seasonally and over time could provide more timely and granular data, help improve small area estimates, and better information on short term population movements.

Example 3: Analysis of public opinion around topics or events e.g. acceptability of using data science in official statistics

Better Statistics

To explore public opinion and sentiment towards topics of interest, compliment public consultations and search for literature references

Better Decisions

Understanding public perceptions aids ethical assessments; helps researchers address concerns proactively and transparently.

Social media data can provide a broader view than standard consultation alone, and is more cost effective than traditional tools such as focus groups and in-depth interviews.

Searching references to topics could help assess information needs on evolving or developing aspects of society, and identify ways of accessing that information, whether through traditional surveys, or emerging data science methods.

7.2 Select the most appropriate data for the need

Social media data will only be used if it is the best option for meeting the stated user requirements, and after careful consideration of benefits and risks of statistical research.

When using social media data, we will:
- consider alternative sources of data, and assess the relative accessibility, cost and quality of the options
- use recognised quality frameworks to assess quality
- ensure proportionality by:
    - only collecting and accessing data that are necessary to meet our objectives;

- using the highest level of aggregation possible; and

- being transparent of the data we are collecting and how we are planning to use them including secondary data, e.g. geo-tags

- seek advice from ONS DaaS team, and Web Data Group on what data are already held and the previous use-cases for social media data, including where ONS has decided not to implement a proposed use for ethical or legal reasons.

Figure 2: Examples of quality considerations when selecting data

Example 1: Analysis of sentiment within Tweet content to estimate well-being statistics	Example 2: Geo-located social media as the proxy for population mobility and migration patterns (e.g. Twitter, Flickr)	Example 3: Analysis of public opinion around topics or events e.g. acceptability of using data science in official statistics
Relevance: Twitter users are not representative of the whole population		
Geo-located data is a further subset of the UK tweets which may be linked to some population characteristics.	**Relevance:** Social media users are not representative; difficult to infer migration rates at a single point in time; but could be inferred over time.	**Relevance:** Views expressed may not be representative of the wider population
Accuracy: Sentiment in tweets may only reflect how users feel towards the topic of their content, and not translate to how they feel generally	Geo-located data is a further subset which may reduce the representivity but certain social media platforms might be a good proxy for certain groups, e.g. students	Views expressed may be influenced by social networks and not a true reflection of personal sentiment
Comparability: The availability of Tweets may change over time.	**Accuracy:** User needs to be an active poster in order to derive whether is a migrant or not; at least 3 points would be needed to define a user node cluster.	**Accuracy:** Keywords/Hashtags might not catch all the discussion around the specific topic
Coherence: The measurement of aggregate sentiment is not the same as asking direct survey questions, for example on a Likert scale. It is not possible to match the age ranges of Twitter users and survey respondents, so the coverage may differ.	**Comparability:** Geo-location could be affected by default settings changes imposed by the app itself or smartphone providers.	**Comparability:** keywords might change over time

7.3 Understand and comply with the legal framework for using social media data

ONS is fully committed to protecting the privacy of individuals and to follow good practice. We are governed by various laws including data protection legislation (including the Data

Protection Act 2018 and the General Data Protection Regulation) and the Statistics and Registration Service Act 2007. Data protection is important, not only because it is critical to the work of the organisation, but also because it protects individual privacy and maintains confidence. To ensure compliance, all use of personal data will be fair, lawful, proportionate and transparent. In addition, personal data will be held and used with the appropriate levels of technical and organisational security.

Section 39(1) of the Statistics and Registration Service Act 2007[6] states that data which identify businesses or 'bodies corporate' and which have not been lawfully made public should be given the same level of protection as data which can identify individuals.

We may consider sharing aggregate research outputs using social media data with other public sector or academic organisations within our scope of producing statistics and research for the public good, where it is lawful to do so, and in full compliance with the terms and conditions of the social media platform where the data are obtained.

The legal aspects of social media research are developing and terms and conditions and privacy policies might change without notice. Privacy policies and terms and conditions will therefore be checked before each use of the data. Privacy policies may differentiate between services being offered and separate research from other uses.

When using social media data, we will:
- check and abide by the terms and conditions of social media platforms, and contact ONS's Legal Services and ONS Data as a Service team in the event of any uncertainty regarding the terms and conditions;

- abide by all data protection legislation, and other relevant legislation (examples in Annex C). This includes ensuring that personal data are not disclosed in any published statistics or research;

- consider whether the social media data are protected by any international laws and seek legal advice from ONS's Legal Services if there is uncertainty about any aspect of the legal framework;

- comply with data protection legislation by:

  - assessing risks to individual privacy and conducting a Data Protection Impact Assessment if there are genuine risks to privacy which are likely to occur.
  - complying with the data protection principles as set out in the General Data Protection Regulation.
  - removing or anonymising personal identifiers, such as *profile handles* in the research outcomes in line with terms and conditions of service;

---

[6] https://www.statisticsauthority.gov.uk/about-the-authority/uk-statistical-system/legislation/

- conducting analysis at the highest level of aggregation possible, and seeking advice from NSDEC;

- excluding content in our reports and outputs that could identify individuals (such as verbatim quotes) and using appropriate disclosure control procedures. In case we need to demonstrate methods we will need to make sure that quotes used cannot be traced back to the original producer of the quote.

- securing identifying and sensitive data with appropriate technical and physical measures. When linking social media data to other sources of ONS data the sensitivity of that data will be considered; and

- abiding by any clauses in the terms and conditions of social media platforms pertaining to the retention and sharing of collected data;

- make additional consideration when accessing and using data which may contain personal information on minors (children under 16). Seek advice from ONS Legal Services, and key stakeholders, where necessary.

- continue to monitor the legal landscape as it evolves and amend the research's approach accordingly; and

- be transparent about the data being used:

  - where possible using the API provided by the social media platform to collect the data and clearly identify that the data is for ONS i.e. do not use personal developer accounts.

recording our use of data on the ONS Information Asset Register and report it in transparency reports, including what data we hold, how long it will be retained and the purpose(s) for which it has been used.

7.4 Consider ethical issues and risks of using social media data

ONS recognises that there are ethical issues related to using social media data, particularly if they contain personal information about individuals. We may use personal data within analysis, where we can justify that it is lawful, ethical and in the public good for us to do so.

Public acceptability of using social media data for statistical purposes will vary by the social media platform, the demographics and characteristics of its users and the type of content held. We are committed to engaging with the public to understand their views on our use of social media data (see Annex A – Consultation guide).

The decision to use social media data will be made with reference to the balance between risk and negative consequence and efficacy and public benefit, with respect to users' privacy and public acceptability.

Some social media platforms are closed to the general public and not searchable through web browsers. These are usually set up for a specific purpose, have a moderator and require the user to set up an account and log in to see and provide content. Examples include discussion forums for members who share a common interest or concern that requires social support. It would be unethical to access data from closed social media platforms without prior consent from the moderator and its members.

To ensure that our use of social media is lawful and ethical ONS will:

- assess potential ethical concerns when scoping new work using the self-assessment process for ethical consideration and review ethics throughout projects;

- consider the views of the public before using social media data. In projects which may involve personally sensitive topics or minority populations, engage with the public or their representative bodies[7];

- When data may include minors (children under 16):

  - we will consider removing minors from the data. If age is not provided, we will consider the privacy risks of attempting to identify minors and the quality risks of erroneously removing non-minors, and consider alternative sources of data to realise the potential benefits of this research.

  - if the research pertains to minors, conduct public engagement on the acceptability of the research, consider completing a Data Protection Impact Assessment, seek advice from ONS Legal Services, and inform ONS Communications and Media Relations.

- not access data from closed social media platforms without consent from the social media platform.

- when social media data is to be linked with other data which have been collecting using consent as the legal basis for processing, we need to verify if the original consent allows the suggested linkage and generally consider the view of the public on this project. If not, then we may need to obtain additional consent.

---

[7] Researchers should check for existing public consultation and seek advice from DaaS on whether there are similar use-cases within ONS. Consultation is not required for very early discovery work, but should be considered if thinking about developing uses of social media data beyond early proto-types. Consultation should always be conducted with appropriate approval and input from DGC, ONS Communications and Media Relations and any ONS business areas involved in providing statistics, analysis or advice on the research topic.

- for new uses of social media data, request approval by the Data Governance Committee who will consider the organisational perspectives of the research proposal and seek ethical advice from NSDEC before collecting and using social media data.

Figure 3: Examples of ethical issues and risks

Example 1: Analysis of sentiment within Tweet content to estimate well-being statistics

Twitter is an open network and most content shared is publicly accessible via the API; Twitter profiles and tweets are, by default, set to public visibility.

There are many examples of Twitter being used for research.

ONS already produce estimates of wellbeing, so the research topic is not novel and the public may be more respecitve

The analysis is at the aggregate level, providing regional estimates of wellbeing with low risks to privacy, minors and vulnerable individuals.

This research might be however seen more intrusive than collecting the data via survey. This might impact on survey response rates of existing Wellbeing surveys.

Non-personal data from minors could be included in the data, but would be protected by aggregation before analysis.

Attempting to remove minors from the dataset would pose a high risk of re-identification compared to aggregating the data for analysis

Example 2: Geo-located social media as the proxy for population mobility and migration patterns (e.g. Twitter, Flickr)

Although users have chosen to turn on and publicly display their location when posting, and consent through the terms and consitions of use to this data being available to third parties, they may consider the processing of this data locations into migration flows intrusive.

Risks to privacy and identification can be mitigated by anonymising the data during collection; aggregating prior to analysis and applying appropriate disclosure control techniques to outputs.

Due to the higher risks associated with this analysis, ONS would not use geo-location to estimate individuals' location(s) without appropriate public engagement. Data would be aggregated to the highest possible level prior to analysis to mitigate the risk of reidentification in the data.

Example 3: Analysis of public opinion around topics or events e.g. acceptability of using data science in official statistics

This analysis would use publicly accessible content without any personal information. The age limits of each network would need to be verified before collecting the data.

This analysis is at the aggregate level, providing aggregate estimates of sentiment towards the topic, with low risks to privacy, minors and vulnerable individuals.

Risks to privacy and identification can be mitigated by protecting verbatim content with appropriate security and applying appropriate disclosure control techniques to outputs.

7.5 Applying scientific methods and professional best practice

There are significant methodological challenges involved in producing fit-for-purpose statistics and research using social media data. Social media are not always an accurate source of data and the scope of the research should clearly and unambiguously define any assumptions. We will:

- be guided by professional best practice, including the Code of Practice for Statistics, GSR social media guidelines, and NSDEC Principles (Annex A);

- assess, acknowledge and where possible mitigate bias (e.g. users are not representative of populations and may include biases that are not replicated at the same scale outside the platform, on-line behaviour might not be indicative of off-line behaviour, data may contain content from automated bots and professionally managed accounts);

- assess, acknowledge and mitigate any risk of harm to any individuals in the data; and

- seek to understand how datasets are created, and changes in the functionality of platforms, settings and methods to protect the consistency of research across longer timeframes. The pace and scale at which users might create new posts and even remove posts might have a significant impact on research.

7.6 Dissemination of statistics, analysis or advice based on social media data

When disseminating statistics, and research based on social media data we will follow the same professional and organisational standards used for all our outputs. We will also recognise and highlight the unique limitations and challenges of the work.

When disseminating analysis using social media data ONS will:

- clearly communicate the limitations of the research, analysis or advice and until the quality of social media data is better established, any outputs and research based on social media data will be designated experimental;

- invite quality assurance and peer review of experimental methods;

- consult our Communications and Media Relations Team; and

- consider the views of the public and our users

8. Roles and responsibilities

Role	Responsible for:	Accountable to:
Researcher	• Complying with the social media data usage recommendations when conducting research • Consult with ONS Legal Services and seek ethical approval if required	• Responsible Officer
Responsible Officer	• Ensure ONS DaaS team and ONS Legal Services are aware of social media data use • When needed, seek advice from NSDEC	• Data Governance Committee
Data Governance Committee	• Ensure the consistent application of this policy to all ONS staff and assess the organisational risk by conducting research using social media data	• Senior management
Data as a Service	• Remain aware of social media data uses within ONS	• Senior Management
Legal Services	• Provide advice and decision on legal and ethical issues if required	• Senior Management
NSDEC	• Provide advice on ethical issues if required	• Senior Management

9. Compliance

All staff as well as researchers in the wider statistical research community accessing, processing and sharing data should consider the principles and practices before the inclusion of social media data in any research or analysis.

NSDEC will also monitor the ethical use of social media data to ensure that all projects approved by NSDEC have considered the principles and practices in this policy when working with social media data.

There are exceptions to this policy for small-scale exploratory projects. For example, very early scoping / discovery work on no more than 10,000 records to evaluate the potential in a future project. At a minimum, staff should speak to ONS Legal Services, the DaaS team and complete an ethical self-assessment before collecting sample data.

Failure to comply may result in disciplinary action in line with the organisation's Discipline Policy

Staff making a complaint in relation to the application of this policy should refer to the organisation's Grievance Policy

10.   Compliance

Policy Owner:	Head of the Commercial Data team in Data as a Service
Policy Approval:	Data Governance Committee (DGC)
Compliance Monitoring:	Web Data Group
Review and amendments	Data Governance Committee (DGC)

Annex A – Examples of professional guidance for statistics and social research using social media data

[GSR - Using Social Media for Social Research Guidance](#)

[Government Statistical Service – Code of Practice](#)

[DCMS Data Ethics Framework for government (including data science)](#)

[Research-code-of-practice-and-accreditation-criteria](#): Public Interest (paragraph 33.1):

> [..]the primary purpose of a research project must therefore be to serve the public interest in one or more of the following ways, to:

11. provide an evidence base for public policy decision-making;
12. provide an evidence base for public service delivery;
13. provide an evidence base for decisions which are likely to significantly benefit the economy, society or quality of life of people in the UK, UK nationals or people born in the UK now living abroad;
14. replicate, validate, challenge or review existing research and proposed research publications, including official statistics;
15. significantly extend understanding of social or economic trends or events by improving knowledge or challenging widely accepted analyses; and/or,
16. improve the quality, coverage or presentation of existing research, including official or National Statistics.

Relevant ONS data policies

[Consultations guide](#)

[Safeguarding policy](#)

[Web-scraping policy](#)

Annex B – Recognised quality frameworks for assessing data, methods and outputs in official analysis

Civil Service Aqua Book

European Statistical Service Quality Dimensions - Summary

European Statistical Service - Full Quality Framework

UNECE Big Data in official statistics quality framework

Annex C - Legal frameworks to consider

Statistics and Registration Service Act 2007

Data Protection Act 2018

General Data Protection Regulation

Human Rights Act 1998