Using social media for social research: An introduction

Social Media Research Group

May 2016

Contents

Social Media Research Group

The Social Media Research Group was established in March 2014 to achieve the following goals:

- To develop an understanding of robust (and non-robust) social media research and its relevance and application to government.
- To build capabilities within Government Social Research (GSR), and across government, to carry out and critically appraise social media research.
- To raise awareness and disseminate guidance on robust and ethical social media research in government and the appropriate use of findings in policy-making.

The group is co-led by Ant Cooper (Home Office) and Rosanna Mann (Department for Work and Pensions).

This guidance was produced by a cross-analytical subset of the Social Media Research Group, consisting of the following authors:

- Callum Staff, Department for Education
- Hugh King, Department for Transport
- Mary Roberts, HM Revenue and Customs
- Simon Pannell, Department for Work and Pensions
- David Roberts, Welsh Government
- Nathan Wilson, former employee of Ministry of Justice
- Rosanna Mann, Department for Work and Pensions
- Ant Cooper, Home Office

Acknowledgements

The authors would like to thank the members of the *Social Media Analytics Review and Innovation Group* for their peer review of this paper. Thanks are also owed to the various members of the *Social Media Research Group*, without whose help this paper would not have been possible.

1. Executive Summary

Recent years have seen the development and huge growth of 'social media' to a point where it is now regarded as ubiquitous. Just considering social networks (a subset of social media), Facebook had 1.59 billion monthly users (Facebook, 2016) whilst Twitter had more than 500 million tweets sent each day (Twitter, 2015) at the time of writing.

As a result of this proliferation of easily and quickly accessible social media data, analysts and policymakers in government have begun to consider how such data can be harnessed to support robust evidence-based policymaking. This involves a range of considerations, including: the benefits of social media data, the use of social media research as the sole method or in conjunction with other methods, data collection and analysis, ethical implications, and the presentation of social media research findings.

This introductory guidance aims to raise awareness and explore the potential of social media research in government. It is primarily aimed at government social researchers, analysts and policymakers, however it will also be of interest to researchers working outside of government.

The guidance discusses how social media data have a range of attributes associated with them which are not found in 'offline' data and how as such, new approaches and techniques are required. In addition to discussing social media research at a high-level, this document provides a number of working-level resources (for example the Social Media Research Project Lifecycle and the Social Media Methodology Spectrum) which it is hoped will help guide researchers unfamiliar with the topic through their first social media research projects. A guide to resources is also contained in the annex, which contains an index of possible tools and links to social media research centres and will be of use to any researchers attempting to collect and analyse social media data.

Ultimately, this guide is intended to be a starting point for those interested in finding out more about social media research, and should complement regular training and knowledge-building sessions and seminars on the topic.

2. Introduction

This document is intended as a light-touch reference document for those working or interested in working with social media as part of a government social research project. It contains a background to social media and its use within government, data collection and analysis considerations, ethical principles applied to social media research and considerations for the future. We hope that this guidance document will stimulate discussion and thinking around these challenges within government, to enable robust and reliable research to take place using this emerging source of data in the future.

2.1. The history of social media

Recent years have seen the development and huge growth of 'social media' to a point where it is now regarded as ubiquitous. At the time of writing, Facebook had 1.59 billion monthly users (Facebook, 2016) whilst Twitter had more than 500 million tweets sent each day (Twitter, 2015).

Despite recent increases in social media usage, the use of social networks predates the twenty-first century - the world's first social network can probably be traced back to the late 1990s, to a site called 'Open Diary', which allowed users to post and share diary entries with the rest of the online community. From this group came the term 'weblog' which became 'blog' after one user decided to jokingly use 'we blog' in an online post (Kaplan & Haenlein, 2010). Prior to the development of Open Diary, early variants of internet services, such as usenet, allowed users to post and share public messages. As such, the recent increase in the popularity of social media, and social networks in particular, can be seen as a move back towards the historic purpose and use of the internet and the web – facilitating communication between users. (Kaplan & Haenlein, 2010).

2.2. Defining social media

There are many differing definitions of what social media are (and are not). At their root, social media are understood to be web-based platforms that enable and facilitate users to generate and share content, allowing subsequent online interactions with other users (where users are usually, but not always, individuals).

Platforms within this definition can be grouped according to functionality. Kaplan and Haenlein identified six classifications for social media platforms based on their functionality (Kaplan & Haenlein, 2010):

- Blogs and microblog sites (e.g. Twitter, Tumblr)
- Social networking sites (e.g. Facebook, MySpace)
- Content communities (e.g. YouTube, Daily Motion, Pinterest, Instagram, Flickr, Vine)
- Collaborative projects (e.g. Wikipedia)
- Virtual game-worlds (e.g. World of Warcraft)
- Virtual social worlds (e.g. Second Life, Farmville)

Despite being several years old, this framework is a useful starting point for considering the classification of social media, however, as social media expands and evolves it may be necessary to reassess how platforms are classified in the future. Twitter, for example, was designed to be a micro-blogging site and is classified as such above, however more recently has come to be regarded by many as a social networking site.

The Kaplan and Haenlein framework importantly shows that social media is not restricted to social networks (though the two phrases are often used interchangeably). Wikipedia, for example, is not a social network but is a community-created resource designed to facilitate the sharing of content and information. The media (text, images, music and speech) are created, updated and maintained by its users and anyone with access to the web can become a user and make edits to the resource. As such, despite not being a social network, it can be considered a form of social media.

2.3. Social media data and big data

The volume of social media data requires discussion of big data. Big data can be seen to have emerged at the beginning of the 21st century from large scale datasets that private companies began to generate. Some companies, such as Google, eBay, LinkedIn, and Facebook, were built around big data from the beginning (Davenport & Dyché, 2013). Defining big data is difficult; the most widely accepted definition or explanation stems from 2001 when industry analyst Doug Laney proposed the 'three V's' of big data: volume, velocity and variety (SAS Institute, 2015).

There will be numerous times when social media data do conform to, and probably exemplify, this definition of big data. However social media data is by no means always big data. Qualitative analysis of a handful of Tweets is synonymous to textual analysis of a number of person-to-person surveys, and certainly would not fit the above definition. Section 4 provides further detail on analytical techniques.

3. Social media in government

Social media can be used by government in a number of ways:

1) Communication and engagement

This is the use of social media as a communication and multi-participant engagement tool both within government and with external stakeholders. Intra-government use covers internal communication and engagement with other Civil Servants to promote work, share learning and discuss ideas. Government to public communication covers the use of social media as a method of promoting government work and policies (e.g. Foreign and Commonwealth Office's Instagram account, Department of Health's Twitter account). Although still not overly exploited, social media can also be used by government as a public engagement tool to allow a two-way dialogue. For example, open policy-making has explored adopting social media as a way of broadening the range of people government engages with in the process of policy development.

2) Analysis and research

This is the collection and analysis of social media data and includes:

- Analysis of government use of social media.
  This type of analysis monitors and evaluates government social media communication and engagement. An example would be analysing the reach of a departmental Tweet or the dispersal of the information from that Tweet across a social media network of users.

- Analysis and research into the public's use of social media.
  This type of analysis and research can be used to support the development, implementation, review and evaluation of government policy and operational delivery. An example would be researching the attitudes of social media users towards a new service provision, or using public social media data to predict outbreaks of foodborne disease.

4. Social media as a research tool

This guidance document predominantly focuses on the use of social media for analysis and research, rather than for communication and engagement. This is a relatively new area and the majority of social media research projects have only started to gain traction within the last 18 months. Nonetheless, interest in the field of social media research is increasing and opportunities here are wide ranging.

4.1. What is social media research?

The term social media research encompasses any form of research that uses data derived from social media sources. Research in this environment can be classified into two types: using social media as a research tool (such as the use of surveys on social media platforms) and research on the activity and content of social media itself.

Social media research varies from other forms of online research, such as internet-based surveys or webpage reviewing, due to the social nature of the data being extracted from purpose-built platforms. As with any new form of research, several methodological points must be taken into consideration in order to ensure that rigour is maintained; such practicalities will be discussed in this and subsequent chapters of this guidance document.

As with other methods, social media research should be considered one tool in a researcher's toolkit. Whilst some research questions may only use social media research methods, many research questions will benefit from the use of social media research in addition to other methods.

4.2. The social media research project lifecycle

The diagram on the following pages details the main considerations likely to be required during the lifecycle of a social media research project. It uses the Cabinet Office framework for data science projects, as there are numerous parallels here. It also includes examples from two real government projects which made use of social media data. The first is the Food Standards Agency using Twitter to predict cases of Norovirus (Disson & Baker, 2014) and the second is the Scottish Government using social media data to assess experiences of the XX Commonwealth Games (Scottish Government Social Research, 2015).[1]

[1] The XX Commonwealth Games was the 20th Commonwealth Games and was held in Glasgow in 2014.

Social media research project lifecycle

STAGE 1: RATIONALE - BUSINESS/CITIZEN NEED

Considerations:

Social media research is a new area of analysis. It is often far quicker and cheaper than other forms of analysis and data is available in or close to real time.

Based on the above attributes it is suggested that any business or citizen need will be based around these: using insight to deliver a more timely service to the citizen with fewer resources through the support of social media analysis than would have been possible with traditional means.

Twitter/Norovirus Project:

There is a business need to find datasets which allow cheaper/more real time monitoring of potential risks. The citizen need is that by providing an early warning system and a targeted intervention, the number of cases can be reduced, which has both social and economic benefits.

Commonwealth Games:

Because large sporting events such as the Commonwealth Games are consumed through the media as much as by physical attendance, an effective way of understanding the views of the wider population is vital in evaluating the delivery and legacy of such events.

STAGE 2: DATA

Considerations:

Data available from social media sources vary but could include: frequencies (e.g. volumes of posts); user profile information (e.g. demographics); image or text content (e.g. photos and videos); and interactions (e.g. comments; network information).

The primary purpose of this data is not for research so consideration should be given to representativeness, robustness and ethics.

Social media data has the benefits of often being publicly accessible, generated in real time and representative of specific populations of interest

Twitter/Norovirus Project:

Norovirus laboratory reports from Public Health England.

UK-based Tweets containing words relating to Norovirus symptoms.

Both time series ran from January 2011 to March 2015.

Commonwealth Games:

Publically available social media data from the period 14 June to the 6 August 2014.

A limitation was that only English language posts were analysed.

STAGE 3: TOOLS AND OUTPUT

Considerations:

Tools are available for specifically analysing social media data. They help make access and analysis easier for researchers.

Existing general social research tools can also be used to analyse social media data. However, because of differences in the way in which social media is created; some manipulation may be required to render it useful in a social research setting.

Analytical outputs of social media research can range from traditional research reports which present findings, to predictive models which solve real time problems.

Twitter/Norovirus Project:

Tools: Pulsar (Online Twitter Search Platform), SPSS, Excel, Python

Output: Prediction of whether social media data can be used to identify/predict outbreaks of Norovirus.

Commonwealth Games:

Tools: Google Analytics, NVivo, Excel.

Output: Various reports which include content such as analysis of host broadcast coverage and online and digital channels.

STAGE 4: RESEARCH PHASE

Considerations:

Because of the different characteristics of social media platforms, the chosen platform should meet the needs of the research question.

Representiveness of social media data is an area which is still to be properly explored. Whilst understanding of this area is limited, care should be taken to mitigate against any skewing.

The prominence of social media in current society means the volume of data is often on very large scales. Care should be taken to ensure research generates a dataset of a size which can be handled by the subsequent analytics programs.

Twitter/Norovirus Project:

Crowdsourcing of keywords from Yammer, refining of keywords using correlations and factor analysis, construction of logistic regression model, refining on model using Receiver Operating Characteristics (ROC) curves.

Commonwealth Games:

Development of key words across 15 different sub searches on Commonwealth Games related content, data cleaning including the removal of non-relevant content, adaptation of search terms, reallocation of sentiment manually, analytics e.g. volume, most linked media, top hash tags used.

STAGE 5: IMPLEMENTATION/ PUBLICATION/ACTION

Considerations:

Whilst research in government should never be research for research's sake, the infancy of social media research means than work will likely be exploratory.

Successful outcomes or otherwise should be communicated, to allow the interested communities to build on this work.

Twitter/Norovirus Project:

The model will be used to predict when significant rises of lab reports are due to occur. This will notify communications teams in the FSA, Department of Health and Department for Education who will use the early warning to decide whether to enact an intervention in order to reduce cases.

Commonwealth Games:

The work forms part of the Commonwealth Games Delivery and Legacy Evaluation reporting following the Games.

STAGE 6: EVALUATION

Considerations:

Because of the immaturity of social media research versus other research methods, the main focus of any evaluation is likely to be exploring what value to the project social media research added versus traditional methods.

A component of this will be confirming whether or not social media research was specifically needed to respond to a business or citizen need.

Twitter/Norovirus Project:

Has the intervention reduced cases compared to normal seasons? Has the intervention reduced cases more than if an intervention had been enacted without an early warning tool? Does using social media provide any added value to other data sources in an early warning tool?

Commonwealth Games:

It was part of the transfer of knowledge from these Games to the next host and part of the wider Commonwealth Games Legacy Evaluation. It is hoped that social media research reflects views of the wider population on the Games.

STAGE 7: BUSINESS AS USUAL

Considerations:

Many techniques used to analyse social media data are niche skills possessed by analysts. Thought should be given to how these skills can be transferred or embedded to others.

The fast moving nature of social media means that climates for posting, platforms to post on and thus the effectiveness of research will change regularly. Researchers should be proactive in periodically checking what the best approach is, in order to maximise potential.

Twitter/Norovirus Project:

If reductions can be proved to be notable, then this will become background analysis which will be used each Norovirus season (October to March).

Commonwealth Games:

Social Media will continue to be an important part of future Commonwealth Games and other multi-sport events learning.

4.3. Social media data collection

There are a range of ways in which social media data can be collected and formatted. This is determined by the types of behaviour being researched and the platforms which are used as source material (Wilson, Gosling, & Graham, 2012). One significant difference in the collection of social media data is the development of automated technological tools that can collect, clean, store and analyse large volumes of data at high velocity. Indeed, in some instances, social media has the potential to generate population level data in near real-time. These characteristics differentiate social media derived data from material originating from traditional research methods; these are either more intensive or time-bound, as outlined in the table below (Edwards, Housley, Williams, Sloan, & Williams, 2013).

		Data collection	
		Real-time	Timebound
Data analysis	Intensive	Ethnography	Interviews
	Extensive	Social Media Specific	Surveys and Experiments

One frequently used option to collect social media data is to purchase the data from an authorised reseller, for example Gnip facilitate the purchase of customised Twitter datasets. If analysts possess the appropriate computer programming skills, Application Programming Interfaces (APIs) are an effective way of collecting complex datasets, with data samples often available free of charge. APIs provide a set of building blocks, such as protocols, tools and information, to allow programs and processes to be built. Whilst APIs can be used in many circumstances (e.g. Google Maps has an API to allow programmers to embed a Google Map, onto a website), in terms of accessing social media data, APIs will typically provide a real time link to the relevant data, which can then be analysed and/or visualised.

4.4. Social media data analysis

Given the variety and potential size of social media data, new and dynamic approaches to existing quantitative and qualitative research techniques are being developed. The Social Media Methodology Spectrum on page 18 outlines the broad range of social media-derived data types that can help answer social research questions. The diagram also aligns traditional and new research methods that could be useful analytical tools for these different data types.

<u>Quantitative approaches</u>

A wide range of insights can be gained from interrogating the frequency of discrete or categorical variables in datasets:

- **Volume Analysis:** This is the simplest way of looking at volumes of data, either associated with certain groups (e.g. platform users from a certain demographic) or with volumes of mentions of a particular keyword within a fixed timeframe.

- **Relationship Analysis:** Still based on volumes, but looking at interactions between users, this will look at the number of fixed relationships a user has (e.g. friends on Facebook, followers on Instagram, etc.) or the number of responses to a post (e.g. retweets/quotes on Twitter, comments on Tumblr/YouTube, etc.). This is useful for engagement analysis and is often a key infographic displayed on social media analysis platforms.

- **Correlations:** By comparing a social media dataset with another dataset across time or some other independent variable (e.g. location, age), correlations can be obtained. This forms the basis of using social media data as an indicative or predictive tool.

- **Regression/Classification:** By using correlated data, models can be constructed in order to predict categories or values which dependent variables will have. This is particularly useful as a tool to inform interventions. The Food Standard Agency's current work of predicting significant rises in Norovirus using Tweets about Norovirus symptoms in order to inform an intervention intended to reduce cases is a good example of this.

- **Clustering:** This approach is essentially a quantitative version of segmentation; it uses an algorithm to assign data to a cluster where all items in that cluster have similar characteristics. This is useful when looking at demographic characteristics of users talking about topics.

- **Geographical Information Systems (GIS):** The spatial element often provided with social media data (IP addresses from computers/GPS locations when posting from mobile devices) means that datasets can be mapped to provide a real time or historical representation of the spread of an event (e.g. a protest, illness outbreak, etc).

<u>Qualitative approaches</u>

Given the qualitative nature of much data derived from social media platforms, it is understandable that qualitative methods can render a range of analytical insights:

- **Active/Passive Ethnographic Approaches:** Active or passive ethnographic approaches have historically been used, where the researcher collects data from a social media platform. For example, researchers have engaged in Second life and

special interest chat rooms from the position of observing or participating users (Williams, 2006).

- **Segmentation/Group Identification:** Researchers can also actively engage with social media data as an additional source that complements and augments existing qualitative research. Typically this sort of analysis could be achieved through a social media analysis platform such as Cosmos, Pulsar or Topsy. One example of this approach is the identification of hard-to-reach groups for interviewing or supplementary surveying.

- **Thematic Analysis:** Social media data can be coded and thematically analysed to identify the emotive character of content or classify -hierarchical data to identify areas of significance within them (Thelwall, 2008). One example of a tool for achieving this is NCapture, which adapts the NVivo program to enable the extraction and segmentation of Twitter data from a web browser (Edwards, Housley, Williams, Sloan, & Williams, 2013).

- **Sentiment Analysis:** Pre-existing algorithms can be tailored to conduct automated sentiment analysis and identify if text constitutes a positive or negative view (Volkova, Bachrach, Armstrong, & Sharma, 2015). This approach is still limited in its ability to gauge sentiment when processing complex subjects or ambiguous, inconsistent or culture specific material such as colloquialisms or sarcasm.

- **Graphical Media Analysis:** Image and video content is an increasingly important form of online interaction and can provide important data on areas of interest such as recreational drug use or food safety practices (Morgan, E, Snelson, C. Ellison-Bower, P, 2010). In addition to direct semiotic analysis of such material, the manner and reasons for its sharing require more in-depth analysis, particularly given the risk that they become displaced from their original context.

Combining approaches: a methodological spectrum

Social media can be considered ‘qualitative data on a quantitative scale’ (D’Orazio, 2013). As such, traditional methodological boundaries are increasingly blurred when considering the most appropriate tools for addressing a research question. Biographical data such as user occupation or lifestyle interests, can often be clearly defined and statistically correlated with particular patterns of behaviour. However, any inferences drawn from such analysis can be strengthened by a randomised qualitative sense-check of how users have input such data in a non-standardised manner. In the instance of engagement, an experimental approach could be employed to assess the variance in response to identical content which is posted on multiple platforms.

Combining different methods, including offline methods, can also help to establish wider contextual meaning. For example, using Twitter hashtags as sampling criteria leads to the self-selection of cases as only users posting a particular phrase will be studied. In this instance, approaches such as network analysis or additional qualitative case studies could help to establish how counterfactuals can be covered or develop an understanding of the use of a particular hashtag or how the use of a particular hash tag changes over time and between different groups (Tufecki, 2014).

Many machine learning algorithms are being built to do what traditionally would be performed by a human in many of the analytical approaches, both quantitative and qualitative. It is important to understand that whilst the mention of 'machine learning' and 'algorithms' may hint at quantitative techniques, all these algorithms are doing is replacing the work a human would do – the analytical output is still qualitative. The techniques used and the processes taken to apply them must be viewed separately.

Social Media Methodology Spectrum

QUANTITATIVE

QUALITATIVE

Units of volume and frequency
- Number of followers/friends
- Number of users
- Rates of use and interaction
- Searches

Number of reactions
- Views
- Comments
- Likes/endorsements
- Retweets/Quotes

Volumes per unit time

Scores/Other Ordinal Rankings

Deletions

Biographical data
- Age, Name, Gender
- Nationality, Residence
- Occupation or qualifications
- Lifestyle activities or interests

Location
- Latitude / Longitude
- Settlement/Address

Textual Semantics
- Keyword content from posts
- comments on primary posts
- Hashtags

Influencing
- Patterns of reaction

Visual and Audio content
- Photo tags
- Media tone and content

Tone and Sentiment
- Emotions and feelings
- Tone and opinion

Influence and Clout
- Topics of discussion/search

ASSOCIATED SOCIAL RESEARCH METHODS

- **Regression Modelling**
- **GIS**
- **Correlation and ANOVA**
- **Descriptive Statistical Tests**

- **Network Analysis**
- **Semantic Analysis**
- **GIS**
- **Pseudo-Experiments**

- **Semantic Analysis and Thematic Codification**
- **Ethnographic Observation**
- **Active Research**

4.5. Challenges and opportunities

Social media research can present a number of challenges and opportunities for research validity and reliability. As a result, careful research design is required with clear research objectives and questions and the appropriate selection of analytical tools. If initial findings are statistically significant then they should be verified using at least one more distinct additional dataset which has been collected at a different time, using different methods, or on a different platform. If not, it may be necessary to reduce the scope of a study or reframe its central hypothesis to address a more specific aspect of human behaviour on a given platform. As with all research, any analytical limitations and considerations should be placed alongside findings to ensure research findings are not inappropriately used.

Whilst not an exhaustive list, some of the key attributes of social media data which may have implications for validity and reliability are discussed below:

- **Social media users:** Users of social media are not representative of populations (Ruths & Jurgen, 2014). As such, biases will exist and it may be difficult to infer findings to the general population. However, this characteristic may be helpful if the research is focused on a group that is particularly active on a social media platform, and there is the potential to obtain data on groups who do not tend to respond to other research methods. Platforms can host numerous automated 'bots', and professionally managed accounts which pose as genuine human users. Large studies should therefore attempt to filter out results from such anomalous sources during analysis.

- **'Organic' real-time data**: Social media data is seldom created for research purposes. Whilst this means large amounts of data may be irrelevant or in a format that is difficult to analyse, it has the advantage of removing researcher bias, recall issues and participant burden. However, social media users are still engaged in performative social interactions which will produce observer effects. At the individual level, this can relate to reputation concerns, whilst users with shared interests may collectively promote material. If a researcher was to make their presence on social media known they should be aware of the potential changes to behaviour that could occur.

- **Online behaviour versus offline behaviour:** As a performative action, it is difficult to infer how reflective a user's online behaviour is of their offline behaviour without information on them from other sources. It is generally held that both positive and negative online feelings are over-stated and that interest in a topic may actually not translate into further actions (the value-action gap). This is partly due to the 'echo effect' that is produced as a social media platform skews the content viewed by

users according to their preferences, thus limiting their exposure to alternative viewpoints and encouraging group-think. Further in-depth studies of samples of users could provide insight in this area.

- **Private ownership of platforms and data:** Access to data is governed by companies that own the data and their privacy agreements with users. Many companies do not share a wide range of details about the social media interactions which occur on their platforms. Therefore, whilst interactions can be observed and analysed, important nuances or context may be missing.[2] In addition, there is opacity as to how datasets have been created. Platforms change functionality, settings and popularity regularly, which affect the way data is collected and analysed. Whilst there are frequently positive developments in the opportunities available with datasets (e.g. new variables), ensuring consistency in research across longer timeframes can be problematic.

[2] A key example of this is the denominator of how many users are exposed to material but do not react to it. Without this reference point it is difficult to statistically test the significance of some observed behaviour.

5. Ethical considerations

All social research should be guided by ethics principles[3] and research involving social media is no exception. There are some unique ethical challenges raised by social media research and an emerging body of literature is seeking to address these (Evans et al, 2015; NSMNSS, 2014). Using the Government Social Research (GSR) ethical principles as a framework[4], we have set out some of the key ethical points of considerations below. However, given this is an evolving and complex research area, this is intended as useful guidance and should not replace sound professional judgment / advice from relevant colleagues, e.g. ethical sponsors in each Department. We recommend that a full ethical review should be undertaken for any social media research project.

Core principle 1: Sound application and conduct of social research methods, and interpretation of the findings

Government social research should use appropriate, high quality methods which meet a genuine research need. Any research outputs should be appropriately communicated.

Considerations for social media research:

- Social media techniques / data should be the most appropriate method to use to answer the research questions and not used on any other basis
- Methods should be used professionally and appropriately. Given the infancy of social media research, researchers may need to make methodological decisions based on theory rather than prior practical experience.
- As social media methods often make use of existing, publicly available data, the burden on respondents can be reduced. However, researchers should consider the implications of using existing data on ensuring data is robust and valid. Quality assurance will be particularly important to ensure quality outputs[4]
- Where appropriate, researchers should make details of their project publically available, including the research purpose and the data being used.

[3] The Academy of Social Sciences adopted five guiding ethics principles for social science research in 2015 and has commended these as a foundation for the development of a common framework for research ethics across the social sciences. https://www.acss.org.uk/developing-generic-ethics-principles-social-science/academy-adopts-five-ethical-principles-for-social-science-research/

[4] https://www.gov.uk/government/organisations/civil-service-government-social-research-profession

[4] See the Government's Aquabook for more information: https://www.gov.uk/government/publications/the-aqua-book-guidance-on-producing-quality-analysis-for-government

Core principle 2: Participation based on informed consent

Participants in any research study involving primary data collection must be asked for their consent to take part. It should be clear that participation is voluntary and that they have the right to refuse to answer individual questions or to withdraw from the research process at any point. Any secondary analysis should be conducted within the bounds of the original consent. Covert research must be subjected to an independent ethical review and legal advice must also be sought.

Considerations when using social media:

- The terms and conditions which users agree to when signing-up to a social media platform may cover the use of their data for research. Whilst this can provide a legal gateway, researchers should consider whether specific research projects reasonably meet user expectations of the collection, analysis and use of their data.
- Individual informed consent is impractical for research involving large datasets. In these cases researchers should ensure data use is in line with terms and conditions and care should be taken to protect the identity of users (see principle 5).
- If individual informed consent is sought, researchers should consider appropriate ways to contact users. Thought should be given to the relevant Department's reputation and public trust of the government and its research operations.
- Users can post data to social media platforms and subsequently delete it. If that data has been retrieved by a researcher before deletion, it is not clear whether the user's initial consent for their data to be used remains intact. Depending on the sensitivity of the data and analysis researchers should agree up-front how to manage this issue. For example, it may not be necessary to delete the count of a post from a time series, but it may be unethical to quote an individual post which has since been deleted.

Core principle 3: Enabling participation

Consideration should be given to any barriers to participation and steps should be taken where necessary to enable participation (e.g. provision of assistance with any participation costs). In particular, the effect of research design on groups such as ethnic minorities, those with caring responsibilities and those with physical or mental impairment should be considered.

Considerations when using social media:

- Certain groups are more likely to use social media than others and significant differences can exist between social media platforms. Researchers should consider whether any groups are being inappropriately excluded given the nature of the research questions. Actions to enable participation where possible (e.g. collecting data through a number of different platforms).

Core principle 4: Avoidance of personal and social harm

The physical, social and psychological well-being of research subjects and researchers should be protected at all stages of the research process. This included minimising intrusion and researchers should respect participants' privacy. An objective assessment of potential personal or social harms should be included in any research proposal.

Considerations when using social media:

- Researchers must consider privacy settings to understand whether data is public or private. Any research involving private content should only be conducted with explicit informed consent from the user.
- Whilst it is not possible to guarantee that personal data will not be collected, the collection of unnecessary personal data should be minimised. This could include limiting the amount of information collected, or stripping out personal or irrelevant data after collection.

Core principle 5: Non-disclosure of identity

Efforts should be made to protect the identity of participants throughout the lifecycle of the research. Any personal information should be safely secured and managed.

Considerations when using social media:

- There can be no guarantee of full anonymity within social media research. Aggregated findings may provide anonymity but any raw data will not be anonymous even if stripped of the author field as the content could be searched for online. In addition, removing the author field may be problematic with some social media platforms. For example, Twitter's terms of use state that the individual's username must always be displayed with the tweet text.
- It may be possible to 'mask' content by altering the content for that the meaning is maintained but it is not traceable back to the source but it is unlikely this will guarantee anonymity. There is a trade-off here between accurately quoting what was said and by whom, for academic scrutiny
- If researchers wish to include verbatim content, they should consider contacting social media users to ask them is they would be happy for their content to be cited.

PROJECT STAGES

Concept/Project Scoping → **Performing the Research** → **Reporting/Using to Advise**

ETHICAL CONSIDERATIONS

Policy/Analyst Engagement:

- Is social media the most appropriate/useful data source?
- Does using social media add value to current knowledge?

Data:

- Does the project allow ease of participation from those relevant/interested?
- What is the level of informed consent expected by users of the chosen social media platform?

Tools/Methods:

- How representative of society/the social media population is the tool/method?
- How intrusive/much of a burden is the tool being used?

Data:

- What is the level of informed consent expected by users concerning the data collected?
- How is the data stored and managed?

Publishing Findings/Transparency:

- At what level has the data been anonymised to and why?

Informing Interventions:

- Are interventions reacting to known/ specific users' on social media?
- If answer to above question is 'Yes' are the segmented interventions supportive or punitive?
- Would the consent of users change if they knew their data was being used for this intervention?

ASSOCIATED CORE PRINCIPLES (CPs)

- **CP 1**: Sound application of methods/analysis
- **CP 2:** Informed consent
- **CP 3:** Enabling participation

→

- **CP 1**: Sound application of methods/analysis
- **CP 2:** Informed consent
- **CP 4:** Avoidance of harm

→

- **CP 2:** Informed consent
- **CP 4:** Avoidance of harm
- **CP 5:** Non-disclosure of identity

6. Future considerations

Since the emergence of social media into the mainstream in the early-mid 2000s, entire platforms have entered, conquered and subsequently fallen out of the market as smarter and better-targeted platforms took hold. The speed of this process considerably limits the ease with which it can be predicted; whether the market is maturing and settling; or whether there will be further major market disruptions where entire technologies and platforms are replaced. If a platform were to change its API or core functionality, cease to exist, or its users were to systematically change the way in which they engaged with the service, any research based on this platform could quickly become irrelevant. It may be difficult or impossible to transport existing research onto new or changed platforms.

Recent technological progress in social media has been dominated by extensions to a wider range of devices (e.g. mobile phone, home appliances, cars and 'wearable technology'). These extensions have the potential to fundamentally change and increase the amount of data available on users through new sensors (e.g. location, microphones, cameras). When linked to social media identities, data covering a larger portion of users' lives and opinions may become available, potentially allowing researchers to better understand the demographics of their 'respondents'. The increasingly connectedness of data also raises new ethical considerations, as it could increase the *distance* between the point-of-consent and the data, which the user technically consents to be made available through APIs. Researchers should consider what the user's *informed* consent process is; beyond the legal technicalities of accessing the data through an API.

As social media platforms change and evolve, so too will public perceptions towards the use of social media for research. Researchers will need to engage with this as part of their ethical considerations. A Government Digital Service project as part of the wider Government Data Science Partnership is currently exploring the public's views on data science work, of which social media research can fall into; understanding how projects are perceived by citizens will be important in shaping the sort of project we undertake and how we undertake them in the future. The findings from this project will be published in summer 2016, and will be of relevance to social researchers across government. [5]

[5] Further information on the project, including a high level draft of the framework can be found at https://data.blog.gov.uk/2015/12/08/data-science-ethics/

7. Bibliography

Bollen, J., Mao, H., & Zeng, X. (2010). Twitter mood predicts the stock market. *Journal of Computational Science*, 1-8.

Davenport, T. H., & Dyché, J. (2013, May). *Big Data in Big Companies.* Retrieved June 9, 2015, from SAS Institute: http://www.sas.com/resources/asset/Big-Data-in-Big-Companies.pdf

Disson, J. & Baker, J. (2014). *The social media challenge within the Food Standards Agency.* SRA Social Media in Social Research, 4th Annual Conference, 16th May 2014.

DOMO. (2015). *Data Never Sleep*. Retrieved June 9, 2015, from DOMO: https://www.domo.com/learn/data-never-sleeps-2

D'Orazio, F. (2013). *The future of social media research: or how to re-invent social media listening in 10 steps*. Retrived January 26, 2016, from Pulsar: http://www.pulsarplatform.com/blog/author/francesco-dorazio/page/3/

Edwards, A., Housley, W., Williams, M., Sloan, L., & Williams, M. (2013). Digital social research, social media and the sociological imagination: surrogacy, augmentation and re-orientation. *International Journal of Social Research Methodology*, 245-260.

Evans, H., Ginnis, S., Barlett, J. (2015). *#SocialEthics a guide to embedding ethics in social media research.* Retrived February 09 2016 from Ipsos Mori: https://www.ipsos-mori.com/researchpublications/publications/1771/Ipsos-MORI-and-DemosCASM-call-for-better-ethical-standards-in-social-media-research.aspx

Facebook. (2016). *Company Info*. Retrieved February 12, 2016, from Facebook: http://newsroom.fb.com/company-info/

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, 59-68.

Morgan, E, Snelson, C. Ellison-Bower, P. (2010). Image and video disclosure of substance use on social media websites. *Computers in Human Behavior*, 1405–1411.

New Social Media, New Social Science (2014). *New Social Media, New Social Sciene…and New Ethical Issues!* Retrived February 09 2016 from NSMNSS: http://nsmnss.blogspot.co.uk/2014/02/new-social-media-new-social-science-and.html

Oxford English Dictionary. (2015). *Big Data*. Retrieved June 16, 2015, from Oxford English Dictionary: http://www.oed.com/view/Entry/18833#eid301162177

Ruths, D., & Jurgen, P. (2014). Social media for large studies of behavior. *Science*, 1063-1064.

SAS Institute. (2015). *What is Big Data?* Retrieved June 9, 2015, from SAS Institute: http://www.sas.com/en_us/insights/big-data/what-is-big-data.html

Scottish Government Social Research (2015) *Analysis of XX Commonwealth Games Host Broadcast Coverage, Online Media and Official Digital Channels*. Retrieved February 15 2016 from Scottish Government: http://www.gov.scot/Resource/0048/00482015.pdf

Sloan, L., Morgan, J., Burnap , P., & Williams, M. (2015). Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data. *PLoS ONE*.

Thelwall, M. (2008). Fk yea I swear: cursing and gender in MySpace. *Corpora*, 83-107.

Tufecki, Z. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media, 2014.*

Twitter. (2015). *Company*. Retrieved June 9, 2015, from Twitter: https://about.twitter.com/company

Volkova, S., Bachrach, Y., Armstrong, M., & Sharma, V. (2015). Inferring Latent User Properties from Texts Published in Social Media. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (pp. 4296-4297). Austin: AAAI.

Williams, M. (2006). *Virtually Criminal Crime Deviance and Regulation Online.* London: Routledge.

Wilson, R. E., Gosling, S. D., & Graham, L. T. (2012). A Review of Facebook Research in the Social Sciences. *Perspectives on Psychological Science*, 203-220.

8. Annex – Further resources

Tools

There are a broad and growing range of tools available for gathering and analysing social media data, the choice may seem overwhelming and it is likely that a fairly limited combination will meet most needs. Some of them are free to use whilst others can entail fairly substantial subscriptions; inclusion of a tool in this section is in no way an official endorsement. As we look to update the guide, please do get in touch (through the contacts at the start of the document) with feedback and recommendations of additional tools to include.

Text Analysis		
DiscoverText	Cloud-based text analysis tool allows different data (including, but not limited to Twitter data) to be stored in different project folders.	https://discovertext.com/
NCapture	NVivo 10 add-on allows users to capture data from their web browser, such as segmented Twitter data, for analysis.	http://www.nvivo10.com/
SentiStrength	Freeware analyses short extract of texts in order to determine positive or negative sentiment, and the strength of such sentiment.	http://sentistrength.wlv.ac.uk/
Mozdeh	Gathers and runs time series analyses all or subgroups of tweets based on topics which can be selected by the user. Also works with SentiStrength	http://mozdeh.wlv.ac.uk/
Network Analysis		
Gephi	Well documented open source program, which allows visualisation and exploration of networks (including social media networks).	http://gephi.github.io/
Node XL	Free Microsoft Excel add-on which enables interactive exploration of network graphs. Not necessarily for social media, but easily applied.	http://nodexl.codeplex.com/
SocSciBot	Produce statistics and diagrams explaining the interlinking of pages on websites; can be used to run limited analyses of the text in the websites	http://socscibot.wlv.ac.uk/
Data Acquisition and Management		
Tweet Archivist	Allows users to create 'archives' which tweets are saved to, in a format which can be downloaded and saved to the desktop (e.g. in MS Excel format). Some simple data	https://www.tweetarchivist.com/

	visualisations are also provided by the service.	
ScraperWiki	Can be used to visualise and analyse a range of data from a number of web-based sources.	https://scraperwiki.com/
Multipurpose Platforms[6]		
Pulsar	Time series volumes, topic and sentiment analysis. Strong emphasis on visualisation. Analyses data from internet/social media.	http://pulsarplatform.com/
YouGov SoMA	Overlays demographic data with comments made on social media platforms such as Twitter and Facebook, to identify what an audience is 'hearing'.	http://research.yougov.co.uk/services/soma/
Crimson Hexagon	Geared towards qualitative analysis – sentiment and text analysis and user backgrounds.	http://www.crimsonhexagon.com/
Topsy	Allows users to identify social trends on the web and Twitter including comparing tweet volumes over time for specific terms.	http://topsy.com/
Ripjar	Real-time time series along with focus of geo-location infographics. Customisable graphics.	http://ripjar.com/
Traackr	Focus on influence and relationship analysis between stakeholders.	http://traackr.com/

Resource Centres

Social Media expertise has grown fastest in the private sector with a focus on consumer and brand analysis however; there are a growing number of collaborative academic efforts to engage with more complex social phenomena. These can provide a wealth of resources on the theory and practice of online social research and social media-specific research.

The Centre for the Analysis of Social Media (CASM)	Based on a partnership between Demos and the Text Analytics Group at the University of Sussex. It produces social media research on a broad range of topics to provide new political, social and policy insights.	http://www.demos.co.uk/projects/casm
The Collaborative Online Social Media	Funded by the ESRC and Joint Information Systems Committee (JISC), this cross disciplinary centre studies the methodological, theoretical, and policy dimensions of Big 'Social' Data.	http://www.cs.cf.ac.uk/cosmos/

[6] Time series and real time analysis are essentially integral to all of these platforms

Observatory (COSMOS)		
Digital Methods Initiative	Hosted by the University of Amsterdam, this online academic partnership works with universities across Europe to Generate online courses and case studies of digital research approaches, including Social Media.	https://wiki.digitalmethods.net/Dmi/DmiAbout
New Social Media, New Social Science	This group was established by NatCen and SAGE to look at approaches and issue s within social media research, in particular ethics. Natcen have also hosted various workshops and briefing sessions on social media research, one of which produced a useful summary report.	http://nsmnss.blogspot.co.uk/ http://www.natcen.ac.uk/media/282288/p0639-research-using-social-media-report-final-190214.pdf
Oxford Internet Institute	This academic institute covers a broader area than just social media but offers a wealth of information on methodological developments and considerations in areas such as sampling bias.	http://www.oii.ox.ac.uk/research/
First Monday	This free online academic journal focuses on internet research issues. Obviously a wealth of other key academic papers is published in other journals but this title is regularly of interest.	http://firstmonday.org/ojs/index.php/fm/index
Visual Social Media Lab	The Visual Social Media Lab (@VisSocMedLab) brings together a group of interdisciplinary researchers interested in analysing social media images. It was set up as part of a research project, 'Picturing the Social: transforming our understanding of images in social media and Big Data research', which is funded by the ESRC's Transformative Research programme.	http://visualsocialmedialab.blogspot.co.uk/