

## **Blurring the Boundaries - New Social Media, New Social Science**

Dr Grant Blank, Oxford Internet Institute

Under the joint sponsorship of NatCen Social Research, Sage Publications and the Oxford Internet Institute the “New Social Media, New Social Science” (NSMNSS) network has been exploring the challenges of new social media (NSM). These challenges are described here as a series of strengths and weaknesses. Most challenges are particularly important for quantitative research. NSM provides rich settings for qualitative researchers. In NSM the qualitative focus on meaning and developing understanding through participation and progressive collection of data remains consistent with traditional approaches. Understanding meaning is not simplified by digital data. In general, NSM has had less impact on qualitative research.

### **Advantages of social media data**

The first thing everyone notices about social media data is that they are really easy and inexpensive to collect. Since the data arrive already in digital form, they are also relatively easy to prepare for statistical analysis. For qualitative researchers NSM offers a wide variety of forms of data, including text, video, audio and photos. Since the data are already in digital form they are much easier to gather, store and manipulate, compared to paper. Social media often record actual behavior or actual communication in the form of Tweets, e-mails or blog posts, so researchers don't have to depend on self-reports, and the uncertain memory of respondents. This improves reliability and simplifies certain measurement issues. Ease of collection means that researchers can often get data on entire populations. An entire population of tweets, email messages or Wikipedia edits can be collected with no sampling is required. There are no issues of biased respondent selection or assembling a sampling frame. A large population means that even small subgroups can be studied. Since data collection is simple, researchers can often collect longitudinal data (at least over periods of days, weeks or months). This often improves their ability to detect certain causal processes. In many cases, data can be obtained without the subject being aware they are being observed. This raises ethical issues that are discussed below, but there is no concern that subjects changed their behavior because they were observed.

Respondents can be sampled via NSM and asked questions online using well-established survey techniques. The low cost and easy access allow truly enormous samples, possibly with Ns in the millions. Since there is no interviewer, there are no interviewer effects. Online questionnaires allow complex skip patterns that are easy to control. Since there is no problem with delays, people can consult their records at leisure.

This brief summary indicates some of the strengths of research using NSM. They are offset by a number of weaknesses. Since good research requires taking into account both strengths and weaknesses, we turn to the problems of social media data.

### **Disadvantages of new social media data**

We will begin by talking about sampling problems. A sampling frame is often impossible to create. Except in certain limited circumstances it is impossible to get a definitive list of email addresses, web pages, Tweets, etc. This means that no unbiased sample is possible and most samples using NSM are convenience samples. A related problem is that non-users are unreachable. Since only about 50-60% of the British population use NSM, a large part of the population will never be included in a NSM-based sample. Response rates on NSM questionnaires are poor, usually in low single digits. When someone fills out an online questionnaire, the researchers does not know who it is since who answers is uncontrolled. Finally, it is hard to get reliable responses to knowledge questions, since respondents can just use Google.

There are equally important problems when gathering data on whole populations. To begin with, a researcher may have data on an entire population but it may not be a very interesting population (like the population of Twitter users). Many NSM datasets are collected because they are easy to collect, not because they are important. Compounding this issue is that fact that researchers know little about the population. In particular, they do not know basic demographic information like age, gender, education or marital status of respondents. Since these have an important influence on behaviour, researchers suffer a major loss of ability to understand their results. Some things are simply unobservable. Meaning is an example. Although researchers can observe behaviour directly, it is often difficult to infer what the behaviour means to participants. Attitudes are also unobservable. Since both meaning and attitudes exert a major influence on behaviour, not to know them will weaken prediction and some alternative explanations for behaviour cannot be ruled out. Finally, hard to reach populations like the elderly, the poor and the uneducated remain hard to reach online.

In summary, NSM data are *broad* data in the sense of many subjects, but *shallow* data in the sense that we know little about each subject.

### **Ethics**

If the social sciences are newly rich in data, the fly in the ointment is ethics. There are new and serious ethical problems associated with social media data. Throughout the social sciences, ethics is often based on confidentiality. But anonymity is weak in social media. It is all too easy to Google an anonymous quote to break the anonymity. Furthermore, ethics committees may not have a good grasp of what is different about social media, and they may apply rules based on face-to-face interaction. This works badly. On one hand some committees are too easy and so compromise privacy in social media. Others are so fearful that they don't allow research in social media at all. The best social media ethics guidelines are for marketing from organizations like CASRO and ESOMAR. Social science guidelines are not as complete. This generally reflects a lack of consensus about such ethical issues as lurking without identifying yourself as a researcher or collecting data from listserv archives.

### **Conclusion**

There is much more social media data in our future. It is too easy to collect and too easy to analyse for it to be ignored. But good research is a dialog between ideas and evidence. As we think about data we need to pay attention to the other half of the dialog: the ideas. A

key issue is embodied in the question, how many times have you thought about a theory or a hypothesis and said to yourself, “If I could just collect a million cases I would have the answer”? In general the social sciences don’t have such theories. Effective use of NSM data requires the development of theories that will benefit from large amounts of data. There is theoretical work to be done to take full advantage of our newfound data riches.

**Further Information**

ESRC National Centre for Research Methods: <http://www.ncrm.ac.uk/>

NCRM Networks for Methodological Innovation: <http://www.ncrm.ac.uk/research/NMI/>

Digital Methods as Mainstream Methodology: <http://digitalmethodsni.com/about/>

Blurring the Boundaries – New Social Media, New Social Science:

<http://nsmnss.blogspot.co.uk/>