Social research and Big Data – the tension between opportunities and realities

Kari Steen-Johnsen and Bernard Enjolras
Institute for Social Research
kari.steen-johnsen@samfunnsforskning.no
bernard.enjolras@samfunnsforskning.no

In a passionate argument for the idea that social research must adopt social transactional data generated through new information technologies and new analytical techniques, Savage and Burrows claim that:

> both the sample survey and the in-depth interview are increasingly dated research methods, which are unlikely to provide a robust base for the jurisdiction of empirical sociologists in coming decades. (Savage and Burrows, 2007: 885)

Savage and Burrows base their claim on the argument that digital data on social transactions is data about actual events, while also being data that pertains to entire populations. While the survey depends on representative samples and makes predictions based on such samples, those who analyze web data have direct access to complete data about actions and statements. In other words,

analysts of digital transactional data, or what has frequently been termed «Big Data», evade the problem of representativeness: they can provide actual descriptions of peoples' actions and infer future actions from these. This has proven a strong tool for predictions, as exemplified by how Amazon.com is able to make book recommendations to customers based on the choices made by thousands of other customers. In a situation where response rates are falling, data from social transactions provides a powerful alternative. The argument against the qualitative in-depth interview is that this method is incapable of grasping the complexity that modern life entails. According to Savage and Burrows, this complexity can be better grasped in its entirety by studying large numbers of cultural expressions – pictures, videos and texts on the Internet.

There is no doubt that the use of Big Data in research presents researchers with new opportunities for analyzing social phenomena. Yet the use of such data also has its limitations, and introduces a set of new ethical and practical challenges. Both the opportunities and challenges are not only closely linked to the very nature of data, but also to how ownership and access to data are regulated.

In this article, we will attempt to shed light on the role of research in a field of tension between the new opportunities data offers, the ethical considerations that are necessary when a person carries out research and the limitations that exist in the regulation of and access to digital data. Underlying our considerations is the realization that the way in which we as researchers approach this field of tension has consequences. When we study social transactions through Big Data, we are studying a social reality. Through research, we participate in both constructing a social reality, such as the digital public sphere, and giving society insight into what its social reality is, both of which can have social consequences.

First, we want to describe what characterizes digital transactional data and what kinds of opportunities it offers to research. We will

use the term Big Data in order to underscore the new analytical opportunities embedded in the characteristics of digital data, and use social media data as our main case. Then we wish to say something about the new ecosystem that has emerged around the production, gathering and analysis of digital data, and how it changes the research premises for the production of knowledge. We have borrowed the idea that the use of Internet data must be understood as being a part of the growth of a new ecosystem from boyd & Crawford's article «Critical Questions for Big Data» (2012). By using the term *ecosystem*, we emphasize the idea that it is important to understand the use of such data in a context, where various formal and informal actors position themselves, compete and influence each other in the production and use of Internet data. At the end of this article, we will discuss the role of research, both in relation to representing the social reality that is created by digitalization, and in relation to the fact that research itself is an actor with a place in a new ecosystem that is linked to the production of knowledge.

Big Data – what is it and what can it be used for?

The amount of available digital data about people has exploded in recent years. This has to do with mundane status updates on Facebook, videos posted on YouTube, and Twitter posts that are available to anyone who wants to read them. It also pertains to data from purchases, both those that are made on the Internet and those that are made by credit card. Other examples of digital data include data from Google searches and data that logs phone calls.

The term «Big Data» is a collective term for data that is of such a scope that more data power than normal is required in order to collect and analyze it (Manovich, 2011). The term is often used not only to simply describe the data itself, but also to describe the new

technical, legal and ethical issues such data gives rise to. Big Data are regularly collected in a number of areas, related to natural, technical or social phenomena. For present purposes we understand Big Data to be data containing person-related information. What phenomena characterized in terms of Big Data thus understood have in common with each other is that they involve a registration of actual events, interactions and transactions connected to individuals.

There are two aspects of Big Data in particular that will greatly impact the social sciences. First, transaction data differs from survey data in that such data directly reflect what individuals actually do, instead of drawing conclusions based on individuals' statements about actions. Second, digitalized data combined with cheap data power make it possible to study entire populations, instead of drawing on a selection. Thus, it becomes possible to conduct very sophisticated analyses, and also to predict future actions. In the book *Predictive analytics. The power to predict who will click, buy, lie or die* (2013), Siegel provides examples of such predictions within several areas: within the banking system, in the struggle against crime, and in health-related services. One example that pertains to many people's daily lives is the way in which companies such as Amazon.com are using previous purchase data to market and offer relevant products to the individual user: «Other people who bought this book, also bought …». One example from the field of health is how Google was able to use search trends to predict the development of the swine flu epidemic long before the national health institutes could do so.

The way in which Obama's second presidential campaign used and analyzed data provides examples of both how data may be used to predict behaviour, and how data from different sources can be pulled together and offer powerful analyses.[50] Obama set

50 See http://www.technologyreview.com/featuredstory/508836/how-obama-used-big-data-to-rally-voters-part-1/ for a thorough description of how the Obama campaign worked with Big Data and predictions.

up a separate computer lab that linked data regarding households, previous voter behaviour, previous donations and television use, and then tailored his message to individual voters based on such data. This afforded him the ability to choose where to direct his efforts – and with which message – in a much more efficient manner than would otherwise have been possible using traditional analysis and prediction methods.

The example of Obama highlights an important characteristic of digital transactional data, namely that such data contains several layers of information. Metadata, such as email addresses, make it possible to link the different data that an individual has produced, for example data about purchase transactions and toll transactions. Savage and Burrows (2007) emphasize the importance of the fact that digital transactional data contains information about one's geographical positioning, and claim that many of the classic background variables in sociology can collapse into one variable as a basis for predicting actions. In this argument they rely on a set of studies that have demonstrated that neighbourhood location is a more significant predictor of many outcomes than other person or household-related variables (2007: 892). The possibility of determining one's position grows with the increased use of handheld media devices. In addition, different types of data from various sources are linked together and integrated through social media and the different Internet sites where individuals have user accounts. For example, Facebook integrates several different applications, such as Spotify, Instagram, and Farmville, and the data about activity on the different applications are thereby linked together. A parallel example is the link between Google, Gmail, YouTube, Chrome and Google+.

The result of this linking of data is that it creates new opportunities for assembling very detailed information, not only about individuals, but also about groups and organizations. One characteristic of the

built-in opportunities for action in social media is that they make it possible to establish a social graph – lists of followers and friends (boyd & Ellison, 2011). Data gathered from social media thus contain information about how individuals and groups are linked together. If data about a user is collected on Facebook, data is simultaneously collected about several people in this user's network.

Opportunities and limitations for social research

It would appear that such data represent an obvious enrichment to social research, as they give direct access to people's lives, statements and actions, provide detailed information and can be easily collected. Both access to Internet data and the opportunity to conduct analyses of large datasets based on concrete actions and interactions have caused many researchers to feel that the use of Internet data will revolutionize social research in fundamental ways.

However, in the article «Critical Questions for Big Data», boyd and Crawford caution against believing we can leapfrog over fundamental methodological challenges, such as the issue pertaining to representativeness, when the data that is used is large enough. Analyses of Twitter provide a good example of some of these challenges. Twitter studies have become very popular internationally, especially due to the availability of data. However, questions may be raised as to what analyses of Twitter posts represent. An obvious challenge is that Twitter users only constitute a certain selection of the population. Other issues are linked to the fact that there is no one-to-one-relationship between user accounts and actual people. One person can have several accounts, several people can use the same account, and accounts can also be automated – so-called «bots».

Another problem is linked to the definition of what constitutes an active – and thereby relevant – user on Twitter. According to Twitter, as many as 40 per cent of their users in 2011 were so-called «lurkers», that is, users who read content without posting anything themselves. Brandtzæg showed in his doctoral thesis that the same finding pertained to 29 per cent of those who use Facebook in Norway (Brandtzæg, 2012). The question then is how to get hold of and define a representative picture of social media as both forums for distributing information and as «public spheres». One response to this might be to argue that the use of Big Data makes it possible to analyze social media sites like Twitter or Facebook on an aggregate rather than an individual level, and in this way paint a picture of these social media as public spheres based on whatever and wherever topics are being discussed and distributed.

Another example of the difficulties in determining what constitutes the correct representation of the digital public sphere may be found in our contribution to Enjolras, Karlsen, Steen-Johnsen & Wollebæk (2013). In this book, one of our aims is to study public debate as it takes place in social media. The basis for the book is a web-based, population representative survey that we repeated on two occasions, where we posed fairly detailed questions about where people debate, what kinds of topics they discuss, and how they experience the debate (for example, whether they often debate with people who agree with them).

Based on analyses of the material, we paint a picture of a fairly well-functioning public debate: the debating Internet population is hardly distinct from the population in general when it comes to socio-economic background and political views. Many debate with people they disagree with, few experience getting hurt, and many learn something from the experience – though few change their opinions. We also draw a comparison of political attitudes among

those who debate on the Internet and those who do not, a comparison which is broken down into different forums (Facebook, discussion forums, Internet papers, blogs, etc.). As a whole, there are hardly any differences between those who engage in discussion on the Internet and those who do not. There are differences, however, between debaters on the various platforms. This picture differs significantly from the picture that is sometimes presented in the mass media and the socially mediated public sphere. At the same time, there is little doubt that if we conducted a thorough qualitative examination of the content from selected discussion forums and the debate forums of online newspapers, and studied the political attitudes within them, we would find a different picture. These two approaches would both give valid representations of Internet online discussions, but they would be representative of different phenomena – either the broad picture or the dynamics of particular forums.

This example illustrates the argument that, depending on whether one uses selected web content or representative survey data as the basis for analysis, very different pictures of an Internet phenomenon can be obtained. The same phenomenon is pointed out by Hirzalla et al. (2011) who point out that in studies about whether the Internet equalizes differences between different groups when it comes to political participation, it is often the case that those who conduct case-based research are more optimistic than those who conduct survey research when it comes to their mobilizing potential. Those who study the Obama campaign's social networking sites will most likely find a stronger source for the mobilizing power than those who study the American population's activism through survey-based methods. The representative picture and the picture that is based on significant events thus part ways – but which picture is the most valid in relation to how the digital public sphere works?

Even though the use of Internet data may potentially provide access to complete data and enable researchers to analyze statements and content from a large number of users, this does not eliminate the problem linked to representativeness or the need to interpret and discuss whatever phenomenon a person has captured. In addition, it is important to point out that the assumption of having complete data is hardly ever correct. We can again use Twitter as an example. Only Twitter has complete access to the information in Twitter accounts and the complete set of Twitter posts. Twitter makes Twitter posts available through so-called APIs,[51] which are program sequences that make it possible for outsiders to retrieve Twitter posts. However, it is unclear what underlying factors are at work in such retrievals, for example, on what grounds the selection is made. The selection of Twitter messages can also become uneven due to the fact that Twitter censors certain types of unwanted content.

What characterizes the new ecosystem for the production of knowledge?

In connection with the increased access to digital data, important changes have taken place in the social landscape where research positions itself. Savage and Burrows use the term «knowing capitalism» to describe a new social order in which social «transaction data» become increasingly important, and where such data are routinely collected and analyzed by a large number of private and public institutions. The main point for Savage and Burrows is that research has thereby ended up in a competitive setting. Research is no longer the dominant party when it comes to providing interpretations of society. boyd and Crawford use the concept

51 API stands for «application programming interface» and is a software that can be used to collect structured data, for example from Twitter (boyd & Crawford, 2011: 7).

of a new «ecosystem» to describe the new set of actors connected to the analysis of digital data and the power relationship that exists between them. Several elements of this new ecosystem touch on the potential of social research to represent and interpret society. We will highlight a few of these elements here.

First, the ownership of data has become privatized. While access to data has traditionally depended on an exchange between researchers and individuals who have given their consent, such access now largely goes through large, private companies such as Google, Twitter, and Facebook. Also, several other types of commercial businesses, such as telecommunication companies, banks and online bookstores, are in possession of large amounts of data. This situation creates several challenges for research. The fact that data are owned by private actors makes the access to such data and assessments of the quality of data difficult. Further, a gap arises between these private actors and externally performed research, because the owners of the data have access to complete sets of data. Several of the large Internet actors, such as Facebook and Google, have their own research departments with a high level of competency that enjoy the benefits of having complete access to the data.

As the data are under private ownership, these research departments are not in the same situation when it comes to privacy protection as researchers are, for example when it comes to requirements regarding informed consent. This is because users are required to accept the terms and conditions for the use of their digital data in order to be able to use the service. The result is that actors outside of academic research get a jumpstart when it comes to providing relevant social analyses and interpretations. Besides the fact that researchers inside these private companies are in a position to produce unique analyses, it is a problem that they are not required to let their research be reproduced or evaluated through the examination of data, given the fact that the data are private property

(Enjolras, 2014). Moreover, there is no requirement to let results be examined critically by the international researcher community. In the long run this may lead to a privatization of social research.

Second, the access to new types of data leads to research activity being disconnected from traditional research institutions, which again leads to a situation where «everyone can conduct research». Despite the tendency towards privatization, it is still a fact that a great deal of data is available to the general public on the Internet. A consequence of this is that a much greater number of people can now conduct some form of research or investigations. One trend is that there is a growth in payment services that allow people and organizations to analyze Internet data linked to their own business and to conduct surveys via the Internet. An important example is Google Analytics, which offers tools for analyzing the Internet activity linked to various companies or organizations. The interpretations derived from the research of digital information thus compete with several other narratives coming from individuals, organizations, and analyst agencies. These are not necessarily based on observations of the data's representativeness and quality or on a suitable theoretical foundation.

Third, the development leads to powerlessness among the average Internet user. While digitalization offers new opportunities to many people for gathering information, expressing opinions and analyzing digital data, the users' control over their own data is being reduced. The privatization described above is an important reason for this. To be able to use popular services such as Facebook and Twitter, we have to give these applications access both to basic data and to utilizing these data for their purposes. Such use can consist of generating analyses and statistics, resale efforts and marketing purposes. It can be difficult for the user to understand what rights she or he is really giving up.

As described above, a complicating element is the complexity of digital data. The fact that data exist in many layers, with different

kinds of information, constitutes one such complexity. The fact that different data gathered from different applications and websites are linked together is another. In addition, it is hard to get a full overview of what the network structure in data really entails. For example, those who gather information about you can also at the same time gather information about the users in your network, and vice versa.

The difficulties in understanding both the technical and legal stipulations for the use of an application mean that users lose control over their own statements. A survey conducted in 2009 by Brandtzæg and Lüders in regard to Internet use and privacy protection revealed that many users were concerned about the consequences of sharing personal information on the Internet. The study also showed that most users had limited insight into how social media function and how to handle the privacy settings. boyd & Crawford call attention to the fact that data that have been produced in a certain context may not necessarily be brought into another context just because it is available (2012: 672). The conditions we have outlined here make it important for researchers to take this cautioning seriously. The users produce content within vague frameworks, and do not necessarily have full oversight over what data about them is available, and what it can be used for.

New digital dividing lines

Based on the preconditions that exist in the new ecosystem linked to digital data, it is possible to see the contours of a set of digital dividing lines that will affect what knowledge is being produced by whom (boyd & Crawford, 2012: 674). We claim that academic social research is in danger of losing out in the battle of having access to producing knowledge based on these new types of data. Social research, or at least parts of it, is currently in a disadvantageous

position when it comes to certain disparity structures that will potentially strengthen over time.

An important disparity pertains to access to data and to the resources required to utilize them. As pointed out above, private companies and their analysis departments have privileged access to data. For those who wish to conduct research on such data, access is restricted, and financial resources are required. Digital data from different platforms have been commercialized and can be purchased through so-called «data brokers». Access is thus dependent on the financial resources one has available. Alternatively, one can also collect certain types of data, such as Twitter data, by programming APIs oneself, that is, programs that can fetch data based on certain criteria. However, this also requires resources in the form of data power. As a result of the requirements regarding finances and technological investment, a disparity emerges, not only between private actors and academic institutions, but also between the academic institutions. Elite universities, such as Harvard, MIT and Stanford, have resources to build and equip research environments with technology and resources that allow them to utilize digital data. Smaller universities may not have the same resources. In Norway, we can imagine dividing lines among both the universities and between the university sector and the research institutes.

The use of digital data also creates dividing lines when it comes to competency. In order to utilize such data, competencies in programming, analysis and visualization are required. Such competencies are still a limited resource within the social sciences and the humanities. Building up such competencies requires resources, and larger institutions have an advantage if they are able to connect technical and interpretative environments.

An important disparity-generating dimension has to do with the regulations and requirements that pertain to the use of Internet data. Here we can find several dividing lines. First, there is a

difference between different countries' legislation with regard to privacy protection connected to research on digital data when it comes to collection, information and storage. This creates differences when it comes to getting access to using such data. At present, there are efforts at the European level to harmonize regulations relating to data protection rules, among them regulations pertaining to research.[52]

Second, there is a fundamental gap between those who own digital data and those who do not. When consumers make use of services, such as mobile phones, bank cards or online shopping, or use Facebook or Google, they also give the providers of these services permission to use the personal data they enter by accepting the terms of conditions for the service. This permission allows private companies to stand in a privileged position when it comes to utilizing data, both for research and analytical purposes, as well as for commercial purposes, without having to abide by the same set of ethical guidelines that research does. The companies are not subjected to any requirements to make these data publicly accessible so that other researchers can make use of them.

It is not only reasonable but of some importance that research on digital data are subject to strict ethical requirements, especially considering the users' potential powerlessness when it comes to protecting their own data. At the same time, it can be claimed that current regulation posits some conditions that are not up to speed with the general public's perception of the boundary between public and private information, and the perception of what kind of information should be covered by the privacy clause. One example that was published in *Forskningsetikk* (Research Ethics) 1/2013 has to do with the use of Twitter posts in an aggregated form, for example if one wishes to analyze all Twitter posts

52 Cf. http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm

with the hashtag #bevarhardanger (savehardanger). In relation to research based on such Twitter data, Kim Ellertsen, director of the Law Department at the Norwegian Data Protection Authority, emphasized the importance of informing Twitter users about the research. He also questioned the practical difficulties in reaching the users (Forskningsetikk 1/2013: 8). As pointed out by Ellertsen, the issue of the need to inform research subjects raises questions about whether the results are of general public interest and about the effects on freedom of speech. In our view, one key issue is whether these posts should be viewed as public statements on par with statements in newspapers and edited media or as personal information.

Referring to Manovich (2011), it is possible to claim that «the computational turn» has created a new hierarchy between those who produce data (that is, consciously or unconsciously leave data behind), those who have the tools to gather data and those who have the expertise to analyze them. boyd & Crawford point out that the latter group is the smallest and the most privileged, and it is the group that will have the most influence on how Big Data will be used and developed (2011: 113). There is a danger that large segments of social science research will be unable to contribute to these analyses.

The responsibilities and challenges of research

Internet research does not only present us with a new set of data and methods with corresponding ethical problems, but also with a new ecosystem for the production of knowledge. To define what the responsibilities of research are, we believe it is important to be aware of the fact that research plays a role in at least two different ways: as a producer of knowledge and as an actor with a shared

responsibility to develop and practise a code of conduct adapted to Big Data. Both of these roles are to some extent dependent on conditions outside the research itself, such as the activities of other actors and public and private regulations.

The role of research is thus primarily about providing knowledge about the new way in which information is stored and structured in a digital society, as well as to shed light on what kind of power different types of actors, such as citizens, elites, organizations, and states, possess when it comes to having access to and the ability to interpret information. This means contributing with research-based interpretations of what the Internet is and how it works. Researchers should take it upon themselves to provide understandings of both the structural qualities of the web and the social practices that develop within different social fields and among different groups. A major challenge for segments of the social sciences is addressing the methodological opportunities that the Internet and Big Data present us with in such a way that researchers will be able to provide these types of analyses. This requires an investment in a new competency which is not included in most of the researchers' basic education. A further requirement is a reflection on the methodological challenges and problems, such as the question of whom and what the different forms of web data represent. Through research, we take part in constructing the new information society as an object and providing the society with insight into what this reality is. This implies providing perspectives on such questions as: Do we understand the Internet as being open and free, or as something that is regulated and conditional? Does the Internet really constitute a public sphere, or rather a network consisting of isolated islands? Which voices are being represented in our research? What social interpretation is being produced that will potentially impact what legitimacy and significance the Internet will have in social political processes? Depending on which instruments,

methods and theoretical tools we, as researchers, use to attack the digital research field, we will be able to provide different answers to these questions.

The second role of research is about contributing toward developing an ethic that is adapted to the new premises laid out by digital systems and Big Data. We feel that such an ethic cannot be developed in a vacuum, but must take into account the ecosystem of knowledge of which research is part. The challenge of research is to produce valid research in an ecosystem of knowledge while being under pressure by privatization and ethical regulations that differ from one country to another, and which to varying degrees are adapted to digital data. One way to think about this is through the concept *accountability*, which can be understood as more encompassing than the concept of privacy protection (boyd & Crawford, 2011). Accountability is not only directed towards the research subject, but also towards the research field in a broader sense. Accountability implies reflecting on the consequences of research related to individuals, organizations and the public sphere, or towards potential shifts in the ecosystem regarding the production, collection and analysis of digital data. If independent researchers could get the same access to using and connecting private data that Facebook has, a researcher would perhaps not choose to summarily use them to analyze statements about, for example, religion and sexuality. The researcher must weigh her interest in providing solid, academic knowledge about a social phenomenon, against people's perceptions of digital forums, and their faith in them. If researchers choose to use material that is perceived as private or as taken out of context, this might violate both the legitimacy of social media and the legitimacy of research. At the same time, *not* using these data means that the researcher leaves it to Facebook's analytical department to interpret them. Such considerations must be carried out specifically in relation to different kinds of data and

linking of data, different kinds of subjects and the public's perception of these data.

Presenting interpretations of society is nothing new in social research, nor is the fact that these interpretations may have social consequences. Still, our claim is that the Internet presents us with a new set of challenges. It requires that we both understand and critically evaluate what kind of data Internet data is and what it can produce knowledge about, and that we understand the social context of this type of research The ethical dilemmas that Internet research presents cannot be resolved without a deeper reflection on and establishment of ground rules for how Big Data should be handled in society, where research and other forms of public and commercial data use are put into context.

The rules that pertain to digital data are adapted to a «small data» world, where both data and computing power are accessible in large quantities. The ethical challenges pertain not only to research, but also to industry and administration. Therefore, we need new forms of accountability for both research and society.

References

boyd, d. & Crawford, K. (2012). Critical Questions for Big Data. Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*. Volume 15, issue 5. pp. 662–679.

boyd, D. & Ellison, N.B. (2007). Social Network Sites: Definition, History, and Scholarship. Journal of Computer-Mediated Communication. Volume 13, Issue 1, pages 210–230.

Brandtzæg, P. (2012). Sosial Networking Sites: Their Users and Social Implications- A Longitudinal Study. Journal of Computer-Mediated Communication. Volume 17, Issue 4, pages 467–488.

Brandtzæg, P. & Lüders, M. (2009). Privat 2.0: Person- og forbrukervern i den nye Medievirkeligheten. SINTEF-Rapport.

Enjolras, B. (2014). Big data og samfunnsforskning: Nye muligheter og etiske utfordringer. *Tidsskrift for samfunnsforskning*, 55 (1). p. 80–89.

Enjolras, B., Karlsen, R., Steen-Johnsen, K. & Wollebæk, D. (2013). *Liker, liker ikke. Sosiale medier, samfunnsengasjement og offentlighet*. Oslo: Cappelen Damm.

Hirzalla, F., van Zoonen, L. & de Ridder, J. (2011). Internet use and political participation: Reflections on the mobilization/normalization controversy. Information Society, 27(1): 1–15.

Manovich, L. (2011). Trending: the Promises and Challenges of Big Social Data. http://lab.softwarestudies.com/2008/09/cultural-analytics.html#uds-search-results, retrieved 09-01.13.

Savage, M. & Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology*. Vol. 41, no. 5. 885–899.

Siegel, E. (2013). *Predictive Analytics. The Power to Predict who will Click, Buy, Lie or Die*. Hoboken, NJ: John Wiley & Sons.