2013

Data Horizons

New Forms of Data For Social Research

Mark Elliot, Kingsley Purdam, Elaine Mackey
School of Social Sciences, The University Of Manchester
CCSR Report 2013-3 12/6/2013

Acknowledgements

The research described in this report was part funded by the ESRC grant number: RES-149-34-0001-A.

We would like to acknowledge the input of the participants of the Data Horizons Workshop which took place in Manchester in June 2012.

Francis Aldhouse	Bird and Bird Solicitors
Andy Brass	University of Manchester
Keith Cole	University of Manchester
Noshir Contractor	Northwestern University
Louise Corti	UK Data Archive, Essex
Libs Davies	UK Information Commissioners Office
David de Roure	Oxford University
Tanvi Desai	University College, London
Chris Dibben	St Andrews University
Bill Dutton	Oxford University
Rachel Gibson	University of Manchester
Mark Handcock	University of California, Los Angeles
Justin Hayes	University of Manchester
John Keane	University of Manchester
Sarah King-Hele	University of Manchester
David Martin	University of Southampton
Karen McCullagh	University of East Anglia
Mary McDerby	University of Manchester
Megan Merredith-Lobay	Oxford University
Robin Mitra	University of Southampton
Suzy Moat	University College, London
Gareth Morell	NATCEN
Paul Norman	University of Leeds
Nazmiye Ozkan	Policy Studies Institute
Simon Peters	University of Manchester
Rob Proctor	University of Manchester
Felix Ritchie	Office for National Statistics
Natalie Shlomo	University of Southampton
Richard Wiggins	IOE London
Farida Vis	University of Leicester

We would also like to thank the stakeholder interview participants, and the survey respondents.

CONTENTS

“There is more data for social research but can people use it, under what conditions and do they know how to?” (Stakeholder interview i3).

“There will be both more open data and more restricted access data” (Stakeholder interview i10).

“Different data, different quality, different time frames but same methodological challenges in relation to producing robust findings” (Stakeholder interviewi6).

“We need to challenge the assumption that new data means we know everything...there are gaps for certain populations. Some things are still unreported!” (Stakeholder interview i2).

“There is data overload. There is no time to analyse it” (Stakeholder interview i4).

“There is a growth of under theorised empiricism in social science…Uncritical use of data with limitations in coverage or definitions and the steering of research to things that happened to be measured” (Survey respondent).

“Lack of clarity on how to handle new types of data with regard to data protection and copy right” (Survey respondent).

“Social scientists need to be at the forefront of setting the standards for analysing new types of data” (Stakeholder interview i12).

“Research design and ethics training need to be a priority” (Stakeholder interview i7).

Executive Summary

This report presents a view of the social *data horizon*, in order to map the development of likely and possible data forms and sources and to develop an early view about what methods might be needed to analyse the new data. The report draws upon a project conducted between January and June 2012 under the Digital Social Research programme funded by ESRC. The project included: (i) the setting up a stakeholder group, (ii) a series of stakeholder interviews, (iii) an exploratory on-line survey of different types of data and their use and (iv) a workshop.

The New Data

- The last two decades have seen a step change in the data that is collected as part of everyday life and for social research. We are in the *age of data.*
- Data on human activity, decisions, preferences and behaviour is being routinely and automatically collected and collated as people go about their lives as part of a service or transaction process, or as a secondary process or self-released/published by the data subject.
- New types of data continue to be created. These cover an increasingly rich array of human activity including: health, genetics, location and movement, transaction and communication.
- A new theoretical framework is required in order to understand the changing data environment. Here we outline a typology of data based on data generation processes which provides one tool for use in such a framework.

The Challenges and Opportunities for Social Science

- New types of data present both new opportunities and new challenges for social research.
    - For example, Twitter data can enable the instant tracking of attitudes. However, representative data is needed to examine the attitudes of the wider population.
- A key feature of the new data is the potential for data linkage. The opportunity for social research is in bringing together the different data (types) to achieve the right mix of evidence to address a research question rather than a relying on a single not-quite-fit for purpose data set.
- Increasingly people's identities are defined and played out on-line with traces left of their behaviour. Social science research techniques should become responsive to the immediacy of people's digital lives.
- New tools for data collection and analysis include: automated website searching and data collection software and social media archiving and coding software. New tools will continue to appear and evolve.

- Social scientists could - and perhaps should - both be championing access to new types of data and also brokering, supporting and leading new partnerships and archives.

- Even within positivistic frameworks, there is a blurring of the orthodox boundary between the researcher and the researched and analytical processes are going to be increasingly less divisible from the data that is analyzed.

- Social research is becoming an area of activity for the private sector and citizens themselves are also engaging. As such, the boundary of social science has itself become the subject of debate.

- Social scientists face competing voices and narratives for researching and examining social issues.

- New research design frameworks and good practice guidance are required in relation to the collection and use of social media and administrative data.

Data Quality Issues

- Social media and administrative data can be partial in terms of both issue and population coverage.
  - For example, demographic information will often be missing/access restricted from Twitter data whereas information on individuals' attitudes is often missing from administrative data.
  - Vulnerable and hard to reach groups can often be missing from social media and administrative data sources. Other data and other research designs will still be important.

- Data standards for digital data including data quality are limited. The data can often be unformatted and non-standard. Though this can also be a problem with more traditional datasets especially when linking data from different countries.

- There is no established framework for social media data and few standards for coding and use.

- The key tenets of social science practice are theory lead research questions, validation, replication and effective peer review and these will remain important.

The Old Meets the New?

- Whilst remaining important, social surveys will increasingly exist alongside data gathered from social media communications, people's digital lives such as on-line transactions and their administrative records.

- There is still only limited awareness of the rich sources of survey data available in the UK and beyond and there is evidence of only limited use of certain newly released administrative datasets.

- The existing social research good practices of quality assurance, peer review and ethical data collection should apply to the access and use of the new data but new challenges are posed: methodological, ethical, and theoretical.

- Orthodox data is also changing. Increasingly, different types of data are being collected within the same instrument combining, for example, demographic, attitude, physiological and genomic data.

- A possible way forward is not to think of the new as replacing orthodox data but rather to think in terms of the added value that the new types of data and associated methods bring.
    - Different data and methods can be used to cross validate one another and also as part of calibration tests.

Data Archiving and Access

- We are moving from an era of social data only being available to government researchers and - to a degree - academia to one of much wider access.
    - The UK Government is providing wider access to certain types of administrative data. It will be important to make best use of such data to justify the costs associated with increasing access. At the same time this data should be made available at reasonable costs.
    - Citizens are self-generating data (including data generated as part of citizen science).
    - New data archives are being set up, some as part of commercial activity. These tend to provide only limited or no academic social research access.

- Social media platforms are in a state of continual development. The data that is collected and made available over time can change. Such sources are not stable archives for those accessing them on-line or from outside the organisation.

- Public bodies are creating archives of social media and administrative data for research use.

Training, Ethics and Good Practice

- New types of data require new training for social researchers in relation to data access, anonymisation and confidentiality as well as research design, data handling skills and computer programming and skills in understanding of the limitations of different types of data.

- The tool kit of the social scientist (including funding council training) needs to develop to reflect changes in the data landscape. This applies to the present generation as well as the next generation of social scientists.

- New ethical challenges are posed in relation to the use of social media and administrative data in the areas of ownership, access and disclosure.
    - Questions concerning what "public data" is and what type of data protection is needed should be re-evaluated.

- Debates regarding *informational citizenship* should be initiated to ensure fair use of data and to reflect the growing economic value of individual level data and its inherent social research value.
    - This should incorporate the clarification of the rights of secondary use both in the public and private sector.

1. Introduction and Context

More than a century since the ground-breaking social surveys of Booth in London[1] and Rowntree[2] in York and the subsequent development of Mass Observation methods in the 1930s we are now in an age of almost overwhelming data volumes about people's circumstances. Such data includes information on: attitudes, images of people and places, people's movement and communications. This data revolution includes: life-long health and prescription records, brain scans, genetic, bio marker profiles and family histories, satellite images, digital passports, databases from product warranty forms, consumption transactions, online browsing records, email and web communications (including self generated blogs and Twitter postings), geo-coded information on movement and mobile phone use, and synthetic data. Access to administrative record data held by public bodies including government departments is also being widened[3,4].

This report presents a view of the social *data horizon*, in order to map the development of likely and possible data forms and sources and to develop an early view about what methods might be needed to analyse the new data. The report is based on a consultative project conducted between January and June 2012 under the Digital Social Research programme funded by ESRC. The project included: (i) the setting up a stakeholder group, (ii) a series of stakeholder interviews, (iii) an exploratory on-line survey of different types of data and their use and, (iv) an interactive workshop.

[1] Booth's original work was published in seventeen volumes over the turn of the 19th Century. It has been subsequently summarised, reproduced and interpreted on many occasions. See O'Day and Englander (1993).

[2] Poverty: A Study of Town Life by Benjamin S Rowntree (2010)

[3] Within the UK, see www.adls.ac.uk for general information about new access opportunities (accessed 28th May 2013).

[4] At the time of writing, the UK Administrative Data Taskforce had reported (http://www.esrc.ac.uk/_images/ADT-Improving-Access-for-Research-and-Policy_tcm8-24462.pdf accessed 28th May 2013. As their recommendations are taken up by the UK government this is likely to lead to a step change in research access to administrative data.

Context: The Age of Data and Changing Orthodoxies

Data is information or knowledge on an individual, object or event. Data can be numerical values or text, sounds or images, memories or perceptions.

As we have pointed out elsewhere (see Elliot et al. 2008, Mackey and Elliot 2013) no data exists in isolation but is situated in a complex web of local and global *data environments*. The global data environment has always been an organic, dynamic milieu. In the last twenty years that environment has changed at a previously unimagined rate. As well as the documented expansion in the quantity of individual level data (see, for example, Purdam et al. 2004, Elliot et al 2010), categorical shifts in the type and form of that data are happening and will continue to do so.

Often the concept of data suggests information that has been through some kind of processing and having a structure. However, many examples of new types of data that might be used for social research have very different and often unstructured formats and size; for example, millions of Tweets or a corpus of public documents in PDF format. See Thelwall and Viz (forthcoming) for a discussion. Of course, even orthodox social data types such as surveys can have issues of format variance especially when trying to link data from different areas or countries. However, the proliferation of data types has significant consequences in how that data is merged and what it means to be a social scientist.

Social science and the societies that it studies have undoubtedly entered the *age of data*. The term 'Big Data' has been much used in this context and certainly moves us forward from Sweeny's (2001) discussion of the 'information explosion' and captures the growth in the collection and availability of information (See Boyda and Crawford (2013) and O'Reilly (2011) for discussions). 'Big Data' denotes volumes of data so large that it is kept in so-called data warehouses, which are essentially large data storage facilities often transcending different national borders and data regulation regimes. The term 'Big Data' does include what can be termed orthodox/well-established forms of social data such as survey responses and focus group transcripts. By orthodox social data we mean data collected with the specific intent of doing social research. Though, of course, even these data/methods are continually being developed and renewed; we have travelled some distance from the purposive cross-sectional surveys of early 20th Century in the UK to the cohort studies and experimental controlled trials increasingly used today. However new types of data can have very different origins and structure. Some might be collected primarily for research use, whilst others might be produced as a secondary outcome to another activity, for example, buying a product on-line or posting views on a blog.

Arguably, the term 'Big Data' represents a failure of imagination; a failure to capture the all encompassing nature of the socio-technical transformation that is upon us. Many who use the term quickly caveat that use by saying that 'Big Data' is not just about volume but also other features: that data can be captured, updated and analysed in real-time and that it can be linked through multiple data capture points and processes. However, such caveats are not sufficient, they still express the notion of data as *something we have* whereas the reality and scale of the data transformation is that data is now something that we are *becoming immersed* and *embedded in*. We are generators of, but also generated in, the *data environment*. Hence, we use the term *the age of data to* capture the historical phase that large parts of society has just entered to evoke the reality of the new relationship between humans and what is known about them – the data.

2. The Data Horizons Project

The data horizons project was funded under the Digital Social Research programme. The purpose of the project was to elicit and collate views from a variety of stakeholders regarding the future of social data. The project which ran between January 2012 and April 2013 consisted of four components.

1. Identification of Stakeholder Network Group
Through snowballing from contacts, web searching, attendance at linked events and conferences a list of possible interested individuals/organisations was identified and invited to the workshop. The aim was to obtain a cross-section of knowledge, expertise, discipline and research interests amongst the stakeholder group. The network supported the web survey and the workshop organisation.

2. Qualitative Interviews with Stakeholders
Semi-structured interviews were carried out with stakeholders. The aim of these interviews was to elicit a preliminary view of various perspectives on the data horizon. The objective for each interview was to produce a map of the present and future data types and use. In total twelve interviews were conducted. See appendix 2 for the interview question schedule.

3. Exploratory On-line Survey
Using the insights gained through the interview process an on-line survey was developed and delivered. The survey canvassed views on the data

horizon from the wider academic social science community. It was a purposive sample designed to provide a wider perspective on the present and future use of new types of data and the methods for analysis. In total over 300 usable responses were received. In terms of occupation, respondents were as follows: PhD students (34), Researchers (100), Lecturers (37), Professors (69) and Other (49). See appendix 3 for the survey questionnaire.

4. Workshop
A one–day discussion workshop was hosted at the University of Manchester in June 2012 with four semi-structured sessions on: new data sources, data linkage, data infrastructure and methods training requirements (see Appendix 1 for the workshop agenda).

In total, thirty people participated in the workshop including academics, research practitioners, data archive managers and funding council representatives.

3. Scanning the Data Horizon

"There is more data for social research but can people use it, under what conditions and do they know how to?" (Stakeholder interview i3).

We are now in an age of almost overwhelming volumes of data about people and their circumstances. That data comes in many types and forms and is collected and analyzed for many different purposes both within and outside of social science. Understanding and classifying the data environment has itself become a major undertaking and there are many different approaches one can take to this task. In this section we will consider the changes to the so-called orthodox social data as well the new data before introducing a functional typology, which could act as an organizing framework to help the social scientist navigate the data environment.

3.1 Orthodox Social Data

It is worth reflecting on the expansion and enrichment of orthodox social data sources. In the UK, the longitudinal surveys such as the British Cohort Study, the English Longitudinal Study of Ageing and the Census Longitudinal Study now constitute very rich sources of data for understanding change over decades of people's lives. International surveys such as the World Values Survey provide insights into global opinion. Such surveys and the analysis conducted on them often combine contextual data with survey responses. The data can often be analysed on-line through user interfaces such as Nesstar[5].

3.2 Administrative Data

One core feature of the data horizon is the increasing access that data users will have to official administrative records such as patient health records and school performance records through initiatives such as the Administrative Data Liaison Service.[6] Research access to this kind of data is supported by the UK government's Open Data[7] agenda which requires greater transparency of (and promotes accountability in relation to) public sector service use and performance. (Open Public Services - White Paper 2011). For an overview see Wind-Cowie and Lekhi (2012). Less obvious, perhaps, is that

[5] http://nesstar.esds.ac.uk/webview/index.jsp; accessed 28th May 2013
[6] See www.adls.ac.uk; accessed 28th May 2013
[7] The Open data agenda is another major force in data innovation which is itself impacting on the movement of the data horizon. There are arguments that some forms of research data should be made open. The UK is at the forefront of this movement – most notably through the work of the Open Data Institute (see: http://www.theodi.org/; accessed 28th May 2013) but there are equivalent drivers in place across the world.

administrative data can include physiological measures and genetic data. The use of such data is likely to be new to many social scientists and, for many, require new skills and interdisciplinary working.

In our survey nearly two thirds of the (admittedly self-selected) respondents had used administrative data in their research. However, a similar proportion (61%) reported encountering barriers when trying to access administrative data. Respondents highlighted how access can be very slow and whilst some small organisations may be willing to provide access to data the preparation costs involved can be prohibitive. One theme emerging from the workshop was that many datasets are being used in only a limited way.

The survey respondents also identified many examples of administrative data they would ideally like access to including: tax records, benefits application records, business service data, government decision making, purchasing data, energy consumption data as well as asking for more detail in existing administration data that has been released such as the School Census data.

In the UK legislation has also made public sector information increasingly available for research and scrutiny purposes. Under the *Freedom of Information Act 2000* (FOI) requests for detailed records of what we might term consequential data held by public bodies can be made. Unless there is a reason not to, public bodies must provide the information within twenty working days.[8] Reasons for refusal include: costs, whether the request is vexatious or if it would prejudice a criminal investigation. The legislation has been widely used to examine transparency in government. Thousands of requests have been made since the introduction of the act many in areas that social research has a track record of examining such as public spending and public office decision-making. Access to new types of data has facilitated research breakthroughs in these areas including notably: information on MPs expense claims, records of donations to political parties, extent of care home abuse allegations, detention of children in police cells, links between police forces and commercial companies, police work-force demographics and gambling spending levels. However, as reported in Lee (2005), the majority of such requests are not for what might be considered standard social research purposes. Nevertheless, some recent examples in the UK context include: local authority data on business cases for new schools (Khadaroo 2008); Ministry of Defence Medical Data (Seal 2006), Department of Health data on drug addiction policy (Mold and Berridge 2007) and Police force crime data (Hutchings et al. 2006). At the same time there has been a growth in the leaking of public records data, including most notably Wikileaks which have been used to examine issues of governance and accountability and publishes what it claims to be 'original source material'.[9]

Research access to administrative data creates ethical and legal questions of confidentiality and security. One aspect of dealing with these – and one that has its focus within the Administrative Data Taskforce (ADT) report is the use of safe or secure data settings. Example of safe settings currently in use include the Secure Data Service (SDS)[10], the HM Revenue and Custom's (HMRC) Data Lab and the Minister of Justice (MoJ) Data Lab. SDS allows access to individual level data that is more detailed than that available under standard licensing and so provides potentially richer sources of evidence for social research. The

[8] http://www.ico.gov.uk/for_organisations/freedom_of_information/guide/refusing_a_request.aspx; accessed 28th May 2013

[9] http://wikileaks.org/; accessed 28th May 2013

[10] See http://securedata.data-archive.ac.uk/; accessed 28th May 2013

user analyses the data remotely rather than downloading them. The analytical outputs are then checked by the data provider. The conditions of use are based on special licensing agreements with users and user accreditation, individual training and trust.[11] The HMRC's data lab allows access to individual tax records under controlled conditions and the Ministry of Justice (MoJ) Data Lab allows organisations working with offenders to have their data linked to the MoJ re-offending data. In our survey 26 per cent of our survey respondents had used virtual safe settings to access and analyse data.

There are also increasing opportunities for data linking using statistical matching and drawing on multiple data sources to address research questions.[12] This often involves linking orthodox and administrative data sources together. Well known examples of already linked datasets that are available include: the linking of the hospital data to the Millennium Cohort study (see Calderwood 2007), the Work and Pensions Longitudinal Study (WPLS) which links benefit and programme information held by the Department of Work and Pensions (DWP) with employment, earnings, savings, tax credit and pension records from HMRC and the Longitudinal Study of Young People in England (LSYPE) which links annual survey data to data from the national pupil database.

Some surveys (for example, the National Survey of Wales and the Scottish Longitudinal Study) now ask for the respondent's permission for the anonymous use of their responses for the purposes of linking with other data sets. The ESRC is presently reviewing the area of data access and linkage as part of its Administrative Data Task Force (See Boyle 2012).[13] With the recommendations in the ADT report looking likely to be broadly adopted by the UK government (reported by Boyle 2013) and significant public funds being invested to facilitate access (through four administrative data research centres, a dedicated information gateway and network of safe rooms distributed across the UK), it seems highly likely that the use of administrative data is set to expand significantly. This in itself would represent a step change for social science and that is even before the many other new data sources and types are factored in.

3.3 The New Data

We use the term *new data* to denote data which has only recently started to be utilized for social research. Most of this data has been around in some form and quantity for some time, but its use for social research has been limited perhaps because of access and infrastructural constraints, methodological uncertainties and a lack of interest in, or opportunity for, social research use. New data types include: (i) movement/geocoded data (such as mobile phone and satellite tracking data), (ii) transaction and consumption data, (iii) genetic markers (such as BioBank data), (iv) physiological data such as brain imaging data and eye tracking movement data, (v) communication data including blogs, Twitter and You Tube postings and mobile phone use, (vi) on-line browsing data, (vii) synthetic data, (viii) experimental data from policy impact studies, and (xi) crowd sourced data gathered by volunteer/citizen scientists. As outlined above, new information sources also include public records accessible via Freedom of Information requests.

A critical functional difference of many of the new types of data is that the data is often generated by direct data processes. Citizens and organisations

[11] 50 per cent of our study respondents had used special license data.

[12] 54 per cent of our survey respondents had linked individual records from different datasets for their own research.

[13] http://www.esrc.ac.uk/_images/ADT-Improving-Access-for-Research-and-Policy_tcm8-24462.pdf ; accessed 28th May 2013

and service providers are creating their own digital archives either deliberately or by implication of, for example, people living their lives using social digital media. Such data can be collated in real time, visualised and analysed instantly and updated continually. Rather than waiting for fieldwork, real time data opens up opportunities for compressing the data collection-analysis-dissemination-impact process. These *data-streams* might be directly captured (for example through sensor systems) or *crowd sourced* where collective intelligence and effort in the form of observations, data preparation tasks and ideas generation is deposited and uploaded by volunteers. Information posted on-line can be gathered and collated automatically using screen scraper software and then used to build databases. In relation to social media data free on-line software tools now exist which enable you to collect, analyse and visualize Twitter data by specified topic and key words during a fixed time period[14],[15].

The notion of research based on real-time analysis of *data-streams* clearly challenges, and presents an alternative to the standard practices and timescales for data gathering, checking and for peer review. Undoubtedly such research is providing an opportunity to do research that would not have previously have been possible.

An example of such research - using Twitter data - is the collation and coding of Twitter postings during the civil disturbances in the UK, in 2011 (see Reading the Riots 2011). This involved the textual analysis of large volumes of Tweets to code for attitudes, to look for networks and contextual patterns and movement. Another example is that of a UK police force using Twitter to announce all its emergency calls to highlight their work in a given period. Over 3,000 Tweets were posted. Example Tweets are shown in Box 1.

Box 1: *Example tweets on the Greater Manchester Police Twitter account*

Call 215 stolen vehicle heading towards Manchester #gmp24 Thursday October 14, 2010 5:03;

Call 216 harassment report in Bolton #gmp24 Thursday October 14, 2010 5:03; Custody update 101 in police cells at 5 am #gmp24 Thursday October 14, 2010 5:04

Call 218 neighbour dispute in Wigan #gmp24

Call 219 nuisance call #gmp24

Call 220 aggressive shoplifter held at supermarket in Stockport #gmp24.

This type of twitter data can be coded by the social researcher for: incident, time, location and language and outcomes analysed. Though of course there needs to be an awareness of the limitations of this type of data i.e. its representativeness. To address this follow up qualitative research with people who live in the area and with police officers could be conducted. Such data is however much more detailed and geographically specific than that which might be available from a national representative survey. Comparisons with evidence from other police forces could also be made. What is clear and most exciting is that the social researcher is presented with many more options in terms of data and methods than in the past. The United Nations (UN) is a good example of an organisation leading in this area. The UN has embraced the use of digital data in relation to human rights and policy impacts in real time by: monitoring food

[14] Only samples of the data and subsets of the variables are usually available.

[15] See, for example, Webometric Analyst http://lexiurl.wlv.ac.uk/; accessed 28th May 2013

price discussions, money transfer patterns via mobile phone or tracking health concerns using digital signals in Twitter or Internet searches. These techniques can include feedback loops where people's attitudes and behaviour can be followed up and then captured again. To address issues of data quality tools are being developed for data checking involving volunteers.

An extension of these approaches is the involvement of citizens in the delivery of the research project itself including data gathering. This might be termed citizen social science. A well developed example is, that of, the Open Street Map which is an open source map of the world generated from thousands of volunteers submitting data including images. This initiative was driven by a desire to make such information freely available.[16] It is notable that Google has recognised the value of crowd sourcing techniques as it increasingly allows registered volunteers to add detail to Google maps where the suggested content is checked and then added as part of Google Map Maker.[17]

Other examples of citizen involvement in research are The Satellite Sentinel Project and the Everyday Sexism Project. The Satellite Sentinel Project is essentially a data analysis project which asks citizens to look at images, and code, for evidence of human rights abuses in Sudan. These images maybe of military activity or signs of explosions sourced from a network of private satellites.[18] In contrast, the Everyday Sexism Project is essentially a data gathering project which asks citizens to report their experiences of sexual harassment[19]. Citizen generated data has both its strengths and weaknesses. The project has gathered rich data that could challenge peoples' everyday understanding of sexual harassment. It is easy to see the appeal of data gathered in this way not least because it is relatively low cost, easily accessible and can potentially have an immediate media impact. Such impact can lead to respondents engaging in follow-up discussions both between themselves and with the media and policy makers. However, the sample is limited and there is no verification process; as such there can be no straightforward extrapolation to a measurement of prevalence. Though the evidence suggests prevalence is non-zero. The research design could be developed further to overcome some of these limitations. By, for example, asking respondents to report key demographics, change over time and to describe how their experiences compare to people they know in their social networks.

In parallel with these developments commercial data companies are increasingly creating and providing access to highly detailed individual level information products which include data on: name, address, full postcode, age, gender, income, occupation, number of children, household income, house type, tenure, education, consumption, length of residence, car ownership, insurance packages, ownership of ICT products, holidays, smoking, leisure activities and social attitudes. The data is compiled from different sources including: primary surveys, warranty forms where citizens agree to the shared use of their details, public records, administrative records (such as the electoral register and house sale information) and consumption records. Imputation techniques are used to estimate missing data and attitude profiling is used where demographic information is missing. Some data sources are available instantly and techniques have been developed to combine different types of data often involving large numbers of individuals and variables (see Purdam et al. 2004). This individual record data can be purchased at

[16] http://www.openstreetmap.org/; accessed 28th May 2013
[17] www.google.com/mapmaker; accessed 28th May 2013
[18] http://www.satsentinel.org/; accessed 28th May 2013
[19] www.everydaysexism.com/; accessed 28th May 2013
See also: http://www.bbc.co.uk/news/uk-21520385; accessed 28th May 2013

relatively low cost although access processes are controlled and users are now required to explain what they are going to use the data for.

It is clear, in the data environment there are increasingly detailed records of actual behaviour alongside survey data on peoples reported behaviour and scope for reporting and monitoring behaviour in real time alongside and/or as opposed to more traditional data such as that gathered through diaries or in survey questions etc.

3.4 A New Typology of Data

To aid social scientists we suggest a new typology of data based on the generation process of the data. Given the complexity and changing nature of the data environment it can be argued that mapping the data generation process or the origin of the data is the only stable way of understanding it.

1. **Orthodox intentional data:** Data collected and used with the respondent's agreement. All so-called orthodox social data (e.g. survey, focus group or interview data and also data collected via observation) would come into this category. New orthodox methods continue to be developed.

2. **Participative intentional data:** In this category data is collected through some interactive process – this includes some new data forms such as crowd-sourced data and is a potential growth area.

3. **Consequential data:** Information that is collected as a necessary transaction which is secondary to some (other) socio-physical or virtual interaction (e.g. administrative records, electronic health records and commercial transaction data all come into this category). Many consequential data sources are in theory complete rather than being based on samples. However, like all data sets there are likely to be issues of missing data and incompleteness that the social researcher needs to be aware off including individuals who have not been traced or recorded and duplicate records.

4. **Self-published data:** Data deliberately self-recorded and published that can potentially be used for social research either with or without permission (e.g. Blogs, on line CVs and profiles).

5. **Social media data:** Data generated through some public social process that can potentially be used for social research either with or without permission due to it being already published (e.g. Twitter, Facebook and perhaps on-line game data).

6. **Data traces:** Data that is left (possibly unknowingly) through digital encounters such as on-line search histories and purchasing that can be used for social research either by default use agreements or with explicit permission.

7. **Found data:** data that is available in the public domain such as, for example, observations of public spaces and can include covert based research methods.

8. **Synthetic data**: where data has been simulated, imputed or synthesised.

This typology will aid in the development of new frameworks for social research, new research methods and new models of consent from data providers/respondents which need to involve research use of self published data and linked data. For further discussion of the application of this typology see Elliot and Purdam (forthcoming (a)).

3.5 Summary

The range of data will undoubtedly continue to grow and as a result very detailed records of peoples' lives are developing. Social researchers will potentially be able to access huge volumes of individual level data. A key challenge in this area is the issue of data access and use particularly in relation to new types of data such as social media data. Such data might in theory be available in real time but how can the data be used and by whom? There is a lack of clarity of the ownership and regulation of the use of many types of social media data.

Moreover, in methodological terms what population does the data represent? What generalizations can be made? How robust is the data in terms of being from real people and in relations to issues of performance - including where respondents provide answers driven by response bias effects or where people write Tweets intended to present a certain image or where they use fake Twitter accounts? We consider these issues in our discussions below.

The developments in the data environment including new types of data open up new opportunities for researching intractable social problems from different angles and perspectives with highly detailed data. Existing research, methods and data will not only become increasingly subject to augmentation but also completion from new approaches involving new types of data and new types of social researcher.

What is clear is that we need to understand the form the data environment will take in the short, medium and long term, in order to plan and develop the social science resource base and train both the present and next generations of social science researchers. In parallel with this there is a need for methodological innovation so that explanatory power of the new data and methods can be optimized.

4. Methodological Developments

> *"Different data, different quality, different time frames but same methodological challenges in relation to producing robust findings" (Stakeholder interview i6).*

Alongside the growth of new data sources has been a concomitant growth in new methods to analyze that data. Example new approaches include: software that scrapes and collates website content, automated image linking using on-line website tags, automated textual analysis of large numbers of samples of Twitter postings, analysis of on-line video (coding for content and views), combining/linking administrative and survey data using probabilistic and/or knowledge based matching, combined analysis of genetic and social data, recruiting citizen volunteers to collect data and/or code data including movement, images and text. Existing research methods can also be adapted to utilise social media technology and data such as, for example, using virtual ethnography to draw on the growing self generated archives of everyday life as highlighted by Beer and Burrows (2007).

The potential here is for research questions to be tackled through radically mixed methods applied to multiple data sources. A hypothetical example will help highlight how social science may develop:

An Example: Researching Anti-Social Behaviour. *A social researcher interested in anti social behaviour has a range of data sources to choose from and combine according the needs of their particular research question(s). For example, they could combine administrative data on anti-social behaviour alongside data from the British Crime Survey and commercial* *area profiling data such as that from Experian or CACI. The researcher might also analyse social media data from police forces and officers. This in turn could lead to follow up interviews. The researchers might also follow links to Facebook and any You Tube postings and the analysis of content. This might then be linked with information on criminal proceedings, prosecution and data on offender rehabilitation and victim support.*

Given this inter-diffusion of methods and data, it is perhaps more accurate for a social scientists to be thinking in terms of *data arrays* rather than data sets. One could expect the social researcher to combine data from different parts of the array and to transcend traditional divides such as qualitative vs. quantitative methodologies.[20]

However, as we have indicated previously, ubiquitous data will not compensate for lack of methodological robustness; the issues of data quality, data structure and standardisation remain. Horizontal and vertical limitations (including generalisability) remain as does the need for theory/hypothesis driven research. As one survey respondent stated: *"there is a growth of under theorised empiricism in social science…uncritical use of data with limitations in coverage or definitions and the steering of research to things that happened to be measured"* and another expressed concern *"about more data being widely available than people are able to properly analyse….there will be misinformation, with people putting analysis into the public domain which is inaccurate sometimes deliberately but more often through error"*.

Any *data array* is likely to contain mixed signals with lots of noise and will only be loosely structured. This presents a heightened risk of bias, misreporting and limitations to generalisability. Criteria for understanding samples in relation to social media and volunteer data and the significance of findings need to be developed. This might involve Bayesian and non-parametric statistical techniques, triangulation, splitting samples, blind analyses, automated quality assurance and data checking perhaps also using volunteers and/or citizen science based approaches.

As with so-called orthodox data, the research users of new data need a good awareness of issues of data use, data quality and coverage. For example, Twitter data is often subject to horizontal and vertical coverage restrictions by Twitter (or secondary suppliers); there can be limits on not only the number of Tweets that can be purchased and the variable coverage of the Tweeters (specifically the demographic information Tweeters provide when opening an account may not be made available). A further specific key concern here is the issue of fake and multiple accounts, or Tweets people have been paid to write where the data maybe largely performative (where responses are contrived in some way to create an impression (Law 2009), as well as the more substantive issue of differences between people's socio-physical and online personas. This could be particularly problematic for attitudinal data as attitudinal presentation could well be a key part of constructing an on-line persona. However, no reliable data exist on the prevalence of this phenomenon or how it impacts on the "real" attitudinal data.

New analytical software also needs to be quality assured and peer reviewed. As one workshop contributor noted *"you can't just make it up as you go along"*. At the same time there needs to be space for new approaches and methods.

A key further aspect of the implications for

[20] Tim Berners-Lee has famously predicted that web 3.0 will be the web of linked data (http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html; accessed 28th May 2013). If his prediction is accurate (and there appears to be no reason to doubt him), then we can expect the transformative process we have been describing here - the blurring of traditional social science dichotomies - to intensify further.

methodology is the potential change in the traditional boundaries between researcher and subject, where research is increasingly done *with* participants rather than *on* them. It is easy to envisage research being led by a particular population such as service users or organisation members reflecting the traditions of Active and Participant Research (Emmerson et al. 1995 and McCall and Simmons 1969).

Martin (2012) predicts that the distinction between data and analysis will become less clear. Many forms of new data provide opportunities for real time analysis and condensed dissemination time frames through the shortening of conventional research workflow based around data collection, analysis and reporting. Rather than researchers analysing data and then the results feeding through into policy impact in a lagged and somewhat ad hoc manner, one might envisage researchers-cum-policy analysts directly intervening in social processes using real time data systems as a tool and combining what in the past might have been seen in very different data types.

In summary, there is an urgent need for rigorous methodological work that allows social research to utilize the undoubted benefits of the new data in a robust and principled way. Moreover, as we discuss below, whilst new skills will be needed, the existing skill set of social scientists in terms of research design and understanding data will still be vital. It is notable that in our survey, nearly three quarters of those who took part thought that conventional social research methods such as surveys would not be used any less in the future.

5. Data Resource Infrastructure and Data Sharing

A key concern identified in our research was the issue of access to administrative data held by government departments and agencies and other social data held by commercial organisations.

Such concerns cover not only data that is in the public domain but also access to other aspects of data held on individuals, (for example, the socio-demographic information captured by twitter). It is perhaps naïve to expect research access to commercially held data without a charge or constraint; such access raises issues of commercial sensitivity as well as personal privacy. However social media data perhaps presents a new opportunity for access given that it is directly user generated.

Social scientists should be championing the access to social data held commercially and the setting up of partnerships. A historical precedent lies in the opening up of access to government survey and census data in the early 20th Century. The case made for access to UK Census data for research use and the Samples of Anonmymised Records may be a useful model to follow. Key here was not only access for research purposes but also as a way of developing transparency in governance as researchers can analyse the data used by government departments. Such data is released for research use after extensive data preparation and confidentiality protection and it is likely that such a model may have some value in the access to commercially held data too. One can imagine a model where rather than funding primary survey collections, the ESRC or BIS funds the processing of commercially held data to make it fit for secondary research use.

Legal clarification is required in relation to the use of new data, its archiving, privacy and consent for re-use. In particular, there is also a need for clarity about which aspects of social media data are available for research use and secondary publication. For example, one of the barriers to setting up public archives has been the issue of copyright, though these have now been overcome for certain public archives following legal clarification about information use and the permissions required under the *Legal Deposit Libraries Act (2003)*.

A radical extension to FOI legislation could legally require anonymised forms of commercial data to be made available for research purposes at marginal cost and/or perhaps after an embargo period to allow commercial exploitation. Public resource mechanisms should be set up to enable this to work. Moreover many governments and state agencies themselves are legally allowed to record electronic communications such as, for example, email and web browsing sometimes in real-time. Yet access to such data for social research purposes is likely to be limited as things stand.

Social researchers themselves perhaps need to lead the way by creating their own community archives. There are some innovative examples already including archives of web pages. The Internet Archive[21] service allows users to create their own archives. Information captured includes: text, audio, moving images, software and archived web pages. Data includes on-line records of web publishing in relation to September 11th terrorist attacks. The initiative recently launched by British Library to archive online materials including webpages and blogs is welcomed.[22]

A major challenge for social researchers is the many different formats that the new types of data can be in. As one stakeholder commented there are just *"huge volumes of unstructured data" (i7).* As we discuss below, this poses new challenges for social researchers. Our consultation suggests that existing UK data archives such as the UK Data Archive are not yet actively leading the archiving of social media data though they do archive such data if it is offered as a result of funding council research. Whilst they may not desire or be able to compete with, for example, You Tube the UK archives could set out a framework for the archiving of such data and drive forward good practices for data quality and for the use of such data in social research.

A further issue for data archiving relates to the continual development of social media content. Such platforms are not stable or fixed as people continue to post information and the commercial organisations can refine the information they collect and or release. Moreover, social media data platforms can themselves change and this can lead to secondary changes in the data that is collected and made available over time. Such sources are not primarily archives and should not be treated as such. There is a need for clear time stamping standards for archived data.

One stakeholder commented: *"There will be both more open data and more restricted access data" (Stakeholder interview i10).* Such bifurcation is likely to create inequality and indeed social scientists may not have affordable access to the best data; this includes certain types of public administration data where prohibitive charges are currently levied. This will severely challenge the explanatory power of the research they conduct and as a result the impact they can have in policy terms. Of course large social media companies do already have social research initiatives - working in partnership with social researchers.[23] However, it is not clear that there is an appetite for wider data sharing and access.

[21] http://archive.org/; accessed 28th May 2013
[22] See http://www.bbc.co.uk/news/entertainment-arts-22037199; accessed 28th May 2013

[23] such as, for example, Microsoft's Social Media Research Lab: http://socialmediacollective.org/about/ ; accessed 28th May 2013

Related to this, as Savage and Burrows (2007) outline, is the commercialisation of social science where the research driver is ultimately commercial advantage and profit rather than social value. Apart from creating moral and social policy conundrum there are critical issues of research quality to be mindful of: verification, replication and review in particular become much more problematic. To ameliorate these problems, the case needs to be made - probably through government - for regulatory processes regarding the provenance, value, reliability and validity of the new types of data and the analyses and claims made of them.

A further priority is the development of fair data use policies and effective data sharing protocols. This will include clarifying issues of the definition of personal data, anonymisation, confidentiality and data ownership with the new data. The existing data protection legislation and more recently the UK *Statistics and Registration Service Act 2007* (SRSA)[24] has been primarily designed for orthodox and consequential data. The proposed changes to the European data directives are not particularly helpful in this respect and there is a pressing need for a step change in legislation and agile regulatory processes. This would require significant and probably international funding.

To give a simple example, social researchers need clear guidance on the ownership and right to use Twitter data. This includes risks posed by linking posts and names and risks to third parties such as people named in Twitter posts. Facebook and Twitter postings occur in public but only certain aspects of the resultant data are available to the public. Twitter claims that Tweets are owned by the people who write them, but then treats them collectively as a saleable commodity. There is a process of consent as part of the process of creating a Twitter account.[25] Despite this it may not be entirely clear to the account holder how their Tweets might be used for secondary purposes including social research. There is only limited research on how such terms of use compare with other data types and forms of collection such as, for example, intentional data collected via survey or the UK Census.

To complicate things further, the individuals that the social media data *is about* or refers to may not be aware that the data about them exists or that the data is public and what this means in practice. Even if a person is aware that the data exists they may not realise that it is being used for secondary purposes (research or otherwise) and that it has a commercial value.

Volunteer and crowd sourced data from, for example, observing events may include information on other people taking part in the event. As Gross (2011) argues, the existing frameworks of ethics and particularly consent are

[24] Where 'personal' information is defined as *"information which relates to and identifies a particular person (including a body corporate)".* Information identifies a particular person if the identity of that person - *"(a) is specified in the information, (b) can be deduced from the information, or (c) can be deduced from the information taken together with any other published information".*[24] The disclosure of personal information held by public bodies such as the ONS is a criminal offence punishable by up to two years in prison.

[25] The account holder is prompted that:

"You are responsible for your use of the Services, for any Content you post to the Services, and for any consequences thereof. The Content you submit, post, or display will be able to be viewed by other users of the Services and through third party services and websites (go to the account settings page to control who sees your Content). You should only provide Content that you are comfortable sharing with others under these Terms…. You understand that through your use of the Services you consent to the collection and use (as set forth in the Privacy Policy) of this information, including the transfer of this information to the United States and/or other countries for storage, processing and use by Twitter…. By submitting, posting or displaying Content on or through the Services, you grant us a worldwide, non-exclusive, royalty-free license (with the right to sublicense) to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute such Content in any and all media or distribution methods (now known or later developed) Twitter; Terms of Service: https://twitter.com/tos

limited and they need to be overhauled if they are to cope with the scale, intensity and immediacy of the constantly evolving data environment. Such changes are crucial for the ethical development of social research using new types of data including trace data and for the effective regulation of uses of the data as the relationship between citizens, state and the commercial sector changes.

It may well be that we need to move from a legal culture of regulated *data protection* to one of policing *data abuse*. In a data abuse regulation framework one is less concerned about the control of data flows and processes and more with the consequences (and specifically) harms caused by the actions and choices of data processors. Mandatory social research access by approved researchers would be one aspect of such a framework. This would not necessarily jeopardize the commercial value of the data to businesses but could be part of a legal and ethical responsibility to the 'customer' and their welfare and a natural balancing of the power of large commercial data controllers.

It is clear that good practices for using, accessing and archiving social media data are at present limited. As several interviewees and survey respondents commented:

> *"There are still issues of territorial data custodianship and a lack of sharing" (Stakeholder interview i6).*
>
> *A risk is… "data in the hand of private firms without access for public research" (Survey respondent).*
>
> *"Most organisations haven't got a framework for data sharing" (Stakeholder interview i8).*
>
> *"Data access is still reliant on good relationships with individuals rather than just policies" (Stakeholder interview i1).*

A reliance on personal contacts for data sharing suggests a lack of openness. There is also a risk that social research will become increasingly partial if commercial companies hold and restrict access to rich sources of data that can help tackle social research questions or if data cannot be accessed to enable validation, replication and effective peer review. There is an indication of a mobilization in this area in the work of the Web Science Trust which is focused on sharing expertise and resources to enable research on and strategic thinking about the Web.

In summary, it is clear new frameworks for archiving and accessing new types of data need to be developed. These need to include variable descriptions, population and sample descriptions, time/version stamps and coding information. It is important that publicly and commercially held data is accessible, within efficient time frames at limited costs and is made available in non-disclosive, high-quality formats.

There is also a pressing need for genuine legislative review which does not merely tinker with the existing framework but seeks to create robust and up-to-date regulatory processes that can deal with the rapidly evolving data environment in a manner that ensures that the opportunity that the new data represents is not lost.

Despite the concerns highlighted above there is a genuine opportunity for alternative narratives and perspectives to emerge from the changing data environment, as evidence gaps are filled and new evidence often highly specific and detailed is exploited.

6. Training Development Needs and the Development of Good Practice

> *"Social scientists need to be at the forefront of setting the standards for analysing new types of data" (Stakeholder interview i12).*

The new types of data and methodological approaches bring into scrutiny the skill set and training of social scientists now and in the future. As one survey respondent stated a key risk was the *"inappropriate selection and evaluation of data for analysis".* Another respondent added that the *"misuse of poorly understood data"* was a risk and that, therefore, skills in the: *"evaluation of data and the appropriateness of inferences that can be drawn, particularly focusing on sample selection"* are important.

Of course social scientists should be continually learning and building the evidence base and the explanation of social research questions in a coherent, cumulative and critically engaged way. Engaging with the new types of data and methods is an important aspect of this.

Our survey has provided an insight into new types of data use and highlights how it is the current as well as the next generation of social scientists that will use such data. See Appendix 4 an overview of current and predicted use.[26]

Social researchers in the future are likely to find themselves working with very different and more diverse types of data from those they are used to at present. Current social scientists may well be unaware of the analytical approaches required in relation to genetic data for example. Social research may necessarily involve, even more than currently, working in interdisciplinary research teams.

At the same time as accepting that new types of data will become a part of what social science is, social scientists should not discard existing social research good practice standards, quality assurance, peer review and ethics processes. The challenge is to apply these to the new types of data and to develop research training that reflects the methodological, ethical, and theoretical challenges that the new data presents. Given the apparent blurring of the boundaries between researcher and researched and the collapsed time frames of social research (through real time data), it is all the more important that robust research practices are developed and integrated into social research training curricula and continuing professional development.

The collection and use of the new data for social research requires additional skills for social scientists including new skills in processing, analyzing and linking different types of variables such as Twitter posts and genetic data.

A key training issue lies at the interface between quantitative data skills and digital literacy. The quantitative data skill gap, which is widely acknowledged in the UK (and elsewhere), is being addressed by the research councils as part of ongoing initiatives. The need for such initiatives is only like to grow with the advent of the new data. Also the range of skills under this umbrella is only likely to diversify; even now statistical modelling, longitudinal analysis, network analysis, data mining, causal analysis, data simulation, data visualization, geographical information systems and mapping are amongst the mainstream quantitative skills for social researchers.

[26] It is noteworthy that some older respondents see only limited potential for their research in some of the new types of data. These findings do reflect the nature of our sample but perhaps suggest that at present that new data do not have much value for certain types of research.

In relation to digital data literacy, skills in web analysis, computation and software engineering would be of value alongside more traditional social science research skills. There is a direct link here with computer science and a growing group of *computational social scientists*. As one survey respondent stated: *"computer science skills will be more important than ever before, such that one can program tools in a way to collect the data on the web you want."*

A challenge when designing a study using the new data will be to achieve a complete understanding of the data: its origin, the sampling process and the limitations of the data (which are likely to relate to issues of bias, coverage and generalisability). This would imply that skills in data cleaning and processing, and an understanding of data quality in terms of non-response and other forms of missingness, would be far more critical than with orthodox data. Skills in archiving such data will also be important.

As we have outlined, social scientists maybe increasingly using mixed method approaches and drawing on *data arrays* rather than just one type of data or one data set. Where there is scope for combining and/or linking different types of data including administration data, skills in probability and statistical data linking/matching would be of importance. This might involve skills in various forms of matching and combing data both statistical and otherwise.

It is clear whatever type of data, or combination of types of data, are being used robust research design skills remain central. This includes the importance of hypothesis driven social research where theories can be tested. Whilst new data provide opportunities for more exploratory, inductive and data driven research, the importance of a clear set of research questions is key especially in relation to the opportunities and limitations posed by the new data types and volumes.

Established ethical standards will also remain important for social research. Indeed, the potential depth, breadth and complexity of the new data raise new, more acute, ethical and legal issues. As one survey respondent stated there is a: *"lack of clarity on how to handle new types of data with regard to data protection and copy right"*. Thus, a key area of development is good practice training in data handling and data confidentiality. Social research training content needs to address these issues more thoroughly than it does at present. It will no longer be possible to assume that secondary data use is ethically unproblematic.

Some of these issues will require expert legal clarification at a UK and international level regarding data protection and fair data use. The focus should arguably take account of concerns about an over reliance on security software as opposed to good data handling practices. It may also require new forms of respondent consent and social media permissions need to be developed. Whilst existing good practice guides are of use here such as those provided by the ESRC and the Social Research Association some issues require further legal clarification. New questions concerning what is public data are pressing. As one survey respondent stated a key risk is that *"careless data protection practices will lead to restrictions on further use"*. Social research training in the future should develop skills in data management, data security, anonymisation, encryption and safe data sharing including using virtual tools.

The new techniques for dissemination and publication also need to be part of the social scientists skill set. Whilst peer review still needs to be at the heart of the academic research process, social media and the internet provide opportunities for wider engagement with social research often with rapid turnaround times. Social scientists will need to be increasingly adept at speaking to different audiences without oversimplifying or diminishing the content of their findings.

A related aspect of this are skills and training in intellectual property rights, patents and knowledge transfer where social research involves the development of marketable products. The social scientist will increasingly need to protect their own intellectual property in the context of open data and publishing.

As indicated above the new data will make working in interdisciplinary research teams both more viable and more desirable. As one survey respondent stated: *"The next generation of social scientists will need to produce information in a more timely manner, which may require more training in managing teams of analysts".* Related to this will be the increasing number of voices in *competition for impact*.

In summary it is evident that the skill set and tool kit of the social scientist and funding council training needs to evolve and reflect changes in the data landscape. As we have outlined, key areas include: skills for collecting, processing, analysing and archiving new types of data, a greater focus on interdiscipliniarity and of engaging with different audiences and a deeper understanding of the ethical and legal issues associated with data use.

Moreover, a critical awareness of the opportunities and limitations posed by the new types of data needs to be developed. At the same time, the fundamental skill set of social scientists highlighted in the training focus of the research councils (research design, literature reviews, an understanding of quantitative and qualitative data types, sampling, data gathering, research ethics and peer review) will remain important.

7. Conclusions

The last two decades have seen a step change in the types of data available and potentially available for social research. Moreover, the social data horizon is moving rapidly. New types of data and meta-data are emerging including: health, genetic, movement, transaction, communication (including blogs and Twitter postings) and geo-referenced data. New tools for analysis continue to be developed leading to new opportunities for social research. At the same time there are new types of organisations collecting, archiving and using data that could be of value to social research.

Given that many social data processes are happening in real time, they begin to resemble control processes that one sees in manufacturing system. Scanning further forward, it is not too farfetched to envisage researcher-cum-policy-analysts directly intervening in social processes using real time data systems as a tool and combining what in the past might have been seen in very different data types. But notwithstanding such speculation and given the present rate of innovation, it seems certain new types of data will continue to emerge, and in response, new ways of doing social research will be developed.

As people's lives become increasingly digitized, so there is a need for social research to develop methods and tools to process and analyse the data that arises from and indeed constitutes those digital lives. The key methodological development area for social science lies in the potential for identifying and bringing together the different data types and sources. This is likely to include data generated by user groups and individual citizens sometimes replicating more standard social research techniques such as calls for evidence on-line and crowd sourcing.

Allowing the data to drive the research question is widely regarded as a high-risk strategy for empirical research. However, it is important to explore how the new data can best be utilised and such methodological exploration will necessarily be, in part, driven by the data. Inductive, data driven approaches may also be helpful for some substantive questions providing that there is a theoretical framework for the research. More generally as one survey respondent stated there needs to be a *"new way of thinking about science and the link between science and society…[for] effective and real scientific practice. We need a more complex way of thinking and not only better technical instruments".*

The new types of data provide the potential to explore existing research questions in new ways and to address new questions. As social scientists increasingly work with *data-streams* and *data arrays* rather than data *sets,* it is plausible to posit that, as with many other orthodox distinctions, the boundary between deductive and inductive processes will become fuzzier. Analytical processes may become less divisible from the data that is analyzed. So, whilst it is important that best practice is followed in terms of testing social research hypotheses and questions in a robust way with reference to existing methods and taking account of the existing literature, social scientists should not be closed to the possibility that the new data opens up and perhaps even requires a completely different way of thinking about the relationships between the data, the research, the researcher, the researched and the policy maker.

On a more immediate and less speculative note, we observe that as more consequential and found social data comes to be used in social research, the distinction between primary and secondary data itself will become less central. However, the use and legal status of any data for social research needs to be clearly understood by both citizens and researchers alike. There is a major data literacy issue and training in this area needs to be addressed.

An immediate way forward is to not to look to new data and approaches to replace the orthodox but rather to think of them in terms of adding value. So, for example, whilst Twitter data can enable the instant tracking of the attitudes and networks of a certain population, a representative survey can provide information on the attitudes of the wider population and a longitudinal survey can help examine how people's attitudes have changed during their lives. Similarly, data gathered in real time can be used alongside orthodox survey and administrative data. For example, reports of police call outs could be used alongside contextual data to map reported incidents and model the likelihood of future incidents or to examine patterns in relationships between actual crimes incidents and the fear of crime.

Different data and methods can be used to cross validate one another and also as part of calibration tests. Mixed methods approaches can add considerable value and explanatory power. A key component of mixed methodology research is to develop ways in which various sources of evidence might be combined and weighted. As we have outlined, despite the rise of the new data, there are still gaps in our understanding of different populations including hard to reach and vulnerable groups which may require more traditional research methods to capture.

In terms of training, the collection and use of the new data for social research requires additional skills for social scientists including new skills in processing, analysing and linking different types of variables, data structures and formats. In addition new challenges are posed - methodological, ethical, and theoretical. There has been only limited development of standards for new data use.

Furthermore, new ethical challenges are posed in relation to the use of new types of data including: ownership, access and disclosure. New research design frameworks and good practice guidance is required in relation to the collection and use of new types of data including administrative data. New questions concerning what is public data and data protection need addressing. Legislative review needs to become timelier.

A key aspect of this relates to the potential risks of the data, if used for unethical and illegal reasons including targeting certain populations and/or vulnerable groups. As one survey respondent stated: *"in the wrong hands [the data] can be used in a very harmful way".* Another respondent also highlighted how the growth in the collection of data may lead to a perception of increased surveillance and an unwillingness of citizens to participate in social research.

The key tenets of social science practice will remain validation, replication and effective peer review. However, social science needs to become more agile: (i) in its use of data, (ii) in its engagement with the society that it studies (and serves) (iii) in its willingness to innovate methodologically and take on new ethical challenges. It is undoubtedly true that, if it embraces the opportunities and challenges generated by the new data, social science has a significant proactive role to play in the development of our information society.

References

Aarsand, P. and Forsberg, L. (2010) Producing children's corporeal privacy: ethnographic video recording as material-discursive practice *Qualitative Research* 10: 249

Adkins, L. and Celia Lury, C. (2011) What Is the Empirical? *European Journal of Social Theory* 12(1), 5–20

Alasuutari, P., Brannen, J. and Bickman, L. (eds) (2008) Handbook of Social Research. London, Sage.

Alford, J. R., Funk, C. L., and Hibbing, J. R. (2005). Are political orientations genetically transmitted? *American political Science Review*, 99, 153–167.

Back, L. and Puwar, N. (2013) 'Live Methods'. London, Sage

Beer, D. and Burrows, R. (2007) 'Sociology and, of and in Web 2.0: Some Initial Considerations', *Sociological Research Online* 12(5).

Blasius, J. and Thiessen, V. (2012) Assessing the Quality of Survey Data. London, Sage.

Bowman-Grieve, L. and Conway, M. (2012) Exploring the form and function of dissident Irish Republican online discourses. *Media, War & Conflict* 5(1); 71-85

Boyda, D and Crawford, K. (2012) 'Critical Questions For Big Data. Information', *Communication & Society* 15(5), 662-679

Boyle, P. (2012) Improving Access for Research and Policy, ESRC. http://www.esrc.ac.uk/_images/ADT-Improving-Access-for-Research-and-Policy_tcm8-24462.pdf

Boyle (2013) "Harnessing Big Data across Society" Panel session at launch of e-health informatics research Centres and UK informatics research Network

Bryman, A. (2012) Social Research Methods. Oxford, OUP.

Bryman, A. Quantity and Quality in Social Research. London, Routledge.

Bryman, A., Lewis-Beck, M.L. and Liao, T. F. The SAGE Encyclopaedia of Social Science Research Methods. London, Sage

Cabinet Office (2011) Behavioural Insights Team Annual Update 2010–11. Cabinet Office, London.

Calderwood (2007) 'Methodological challenges in enhancing the MCS through linkage with data from birth registration and centrally collected hospital records', paper presented at 'Exploiting

Existing Data for Health Research' conference, St Andrews, 18-20 September 2007.

De Vaus, D. (2002) Surveys in Social Research, 5th ed. Routledge, London.

Diamond, I and Jefferies, S. (2005) Beginning Statistics. Sage.

Duncan, G., Elliot, M. J. and Salazar-Gonzalez, J. J. (2011) Statistical Confidentiality; Springer, New York.

Elliot, M., Mackey, E. and Purdam, K. (2010) Data Environment Analysis – Annual Report. Office for National Statistics.

Elliot, M. J., Purdam, K. and Smith, D. (2008) 'Statistical Disclosure Control Architectures for Patient Records in Biomedical Information Systems', *The Journal of Biomedical Informatics* 41, pp 58-64

Purdam. K and Elliot, M. J. (forthcoming (a)) 'The Changing Social Data Landscape' in Halfpenny, P. and Proctor, R. (eds) *Innovations in Digital Social Research Methods.* Sage

Elliot, M. J. and Purdam, K. (forthcoming (b)) 'Exploiting New sources of Data' in Halfpenny, P. and Proctor, R. (eds) *Innovations in Digital Social Research Methods*. Sage

Emerson, R. M., Fretz, R. I. and Shaw, L.I. (1995) Writing Ethnographic Field notes. Chicago: Chicago University Press.

Gilbert, N. (2008) (ed) Researching Social Life. Sage

Gill, L. (2001) Methods for Automatic Record Matching and Linkage and their use in National Statistics, The National Statistics Methodology Series, ONS (available at http://www.statistics.gov.uk/downloads/theme_other/GSSMethodology_No_25_v2.pdf)

Gray, J., Chambers, L. and Bounegru, L. (2012) The Data Journalism Handbook. How Journalists Can Use Data to Improve the News. O'Reilly Media http://datajournalismhandbook.org/

Gross, A. (2011) The economy of social data: exploring research ethics as device. *The Sociological Review*, 59(2), 113–129

Herzog, T. N., Scheuren, F. J. and Winkler, W. E. (2007) Data Quality and Record Linkage Techniques. New York: Springer. ISBN 978-0-387-69502-0

Hutchings, J., Bywater, T., Davies, C. And Whitaker, C. (2006) Do crime rates predict the outcome of parenting programmes for parents of 'high-risk' preschool children? *Educational & Child Psychology 23 (2): 15-25.*

Khadaroo, I. (2008) The actual evaluation of school PFI bids for value for money in the UK public sector. *Critical Perspectives on Accountancy* 19(8): 1321-1345

Kurzweil, R. (2005) The Singularity is Near. London, Penguin.

Law, J. (2009) Seeing Like a Survey. Cultural Sociology 2009 3: 239

Lee, R.M (2005) The UK Freedom of Information Act and Social Research *International Journal of Social Research Methodology: Theory and Practice* 8 (1): 1-18

Mackey, E and Elliot, M. J. (2013) Why the Data Environment is Key to Protecting Data Privacy and Anonymity. *XRDS.* October 2013

Martin, D. (2012) Data Technology and Infrastructure. Presentations at the Data Horizons Worksession, University of Manchester. June 2012.

Mason, J. and Dale, A. (2011) Understanding Social Research. Sage

Mason, C.A. and Shihfen, T. (2008) Data Linkage Using Probabilistic Decision Rules: A Primer, Birth Defects Research (Part A): *Clinical and Molecular Teratology.* 82; 812-821.

McCall, G. J. and Simmons, J. L. (1969) Issues in Participant Observation. AWP, USA.

McLuhan, M. (1964) Understanding Media: The Extensions of Man; 1st Ed. McGraw Hill, NY; reissued by MIT Press, 1994.

Mold, A. and Berridge, V. (2007) Crisis and Opportunity in Drug Policy: Changing the Direction of British Drug Services in the 1980s - *Journal of Policy History* 19(1): 29-48.

O'Day, R. and Englander, D. (1993). Mr Charles Booth's inquiry: Life and labour of the people in

London reconsidered. London; Hambledon.

OFCOM (2010) The Demographics of Internet Access - The Communications Market 2010: UK

OFCOM (2011) A Nation Addicted to Smartphones. http://media.ofcom.org.uk/2011/08/04/a-nation-addicted-to-smartphones/. OFCOM

O'Reilly (2011) Big Data Now: Current Perspectives. O'Reilly Radar Team

Ortiz J.R., Zhou H., Shay D.K., Neuzil K.M., Fowlkes A.L. (2011) Monitoring Influenza Activity in the United States: A Comparison of Traditional Surveillance Systems with Google Flu Trends. *PLoS ONE* 6(4)

Pattie, C., Seyd, P. and Whitely, P. (2004) Citizenship in Britain, Values, Participation and Democracy, Cambridge, Cambridge University Press.

Purdam, K., Mackey, E. and Elliot, M. (2004) The Regulation of the Personal, *Policy Studies,* Vol 25, No 4, Dec 2004 pp 267-282

Reading the Riots (2011) Guardian and London School of Economics.

RCUK (2009) Policy and Code of Conduct on the Governance of Good Research Conduct http://www.rcuk.ac.uk/Publications/researchers/Pages/grc.aspx

Rogers, R. (2009) The End of the Virtual: Digital Methods. Amsterdam, University of Amsterdam

Saunders, M., Lewis, P. and Thornhill, A. (2007) Research Methods for Business Students. Prentice Hall

Savage, M. (2009) Contemporary Sociology and the Challenge of Descriptive Assemblage. *European Journal of Social Theory* 12(1): 155–174

Savage, M. and Burrows, R. (2007) The Coming Crisis Of Empirical Sociology. *Sociology*, 41 (5): 885–899.

Seal, A. (2006) Correspondence: UK statistical indifference to military casualties in Iraq. *The Lancet* 367 (9520), 29 April 2006-5 May 2006, 1393-1394

Stanley, L. (2008) It has always known and we have always been 'other': Knowing capitalism and the 'coming crisis' of sociology confront the concentration system and Mass-Observation. *Sociological Review* 56 4, 535-51.

Sturgis, P., Read, S., Hatemi, P.K., Zhu, G., Wright, M.J., Martin, N.G. and Trull, T (2010) A Genetic Basis for Social Trust? In Political Behaviour (2010) 32:205–230

Swan, M. (2012) Crowd sourced Health Research Studies: An Important Emerging Complement to Clinical Trials in the Public Health Research Ecosystem. *Journal of Medical Internet Research* 14.2

Sweeney, L. 2001. Information explosion, In *Confidentiality, disclosure and data access: Theory and practical applications for statistical agencies*, eds. P. Doyle, J. I. Lane, J. M. Theeuwes, and L. V. Zayatz, 43–74. New York, NY: Elsevier Science.

Thelwall, M. and Viz, F. (forthcoming) Researching Social Media. Sage

Thrift, N. (2005) Knowing Capitalism. London, Sage.

Van Zoonen, L., Vis, F., and Mihelj, S., (2011) 'YouTube interactions between agonism, antagonism and dialogue: Video responses to the anti-Islam film *Fitna*', *New Media & Society*, 13 (8). pp 1284-1300

Veneris, Y. (1984), *The Informational Revolution, Cybernetics and Urban Modeling*, PhD Thesis, submitted to the University of Newcastle upon Tyne, UK

Vis, F. (2012) 'Twitter as a reporting tool for breaking news', *Digital Journalism* 1(1).

Whitmarsh, L. (2009). Behavioural responses to climate change: Asymmetry of intentions and impacts. *Journal of Environmental Psychology*, 29, 13-23.

Wind-Cowie, M. and Lekhi, R. (2012) The Data Dividend. Demos. London

Zins, C. (2007) 'Conceptual approaches for defining data, information, and knowledge' *Journal of the American Society for Information Science and Technology* Volume 58, Issue 4, pages 479–493, 15 February 2007

Appendix 1. Outline of Workshop

ESRC Data Horizons Workshop

The University of Manchester

Humanities Bridgeford Street Building

www.ccsr.ac.uk

18th June 2012

Context

As part of the ESRC Digital Social Research programme this worksession will examine the challenges and opportunities posed by developments in the social data environment. We will consider developments in social media, administrative and open data, linkage methods and data support services.

Programme

10.00 Objectives, research reflections and participant introductions (M. Elliot)

10.30 Data Contexts – (N. Contractor)

11.30 Data Horizons - Data sources? (Discussion led by B. Dutton and discussant)

1.00 Data Horizons - Data Linkage? (Discussion led by C.Dibben and discussant)

2.00 Data Horizons – Data technology and infrastructure? (Discussion led by D. Martin and discussant)

3.15 Data Horizons – Methods training requirements? (Discussion led by K. Purdam)

4.00 Overview and next steps (M. Elliot)

Appendix 2. Stakeholder Interview Schedule

Data Horizons - Linked Data and New Forms of Data For Social Research

As part of the ESRC funded Digital Social Research programme, we are conducting a scoping review of the changes and innovations in the use social data and methods and the future. In the last twenty years the data environment in which social science researchers operate and the methods they use have changed considerably and continue to do so.

As well documented expansion in the quantity of data, categorical shifts in the type and form of that data are happening and will continue to do so. Understanding the form that these shifts will take in the short, medium and long term, what the resultant data will look like and how we might use them is vital for the planning and development of the social science resource base and for the training of the next generation of social science researchers.

We would be grateful if you could answer this short questionnaire. All responses will be treated confidentially. It will only take 10 minutes to complete and your input is very important.

1. Reflecting on the last 20 years has the use of data for social science research changed? Y/N

If yes please explain how you see these changes

2. Can you give some examples of innovations in new types and uses of data for social science research?

3. Thinking about the types of research questions that YOU (and your colleagues) tend to tackle imagine that there were no ethical, legal, technical or other restrictions on data – what would your ideal data types be?

4. Imagining now that you can have any data you want how might the research questions that you address in your research change?

5. Thinking of the next few years are you aware of any new data initiatives? Y/N

If yes please give an example(s)

6. Have you collected your own data? Y/N

If yes please describe?

7. Have you ever used Administrative data in your research (add definition)? Y/N

8. Have you linked individual records from administrative data, geographical data, consumer data, something else? to survey data? Y/N

If yes please given an example.

9. Have you encountered any barriers when trying to use administrative data? Y/N

If yes please explain what the barriers were

10. What administrative data would you like to use if it was available?

11. What analytical techniques do you currently use?

12. How do you store the data you use?

13. Have you used any of the special license data sets? Y/N

If yes which data sets?

14. Have you used any virtual safe settings for data analysis? Y/N

15. Do you already use some of these kinds of data or would they be useful for your research? If so which

Electronic Health records - Already use [], Yes would be useful []/No not useful to my research [](add to all)

Sensor data

Individual level records of consumption

Tracking data

Online network data

Tweet Mining

Blog Mining

Clickstream data

Cyberlife data

16. What risks are posed by the growth in the availability of different types of data? (e.g. ethics, access and use, confidentiality, legal compliance issues, good practice, robustness, competing for space)

17. What kind of data training do you feel the next generation of social scientists will require?

18. Do you think that conventional social research methods such as surveys will be less used in future? Y/N

Please explain

19. In twenty years time how do you think social scientists will be working? What data? What techniques? Please describe

Appendix 3. Web Survey Instrument

DATA HORIZONS SURVEY

This survey is an exploratory study of social science data users and those involved in data support services.

Alongside the well documented expansion in the quantity of data, categorical shifts in the type and form of that data are happening and will continue to do so. Understanding the form that these shifts will take in the short, medium and long term, what the resultant data will look like and how we might use them is vital for the planning and development of the social science resource base and for the training of the next generation of social science researchers. As part of the ESRC funded Digital Social Research programme, we are conducting a scoping review of the changes in the data environment and the innovations that will be necessary to use social science data in the future.

We would be grateful if you could answer this short scoping questionnaire.

All responses will be treated confidentially.

It should only take 15 minutes to complete and your input is very important.
You will also be entered into the prize draw for £50 of book vouchers.

1. Thinking about the research questions that you tend to tackle, imagine that there were no ethical, legal, technical or other restrictions on data. What would your ideal type of data be?

Please write in the box below.

2. Imagining now that you can have any data you want, how might the research questions that you address in your research change?

Please write in the box below.

3. Have you ever collected data for your research?

Please tick one box.
Yes
No
If yes please describe the types of data you have collected.

4. Have you ever linked individual records from different data sources?

Please tick one box.
Yes
No
If YES please give an example

5. Have you ever used administrative data in your research?

Please tick one box.
Yes

No
If YES please give an example.

6. Have you encountered any barriers when trying to use administrative data?
Please tick one box.
Yes
No
If yes please explain what the barriers were.

7. What administrative data would you like to use if it was available?
Please write in the box below.

8. Have you used any virtual safe settings for secure data analysis?
Please tick one box.
Yes
No

9. Do you already use some of these kinds of data or would they be useful for your research? If so which?

	Already use	Potentially useful but have not used	Not useful
Electronic health records			
Movement tracking data			
Consumption data			
Clickstream data			
Cyberlife data			
Online network data			
Twitter data			
Blog data			
Special licence data e.g. government surveys			

10. What risks do you think are posed by the growth in the availability of different types of data? .
Please write in the box below.

11. What kind of research methods training do you feel the next generation of social scientists will require? Please provide some suggestions
Please write in the box below.

12. Do you think that conventional social research methods such as surveys will be less used in future?
Yes
No
Please explain the reason for your answer.

13. In twenty years time how do you think social scientists will be working? What types of data? What analytical techniques?
Please write your thoughts in the box below.

FINALLY:

Please indicate your occupation
PhD Student
Researcher
Lecturer
Professor
Other, please specify

Please state the main discipline or policy area in which you work

Please enter your email address if you would like to receive an update of the research. You will also be entered into the prize draw for £50 of book vouchers.

Appendix 4. New Data Type Use by our Survey Respondents

Data Type	% Already use	% Potentially useful but have not used	% Not useful	Total N
Electronic health records	15	56	29	290
Movement tracking data	12	60	29	284
Consumption data	15	62	24	276
Clickstream data	4	40	57	258
Cyberlife data	3	46	51	262
Online network data	13	53	34	269
Twitter data	8	46	47	271
Blog data	8	43	49	272
Special licence data eg government surveys	50	40	10	284