CRESC
Centre for Research on
Socio-Cultural Change

CRESC Working Paper Series

Working Paper No. 74

New Populations: Scoping Paper on Digital Transactional Data

Evelyn Ruppert and Mike Savage (eds.)

CRESC, The University of Manchester and CRESC, Open University

November 2009

For further information:

Centre for Research on Socio-Cultural Change (CRESC)
Faculty of Social Sciences, The Open University,
Walton Hall, Milton Keynes, MK7 6AA, UK
Tel: +44 (0)1908 654458 Fax: +44 (0)1908 654488

Email: cresc@manchester.ac.uk *or* cresc@open.ac.uk

Web: www.cresc.ac.uk

MANCHESTER 1824
The University of Manchester

E·S·R·C
ECONOMIC & SOCIAL RESEARCH COUNCIL

The Open University

New Populations: Scoping Paper on Digital Transactional Data

Evelyn Ruppert and Mike Savage (eds.)

Abstract

In recent years there has been a growing interest in the stance of the social sciences towards the deployment of proliferating amounts of transactional data. This paper is intended to be a resource for stocktaking reflections on these issues. We bring together the views of a number of sociologists, geographers, and business researchers working on these issues, along with input from practitioners with expertise in the use of transactional and administrative data. We do not develop an argument here but rather lay out issues that need to be addressed in future reflections on transactional data, focusing on (i) what is rendered visible and invisible, (ii) embedded temporal and spatial relationships, and (iii) modes of expertise and theoretical resources.

New Populations: Scoping Paper on Digital Transactional Data

1 Introduction

In an era of 'knowing capitalism' the social sciences remain uncertain about how they should respond to the challenges posed by proliferating digital data.[1] It is apparent that there is considerable lack of knowledge amongst social scientists regarding practical issues in the use of digital data. This working paper is designed to be a resource that can aid reflection, being a report of discussions that took place at the CRESC workshop, New Populations: Modulating Transactions and Movements (The Open University, 30th April to 1st May, 2009).[2]

The workshop was organised to launch a cluster of projects related to a new CRESC theme, The Social Life of Methods, which have specific concerns in digital data. It also sought to develop alliances with other interested research parties outside CRESC as a means of developing research interests in the area. Social life of methods (henceforth, SLOM) projects are concerned with how social science methods and data are themselves agents of social change. One of its strands focuses specifically on how social relations are being reconfigured by developments in the collection, storage, networking, processing and analysis of digital data. The workshop aim was to explore developments in the collection and analysis of transactional data in both the public and commercial sectors and how these sources have descriptive potential for the analysis of 'whole populations'.

The workshop began with presentations by representatives from the Office for National Statistics (ONS) on government administrative data and by researchers working on commercial transactional data. On the second day, researchers from different disciplinary backgrounds led roundtable discussions in relation to three themes that were sketched out by the organizers. Part I provides background information on transactional data and a summary of the presentations. Part II is a summary of the discussions organised in relation to three themes identified by the organizers prior to the workshop: (i) Visibilities and invisibilities; (ii) Geographies and mobilities; and (iii) Modes of expertise and theoretical resources for a 'new' population studies.

Part I: Workshop Rubric and Presentations

2 Background to the Workshop

Government practices such as the joining up of administrative data, population and address registers, unique personal identifiers, and biometric passports and identity cards are rapidly becoming central to how individuals are identified and populations are monitored and known. In the commercial sector, users increasingly turn to digital data generated routinely as a by-product of transactions to provide comprehensive or total counts of particular populations (sales data, mailing lists, subscription data, cell phone calls, travel cards). In both cases, the data produced can be understood as on-going and dynamic measurements of the conduct of whole populations: the activities, movements, and transactions of people in relation to both government and the commercial sector. Rather than stable or relatively fixed – as is implicit in Census and even sample survey counts – the population is constituted by these methods as a modulation, continuously changing from one moment to the next. This poses huge challenges to conventional social science methods and theory.

These developments have been facilitated by new information and communication technologies (ICTs), which enable the storing, maintenance, searching and linking of massive volumes of personal identification data, as well as the introduction of new practices such as biometric identifiers. ICTs also make possible practices such as data mining that can reveal patterns and create population profiles, which in turn can be used to make predictions about people and their likely conduct. Through the traces left by individuals in databases, new associations, patterns and correlations can be discovered that were hitherto not visible or predefined.

The technical, legal, surveillance and privacy implications of these practices are the subject of much debate and critical analysis. For example, the political, legal and ethical issues about data being collected for one use only then to be transferred, shared and then redeployed for another use is a matter of much controversy (sometimes referred to as 'function creep') (Surveillance Studies Network, 2006).[3] There are also many methodological and analytic questions related to how such vast amounts of data can be effectively analysed. Issues of data protection and confidentiality are also much discussed, especially due to well-publicized leaks of government data.

Mindful of these issues, the aim of the workshop was focused on conducting a systematic stock taking. First, we sought to interact with government and commercial sector representatives to better understand the development and application of new practices of knowing populations. Second, we sought to engage researchers in a discussion of the kinds of people and populations that these methods make visible, discover and bring into being. That is, if methods do not merely describe but enact and bring into being particular social realities, then what kinds of social relations are being made visible? And what new invisibilities do they engender? What knowledge and governing effects do they produce?

3 Workshop Presentations

3.1 Using administrative data to produce official statistics – Andy Teague and Minda Phillips, Office for National Statistics

The speakers noted that the aim of official statistics is to enable policy makers to make more informed and timely decisions about – amongst other things – services and resource allocation. There are two main ways of collecting the information necessary in order to produce statistics: conducting sample surveys or censuses; and accessing data collected (usually by others) as a part of administrative processes. The latter are extensively used already in producing statistics but largely within specific policy areas (the data is collected and analysed by one department). In recent years, the expansion of Government has created larger reserves of administrative data covering many aspects of individuals' lives. The costs and inconvenience of running surveys, coupled with the greater coverage of administrative sources, has caused the producers of official statistics to consider what scope these data sources offer for improving the relevance, timeliness and quality of statistics. Furthermore, Section 47 of the *Statistics and Registration Service Act (2007)* contains provisions allowing the Minister for the Cabinet Office to make data sharing regulations (secondary legislation requiring parliamentary approval). This enables information to be shared with, and by, the UK Statistics Authority and another public authority for statistical purposes.

ONS itself does not collect any administrative data but is keen to make greater use of data collected by the rest of Government for the reasons outlined above. Andy Teague and Minda Phillips from the Administrative Sources team in ONS described both the current use of administrative data and its future potential as well as outlining the progress made in relation to Neighbourhood Statistics (see www.neighbourhood.statistics.gov.uk) and more recently through the use of the data sharing powers in the 2007 Act. They described some of the lessons that have been learned along the way, including the advantages and disadvantages of administrative data compared to statistical surveys and censuses.

3.2 On the use of name and address records, Richard Webber, Visiting Professor, Kings College London

Given that most transactional databases are rich in transactional information but weak in information about the demographics of individual service users, it is becoming increasingly common for social researchers to explore and mine information implicit in name and address records, which typically form a key element of these databases. From these records it is possible to infer not just the prevailing mores of the neighbourhood in which the data subject lives but also his or her cultural identity and gender.

Such practices are normally treated with suspicion by social researchers who emphasise the need for confidentiality and informed consent, but Richard Webber's presentation reviewed the value of the information that can be accessed from names and addresses, assessed how its relevance and accuracy compares with that of classifiers typically obtained from conventional survey sources and described the areas of social research where this source of information has become a key source of behavioural insight. He particularly focused on the research value of using information on names.

3.3 Andrew Fearne, Professor of Food Marketing & Supply Chain Management, Kent Business School

Andrew Fearne examined the use of Tesco Clubcard data and how this kind of data can be used to understand purchasing behaviour and consumer segmentation. He described the highly detailed nature and scope of supermarket loyalty card data – which can produce individual consumer 'DNA profiles' for every card member – and how it is used commercially, to inform marketing planning and business decision-making, and in academic research to inform theoretical and methodological developments in the area of consumer behaviour.

Through examples of the analysis of dunnhumby data (two years of Tesco supermarket panel data) he illustrated how shoppers can be segmented by lifestage, region, shopping channel, retail format, geo-demographics (Cameo) and lifestyle. While the dunnhumby data offers breadth and depth insights about shopper behaviour it does not directly reveal why shoppers behave the way they do. Andrew Fearne also discussed how further research is required in to understand purchasing drivers (e.g. attitudes, perceptions, motivations).

Part II: Roundtable discussions

Designated discussion leaders led the roundtables and were asked to speak to particular themes and in relation to the presentations noted above. However, the following is not a verbatim account of the roundtables nor does it follow the order of discussions at the workshop. Rather, we have imposed a particular analytic ordering and narrative to reflect the concerns and points raised by discussion leaders and participants. The points developed below therefore reflect the nature of the discussion in which different perspectives were put forward and therefore need to be read as part of an internal debate rather than a coherent or necessarily consistent position.

4 Visabilities and invisibilities

The discussion was led by Kirstie Ball (with Anna Canhoto). The starting point of the discussion concerned who is rendered visible or invisible by linked digital data and practices such as data matching, data mining, predictive analytics and profiling that are used in both the commercial and government sectors.

Despite the advancements in the availability and analysis of transactional data, many social scientists continue to insist that the survey is a powerful research instrument that cannot be completely substituted. It is certainly true that surveys make qualitatively different types of inferences about subjects. However, transactional data also replicates some of the information provided by surveys. A key question then is in what ways does transactional data (and various forms of analytics such as data mining) provide different or greater insight about similar things? Generally, it can be said that the increase in transactional data and digital analytics is related to the rise of tracking and tracing tendencies of surveillance technologies. Whereas surveys focus on (sampled) individuals and hence assume these to be the centre of analytic attention, transactional data focuses more on specific transactions, which are more amenable to be understood in network, associational, and relational terms. For example, whereas surveys might focus on why person A likes to buy bananas, analyses using transactional data might focus on what kinds of other purchases take place when bananas are bought. However, it should not be assumed that this is intrinsically a less important question than the first: it depends on whether individuals or networks are the centre of analytic attention.

What this suggests is that different subjectivities are created and made visible (and others invisible) by transactional data. For instance, transactional data can be considered a register of an ‘actional’ presence or the ‘performative’ aspects of subjectivity in that it consists measurements of what people ‘actually’ do, their habits and preferences. In this regard, transactional data could be considered as evidence of conduct. However, it is not simply a ‘description’ or recording of what people do but a categorisation of that conduct which is built into data-capturing systems, normally at the ‘point of sale’ or ‘point of service contact’. So while transactional data measures what people do, what they are ‘doing’ is pre-classified, pre-formatted and preconfigured, and highlights specific kinds of transactional ‘exchange’ processes.

Transactional data represents a new territory for understanding the production of subjects, an understanding that is no longer confined to separate spheres or domains of social life. There is much ‘bleeding over’ between government the commercial sector data collection. Commercial transactional data produces different categories that overlap with, displace and co-mingle complexly with the data that is produced through government surveys and administrative practices. For example, government surveys at one time used ACORN[4] categories to stratify samples. Likewise all commercial classifications to a certain extent rely upon benchmarking against census data. So there are no clear-cut demarcations between these different systems of data and classification. This also extends to ‘do-it-yourself’ self-categorisations. Web 2.0 applications, mashups and applications like Google APIs[5] people can challenge or reconfigure representations and categorisations of themselves, locations and places.

Additionally, these understandings of subjectivity are performed in relation to objects. In the case of clubcards, subjectivity is in relation to products (bananas, as in the example above). The subject is understood in relation to the things she consumes or does not consume, and it is through the recording of these relations that social distinctions are then made. But subjectivity is not only defined in the relation to the consumer object: cards, scanners and barcodes also mediate it. Along these lines we can consider how other objects track and make visible different kinds of actions, transactions and relations. Devices such as Personal Digital Assistants (PDAs), mobile phones and MP3 players keep records of transactions (searches, applications/music purchased, locations visited and so on) and transmit that data to other networks and actors. One way of thinking of these devices is to consider them as ‘logjects,’ or logging objects that monitor and record usage in some fashion. These things become ‘alive’ and are trackable and trace what Dodge and Kitchin (2009) call ‘permeable knowledge acts.’[6] Indeed, product barcodes and clubcards can also be considered logjects that contain unique identifiers that enable tracking and tracing. While there is a long history of unique identifiers such as the address, logjects are not fixed, are carried around and thus introduce animation and traceability. In relation to subjectivity then, it is movement that creates social distinctions. Instead of where one lives, where one goes, travels and visits become important identifiers.

Such tracking and tracing techniques are also familiar tools through which individuals come to identify who they are. There are numerous programmes and software that people use to monitor and track their dietary habits, training programmes, levels of happiness, children's sleeping patterns, etc. These techniques are also deployed to discover patterns and reveal things about who we are based on conduct. In this way technologies of knowing populations are not alien to everyday practices of knowing oneself. But like logjects and clubcards, these techniques work on the basis of pre-defined categories, which must be taken into account when interpreting data. For example, to what extent does Tesco get the kind of data it wants and in the process excludes 'messy' data or data that may 'skew' results, and thereby decides on what is significant/insignificant? Or in other words, how do clubcards make visible certain aspects about subjects while discarding others?

Beyond organising and classifying the conduct of subjects, categories that make up transactional data also intervene and are involved in constituting the subject. There seems to be a circular relationship and reinforcement happening between categories that the clubcard measures and peoples' behaviour and preferences such that transactional data comes to reinforce the behaviour it has set out to measure. In this regard, like other techniques and practices, transactional data not only inform the corporation but is bound up with the making up of people. We can perhaps think of this as a feedback loop between shoppers and the kinds of data that is discovered about them: the data is used to market particular coupons and advertisements to particular shoppers who then start purchasing other things thus producing new data that is again evaluated and categorized and then used to market new coupons and so on. The ongoing practice of being a shopper is constantly changing because obviously the corporation does not stop monitoring the shopper. The shopper is constantly changing, modulating and has to be tracked. Data systems are thus also in a process of evolving alongside the shopper and are not separate from her. We could say both are co-constituting and thus we cannot speak of the technique and people but about how they are bound up together. It is also possible that the revelation of shopping behaviour enables subjects to think differently and change their way of behaving and to buy different products, for example. Through novel uses of data and through novel ways of thinking about it, new market segments could thus be created as shoppers become more informed about their behaviour. The same arguments and logic could be applied to 'joined-up' government administrative data. The technique of knowing population is bound up with the constitution of groups, which become actionable and governable. For example, practices that identify children at risk lead to programmes that so label and target children and seek to change them. But in either case it is understood that the creation of the subject is not simple or straightforward. There are many knowledge and governing practices occurring at myriad sites and which are sometimes contradictory (think for example about the making of the shopper). Furthermore, there is an assumption that the subject is rational: that once behaviour is revealed, the subject will change. However, the consequences of revelations about behaviour are not straightforward or easily mapped onto behaviour.

These issues point to some of the potential biases of transactional data and data mining algorithms that are embedded in the decisions of statisticians such as what is included and excluded in calculations, the rounding up variables etc. In general social scientists need to pay more attention to how statisticians do this kind of processing work and reflect critically on the robustness of their analyses. But this is no simple task. It is more and more difficult to track and trace decisions given the decentralised, complex, multi-stage and multi-actor processes of data collection and processing. For example, consider two decision-making processes. The first is that of bank loan approvals, which used to be based on a subjective decision by a bank manager who would have taken a decision relying on categories like income, profession, occupation, and what might be considered as a construction of more settled categories in a face-to-face interaction. The introduction of multivariate methods and the generation of credit scores did not necessarily make this process cheaper or quicker but better aligned it with the risk of lending. Whilst there are methodological issues concerning how credit scores are created and how much weight is given to different variables and the sort of data used, the results are deemed more 'objective' and aligned with risk rather than the bank manager's 'personal' judgement. An equivalent example is how a border guard would have taken a decision in the past as opposed to now where decisions are based on pre-screened risk calculations. In both cases though, notwithstanding potential disagreements about which process is more subjective or objective, the question becomes, where is the decision made? Is it the person who designs the algorithm, who collects the data, or is it the bank manager or border guard who still makes decisions but uses calculations that come up on a screen? Where are those judgements and discriminations made? Where do we locate them?

To answer such questions requires identifying the mediators and translators throughout the process, particularly in relation to administrative data and the processes and bureaucratic procedures through which it is collected, which are potentially more complicated and numerous. It is a decentralised, complex system and there are many operations involved in categorising and recording data.

Privacy, research ethics and consent

Transactional data raises a number of ethical issues, specifically in relation to consent and disclosure. Every day people are leaving data trails without their consent because they are carrying a logject in their pockets and are generally unaware that as a result of their use and interaction with objects data is being recorded that someone can exploit. On the one hand, there may be generational differences in concerns about privacy and surveillance that this raises. For example, many young people engage with surveillance as a source of pleasure through disclosures on Facebook, and use such platforms to study their relationships, social structures and social networks. In this regard sociology and social research have become part of popular culture and entertainment.

But then transactional data raises a much muddier question concerning what exactly constitutes personal data? For example, is a person's name personal data? (The question arose in relation to the practice of using names from organisation mailing lists to infer ethnicity). If so, then a researcher wanting to use names would have to approach an ethics committee and address the ethical aspects of the use of names as well as ensure that subjects consented to the kind of data analyses to be conducted. Similarly, a data-collecting organisation can only give their data to a third party if subjects have agreed that the data can be used for research. From the government perspective, ethics underpin all activities related to transactional data mining. Transparency is a key goal and this is what one would expect from the sharing of transactional data, which has been collected for specific administrative purposes but not for research or statistical uses. What is at issue is the use of data for purposes other than that for which it was collected. For every case the ONS has to engage a data sharing order with the Information Commissioner's Office to ensure that what is proposed is fair and complies with the fair processing principles of the *Data Protection Act.* In addition the ONS has to ensure that when respondents provided their data to the original data owner they gave explicit consent that it can be used for statistical and research purposes. If that was not the case, then some steps have to be taken. ONS can't have access to retrospective data and can only have access to data where the confidentiality pledge has been changed and then data can be shared from that point forward. These procedures bring to the fore another difference between methods of data collection. Direct methods of data collection, surveys and censuses involve interaction with a researcher and thus the individual is aware of the process and the method is transparent. New methods of collecting data are not as transparent. The way in which the data is both collected and used is not obvious and that represents a considerable challenge.

Questions of what is private and public data are interesting in relation to cultural comparisons. Population registers and identity cards have long been part of the administrative systems in Nordic countries and the joining up of administrative data is also advanced in these contexts. There is a very different understanding of informational privacy, a sort of social welfare outlook that says information about individuals is a public good; it's about what is good for the collective in terms of the distribution of resources and rights. This reflects a different attitude about government and public trust. Data sharing provisions in the UK also stipulate that the sharing of data is for the public good.

While there are many debates about the ethical aspects of using transactional data the reality is that in the commercial sector several companies are already using transactional data (e.g. data collected through websites) for marketing purposes. The world is not organised as university ethics committee demand and thus the social sciences need to find ways of taking into account the limited power of bureaucratic and professional regulations in defining this world. One of those new realities concerns the ethical consequences of new actors and long chains of decision-making involved in the use of analytic devices such as data mining. Often it is computer scientists who design equations or algorithms along with caveats about their application to particular contexts. But when taken up and abstracted from a particular context and applied further down the line and in different contexts those caveats are lost. Is there an ethical issue here? In relation to this a number of other questions arise which beg further analysis and debate. Given that both are abstractions, does a survey or sample have any greater integrity than associational analysis based on data mining? Survey or census data are also subject to abstraction, many types of computations, etc. so that data becomes ever more autonomous from the subject. Are there more chains of expertise and analysis involved in the mining of joined up transactional data? The point is that there are numerous judgments and decision points involved with all of these techniques. Is decision-making, responsibility and accountability clearer with surveys than with data association types of analysis where algorithms make judgements and thereby remove subjective components? With new inferential types of analysis, the decisions about apportioning, segregating and designing involved in data collection and analysis processes seem to be less visible than with traditional survey data.

Perhaps this is not so much an issue of ethics as it is about the power to generate, organise and utilise knowledge. Decision support systems do not make decisions but support human decision-making. Credit score systems will either put people in boxes or create flags, which then call for human intervention. The same applies to border security systems or GPs who want to prescribe drugs. Decisions about the design of a system should thus be made between analysts, managers, and clients.

Who decides on ethical standards and whose interests are served by particular ethical standards are a matter of some debate. These issues give rise to the more general concern that legislation has not kept up with knowledge practices and technological capabilities. For example, even when confidentiality, anonymity and consent have been acquired, other issues arise such as the use of such data to make inferences from the generation of risk profiles. The UK Information Commissioner recently said that his concern was not so much the use of data for which consent has been acquired, but more about profiles created on the basis of data and how they are used to flag 'risky' individuals. It is not possible to inform somebody at the point of data collection that this is how their data is to be used because such use is far removed from the original transaction and is based on inferences. This is the difference between adherence to the *Data Protection Act* versus the uses of new technologies to interpret data to infer relationships. The distinction is very problematic and legislation has not kept up with this. Legislation also does not recognise that that in addition to a legal subject there is an inferred, projected figure of the legal subject, a refractured subject inferred from data. Systems that data miners call 'humane' meet the concerns of the Information Commissioner's Office because the data is anonymised and unique identifiers are only put back in when a 'risk flag' is raised. It is at that moment that the system says 'this is the person' associated with a certain risk factor. This occurs in a number of practices, from determining who can get mortgages to assessing flood risks. What does data protection mean at this stage when people are identified in relation to risk categories? How can privacy issues be incorporated into practices of data mining?

Thus there is another angle to consider in relation to ethical issues that goes beyond concerns about individual privacy and confidentiality. For one, transactional data can be analysed in ways that involve exploring non-obvious relationships in data and creating equivalences. That is, relationships between categories can be 'discovered' in the data and conclusions drawn and then mapped onto people's lives. Fully anonymised data can be used to identify hitherto 'unknown' groups and on this basis interventions can be defined. The consequences of this can be positive as in the case of the identification of groups or areas experiencing deprivation, which then can be targeted with remedial resources. But the same practices can also be used to define and identify 'risky' groups who then become the target of disciplinary programs or exclusionary practices. Transactional data can be used for deciding on who can receive a mortgage based on addresses: people living in neighbourhhoods classified as deprived are less likely to be approved for mortgages and thus the classification of deprivation is reinforced. People profiled as eating healthy food receive coupons for healthy food, whilst others receive other types of coupons and thus social stratification is reinforced. In this way discriminations and social stratifications can be institutionalised through transactional data. The point is that practices of identifying populations are not benign or objective; they bring into being particular populations in order to render them governable. And in doing so governing interventions may very well reinforce the 'identity' of a population so discovered. Therefore the procedures that have led to the 'knownness' of populations need to be interrogated.

Clearly research is required to better understand the consequences of these new technologies. For one, technocentric discourses tend to emphasise what technologies can do, what they can achieve, and how they can make things better. Ethical implications are not made visible and are seen as add-ons or afterthoughts. Decisions involved in the production and analyses of transactional data thus need to be made more transparent. This is especially so since claims are often backed up by numbers and figures without due consideration of their scientific basis and then dispersed and disseminated through the media. For example, a footnote in Clive Norris's book *Maximum Surveillance Society* stated that there is about one camera for every fourteen people in Britain and that people are captured on CCTV 300 times a day.[7] That statistic has been cited everywhere, however, it was only an estimate based on a hypothetical scenario yet has come to take on a life of its own.

On the implications for the social sciences

While there are important political and ethical concerns about the use of different forms of digital data, the reality is that a large number of government and commercial organisations are extensively using this data. The methodological implications of this data and the various uses and applications to which it is put have not been investigated in sociology and the social sciences generally. Consider for example how logjects become data collection tools whereby the object is the instantiation of the data rather than a researcher who interviews someone or extracts data as an act. The collection of this data and the traceability of subjects and transactions in real time that logjects make possible generates complexity and an amount of data that social scientists are yet to interpret. One reason is that social scientists are not used to working with the organisations that generate this kind of data. The challenge is to examine how we can methodologically innovate, visualise this data, develop theoretical and methodological frameworks to analyse these emerging practices and to give alternative renditions of this data in powerful ways. Social scientists need to align and collaborate with organisations generating transactional data, who generally are not concerned with 'why' questions but only with knowing behaviour. Perhaps then one could argue that the main contribution the social sciences can make concerns insights into ‘why’ people behave the way they do.

Another example of transactional data that is ‘out there’ and routinely being generated is from games. Not only is gaming a major economic activity but it is also now largely conducted on-line and tracked. Eleven million people are spending on average 22 hours a week on ‘World of WarCraft’. Yet there is no developed social science of gaming or sociology of on-line gaming. Much could be learned about tracking and tracing in the non-virtual world through an understanding of these virtual worlds. For example, there is a lot of work in geography on the relationship between gaming and military techniques of tracking and tracing. There is also plenty of literature and work about affects, about the ways that feelings, playfulness and vivacity are at work in gaming. Social science analyses that do investigate these emergent practices often do not provide thorough descriptions of how practices actually operate. They may provide some theoretical insights or conceptual frameworks but little work has focused on investigating and understanding mundane realities and everyday practices. What is needed is mid-range work that involves solid empirical analyses rather than sophisticated theoretical perspectives that are not grounded in robust empirical data.

This concern for developing new research methods and approaches is very much connected to the increasing interest in examining the social impact of research, which is also a stance of the ESRC. As is well known, the traditional research model involves the researcher collecting data, analysing and interpreting it, coming up with conclusions and recommendations, and then disseminating results. Increasingly through practices like the Research Assessment Exercise (RAE) researchers are further required to show their impact. In light of the discussion above we could say that a completely new approach to social impact is required that involves social interaction between the researcher, the object of research and social processes.

But there is also another perspective on the implications of new forms of digital data for the social sciences. It is a perspective that understands the increased use of transactional data as part of a more general expansion of methods of control, discipline, metricisation and surveillance. David Lyon has argued that this move is contributing to a kind of ‘social suicide’ whereby social relationships are being replaced by suspicion, tracking and impersonal forms of monitoring. What then are the implications of transactional data becoming a major source of sociological analysis? Would this constitute a kind of slow ‘sociological suicide’ whereby understandings of social relationships are reduced to transactions, movements and networks? In other words, not only do the social sciences need to re-engineer methods in relation to this new world of digital data but also analyse and identify the ontological, epistemological and governmental consequences for our understandings of the social.

5 Geographies and mobilities: Temporal and spatial dimensions

The discussion was led by Roger Burrows, Louise Amoore and Eleonore Kofman. The starting point of the discussion concerned how in an era of greater mobility and tracking and tracing technologies, digital data is reconfiguring the understanding and governing of social and spatial relationships.

‘Geography is the new sociology’. This is one way of characterising how the social sciences are being challenged and reconfigured by new forms of digital data and technologies. As many of the examples above illustrate, the kinds of analyses of social relationships that are being developed are often based on location as an organiser and identifier of subjectivities. For example, in the commercial sector, analysts have been doing sociology, whether they have realised it or not, through categorisations such as MOSAIC, which are like sociological descriptions of places based on a huge amount of transactional data that social scientists do not routinely have access to.[8] But there are also popular practices that are leading to what could be coined new cartographies of neo-calculism.[9] These practices involve people using the Internet and software like Google Earth, Google APIs and mashups to play around with space and to take up social science and geography as hobbies without realising it. The term mashup comes from music whereby a background track is taken and another vocal track is put on top of it, and thus two things are mashed up together. But the term now refers to any Web 2.0 application that takes two or more different data sources and mashes them up to create something new. Someone with some basic tools can play around with data freely available on the Web to do their own representations of space and to interpret associations that they discover in the data.

There are many good visualisations and tools originally developed by the social sciences that are now available as Web 2.0 applications and are being used by children and students, and which make some social science tools like SPSS look outdated.[10] From simulations to computer games, children interpret the world not through abstract equations but through playing with virtual entities. What was possible to do in GIS (Geographic Information Systems), a 10-year-old can do in 10 minutes. Streetview on Google Maps, though raising issues of privacy, makes it possible to do mashups with photographs linked to Wikipedia. These things are already tremendously helpful tools and are not just play things but also ways of describing and contesting places. They represent a challenge to the social sciences in many ways and are not being regulated by ethical considerations or any kind of research methodological considerations.

The culture of tracking and tracing subjects and objects means that we have entered a new mobility paradigm. We have moved from representations of space to the mobilisation of space. Hence it is necessary to move away from understanding the ‘frozen’ shape and pattern of social structures and spatial differentiations to an understanding of the shape and pattern of the movement of objects and actors, and their differential movements. Data on postcodes is a frozen geography; instead, we need to consider actors as bearers of codes that can be tracked and traced. Setting aside privacy issues for the moment, what tools do we have to understand how different people and things move, where they cluster, and so on? For example, instead of thinking of spatial segregation within the city in terms of where people live consider the segregation of spatial mobilities such as patterns of movement on public transportation systems. The challenge is to get some sense of the animation of social movement.

If it can be said that geography is now sociology, then perhaps geography is no longer geography. It is not in its conventional sense geography – in terms of its origin ‘geograph’ or the graphing of the world. Indeed, contemporary geography is trying to invent novel forms of mapping and drawing lines and one version is a kind of mashup. It involves the analysis of transactions generated by movement and border crossings and how they are being deployed in security practices to make something that’s uncertain in the future - that cannot be predicted in a conventional sense of predicting from data – amenable to security decisions and interventions. It involves identifying associations in data, which is not accomplished by the actions or decisions of any one actor and indeed it is very hard to pinpoint where decisions are made. For example, the UK e-Borders Programme records data on all entries and exits and makes this available to authorities, who then mine the data for associations. They call this the ‘joining up of the dots’, which is really a kind of mashup. For example, the 20 items of information submitted to a commercial travel provider are stored on the PNR database (Passenger Name Record) and shared with public authorities. Analysis of this data does not involve comparing or screening mobile populations against a norm. Rather data is ‘flushed through the analytics,’ which is not a filter that somehow captures mobile bodies that deviate from some kind of a known norm of the population. Instead the norm itself is mobile, a modulation such that populations are a ‘differential curve of abnormalities.’ The mobile norm works according to rules and logics of association, for example, ‘if this and this, then this and this.’ So it’s not the data itself but the drawing together items of data into associations that matters.

These practices also introduce a new temporality. With transactional data the focus is on the future, on populations yet to come, whose dynamics are as yet unknown. In relation to Tescos, the golden key is not about what the customer looks like on a day-to-day basis, but what might be the desires, interests and ambitions of the unknown consumer when she walks in the door. For the e-Borders Programme, the golden key is the unknown terrorist, the person who may or may not carry out violence. This kind of knowing is different from a survey. It is a projection that allows something to appear out of the gaps in the data. This is what risk flags do—they render what is not known (gaps) amenable to management and government. We don’t know, for example, what the relationship is between a particular flight route and method of payment but we can act on the basis of what we do not know. In that way an unknown future is converted into a decision in the present.

This is a challenge to the traditional way that population is understood in the social sciences, as a population where specific individuals have traits and behaviours that can be identified and used for the purposes of planning and intervening. In the governing of borders and security the emphasis is different. First, in the context of the overwhelming volume of data decision-making is not focused on collection but on what data is to be discarded. Furthermore rather than looking at traits and trends or patterns, one looks for potentialities, proclivities as inclinations, as something of the future that we don’t quite have yet.

Another issue concerns how the subject is visualised in the data and what is unique in this type of visualisation. For example, how does the ‘screen’ itself become a space of governing population? As discussed in the previous section on visibilities, the border guard’s screen is where decisions are made on the basis of calculations and algorithms that determine what data to discard, and where a risk flag or profile on a person appears. What happens on that screen and what is the relationship between the screen visualisation and the judgement the border guard makes? Does the screen replace a face-to-face decision with an already screened, programmed calculation? If so, what are the implications?

This is a different kind of visualisation than that of surveillance. To ‘survey’ means to bring particular subjects into play but it is also a particular way of seeing. Techniques such as e-Borders are not about capture and collection but about discarding, projecting and visualising the ‘future yet to come’ in a way that makes present security decisions, financial decisions, etc. possible. Once again we need to ask who makes decisions and where is the human agency in this process. Is it a form of ‘machinic’ or algorithmic agency where algorithms created by humans gain some autonomy and can even generate new algorithms? These algorithms do work that no one really understands. They are not based on numbers but other, non-numeric data and on pattern recognition, which are challenging our power of understanding. Compare the work of algorithms to pattern recognition not in a numerical sense but in having a hunch, a feel. Just having an intuition is no longer used as a way of seeing and approaching data. Thus one issue that digital data and analytic devices such as computer algorithms raise is the ability to bypass more qualitative notions of assessing risks based upon day-to-day interactions rather than probabilistic assumptions.

In relation to security it is interesting that on the one hand there is an appeal to use all of one’s senses to detect unusual behaviour for example, and on the other hand, there is a tendency to substitute senses with these kinds of visualizations. Lorraine Daston and Peter Galison (2007) have written an insightful book about this called *Objectivity*, which is a history of how things come to be seen as objective.[11] For example, in medical technologies how a pulse reader replaced a doctor’s touch. We see it certainly in border controls, where a pat-down search is replaced with forms of risk visualization and on the London Underground posters of people’s eyes, ears and lips enlist travellers to 'use all your senses' to detect the suspicious. So here we have both an appeal to use one's senses in order to ensure security and safety and, on the other hand, practices that evacuate our senses. Furthermore, these techniques have implications for the evaluation and exercise of rights, whereby data and associations in data, for example, instead of the presentation of a passport, determine the right to cross a border.

To be sure, questions about how to govern dispersed, mobile populations are not new but there is a new relationship between mobility and security being forged such that movement itself is becoming a means of securing. For example, at the heart of Olympics planning is how to extend e-Borders pre-arrival screening into the spaces of St Pancras International and Stratford Station ticketing systems via the Oyster card system. These spaces have been described as a smart verification gate that would know when to open and when to close at particular times. The gate is not a fence or boundary in the traditional geographic sense. It is a gate that knows when to be open and when to close. It is a kind of risk-oriented gate. So in that sense a transaction takes on a particular dynamic and comes to mean 'passing through.' In general this constitutes a new way of thinking about population that allows for reconciliation between the image of globalisation, of open gates, and smooth surfaces for mobile people and at the same time the security of the state. As Foucault asserted governing is about freeing up movement and in a similar manner we can think about how joined up government administrative data is also about freeing up the individual and her movement and transactions with government, and about improving her movement through an individualised approach that involves knowing who she is by joining up data about her.

But mobility and immobility is also co-present. On the one hand posters for the US Visit Programme say: 'Keeping America secure but its doors open to business,' which gives a sense of a movement but at the same time a line has been drawn elsewhere, a decision has been made about admission and admissibility. That is, there is also simultaneously immobility determined through a normalisation process that makes 'open gates' possible. In this way, mobility has become a stratifying factor that results in two different groups whose mobility is treated differently. There is a group whose mobility is encouraged (tourist, business person) and there is a group whose mobility is feared (asylum seekers, illegal migrant).

But so far it is the transacting and moving person that has been the subject considered. What about people who are not captured by databases because they do not transact? Such people also happen to be groups that the state would like to know a great deal about but who try to limit their transactions in order not to be detected (e.g., irregular migrants). They avoid crossing borders, are not eligible for benefits and so are not recorded on administrative databases and pay in cash so they do not appear on debit or credit card databases. However, exclusion also has other consequences such as inequities in the allocation of resources and rights: knowledge of groups is often necessary to identify inequities, for example. In this regard, transactional data can be limited because it usually doesn't capture gender, ethnicity or disability. On the other hand, it can be more flexible and provide new categories relevant to peoples' lives and life chances. In sum, there are a number of questions about the nature of transactional data, what it means to combine data from different sources and sectors and the dynamics and consequences of being included or excluded.

6 Modes of expertise and theoretical resources for a 'new' population studies

The discussion was led by Mike Savage and Evelyn Ruppert. The starting point of the discussion concerned what theoretical resources can social scientists draw from to critically analyse, understand and interpret the effects of identifying populations on the basis of transactional data and the related practices of data matching, mining, and profiling.

The foregoing discussion identified that at present there are many different organisations and agencies outside the social sciences that are constructing understandings of the 'social'. It is important to place this in historical perspective, where during the second half of the twentieth century, social scientists lead and elaborated various techniques and methods (such as sample surveys, self-completion questionnaires, the interview, case studies, ethnographies,) that became very powerful across different domains and disciplines. Social scientists focused on procedures for extracting information from subjects in a way analogous to surgery: How do you intrude into the social body to take a sample of tissue? This is how social science technologies developed. The post-war period was a kind of golden age for the social sciences and their engagement in the engineering of the social. As indicative of this consider that in 1947 only 3 percent of university academics were social scientists, and most of them were in the humanities or medical or natural sciences. By 2001 (depending on how you define the social sciences, and if you take a broad definition that includes business studies, etc) they made up a third of a much larger university sector.

However, now with the emergence of transactional data, this expertise about the social is no longer so secure. The difference concerns not only expertise as such but also the way data is obtained. Previously there was a need for going into the field and collecting and extracting data, which now has been substituted by data that is generated as a by-product of everyday actions and transactions, and hence appears to need no special social science expertise to generate it.

The social sciences also invested a lot in procedures for collecting data based on a conception of the individual and the possibility of understanding him through an interview, for example. The technique was based on the psychotherapeutic encounter and became a key technique of social scientists. Nikolas Rose (1999) has written about this in terms of governing the soul, shaping the private self and the relation between political power, expertise and the self.[12] At the time this competed as it were with another approach to constructing knowledge of the social associated with Field Theory, which was developed by the social psychologist Kurt Lewin who borrowed some notions from physics and brought them into psychology. He was not concerned with individuals but with ties, connections, and networks. However, the approach was critiqued in the 1950s by proponents of approaches that understood the individual as the key social unit, and it was this understanding that became enshrined in the techniques of the interview and survey, and in mainstream social scientific analysis. The result was that social network analysis was largely abandoned in the 1960s with the development of large sample surveys and analysis. In relation to transactional data the social network approach and 'association' logic are now coming back in a fundamental way. What appears as the object of interest is not the individual but the connections between things and people. One of the leading American social network analysts is the sociologist Duncan Watts (2003) who wrote *Six Degrees: The Science of a Connected Age* and who is now working for Yahoo on their web technology.[13] It's interesting how this network methodology is now being taken up in relation to new digital data sources. In these approaches it is not the social attributes of the individual (e.g. gender, class) that matter but patterns of association in data, which can be fluid rather than fixed and categorical. Social scientists are rarely involved in the development of these new network methodologies and transactional data analyses. The authorities who are now doing social science and trying to intervene in the social world are drawing on expertise from areas outside the social sciences, and in particular that of Information Technology and Artificial Intelligence experts. However, at the same time many of these methodological developments are also building on techniques designed by social scientists such as factor analysis.

These developments in the role and expertise of social scientists apply not only to techniques and methodological practices but also to the sociological imagination itself. C. Wright Mills wrote that you do not find the sociological imagination in departments of sociology; you find it in history, you find it in journalism and you find it in novels and dramas. Where you find the sociological imagination today is also in popular culture. Television programmes like *The Sopranos*, deal with sociological issues such as structure and agency in a dramatic form and invoke the sociological imagination. With Web2.0 applications, people are playing around with social networks and social issues. The most popular games are the Sims and Reality TV programs as social experiments. Bev Skeggs has written a lot about Reality TV, and has come up with a nice phrase that relates both to the sociological imagination and some of the issues discussed above about the constitution and construction of subjects: 'It's all about the grammar of conduct,' she says. These techniques are all about how to live, how to help people make choices when they have no option other than to choose, and how to map out the dominance of particular middle class taste and preferences for food or clothing and out of this construct identity categories.

However, social science expertise has continued to invest in improving techniques such as questionnaires and surveys in order to know what is understood as a self-eliciting or attesting subject. While the social sciences need to engage with the new empirical reality of digital data and analytics, they also need to develop conceptual tools for understanding the ontological differences that these data and techniques enact and their governmental consequences. For example, how is the subject conceived and understood by these techniques? How is her role and agency being reconfigured, mediated or altered through these different practices? Different technologies and people are engaged in these practices of creating and analyzing transactional data. There are many decision points and actors involved in a long chain of relations that make people legible to themselves and others. We need to think through administrative systems and how in practice these create transactional data and the creation of a legible person and their translation into data involves negotiations between humans and technologies.

For example, if we compare how data is compiled by surveys or censuses to that of government administrative data it is clear that subjects are engaged in different ways and with different consequences. In a survey what is elicited about a particular person is inconsequential and unverifiable. However, with government administrative systems there are major consequences of not being identified in a way recognised by a government classification system and verification is a matter of some interest. Through administrative systems the subject has less opportunity or chance to do anything otherwise than that which the government classification or categorisations demand. What are the consequences of identifying populations on this basis? As Bruno Latour might say, a different set of agencies, objects and subjects are engaged and involved and as a consequence different kinds of data and identifications result. What are the consequences and differences between populations based on classifications of what people say about themselves versus what they do in relation to administrative or transactional systems? With transactional systems there are also variations in the methods or means by which data is collected. In some cases data is collected based on a face-to-face interview such as an applicant for benefits where an administrator asks questions and makes decisions about what should be entered into the system often without much concern for accuracy. Compare that to people transacting on the Internet and applying for services online where validation routines can be built into the system forcing people to categorise themselves in prescribed ways.

There are numerous theoretical resources that can be drawn on to interpret, analyse and conceptualise these developments in transactional data and analytic techniques. Here we outline just a few. Foucault's distinction between two types of surfaces through which the governmental acts upon the social can be taken up to understand the differences between the self-eliciting subject and the 'traced' subject. The former is through the 'public,' by which he refers to acting on people through their beliefs, thoughts, desires, and practices. The other is the 'milieu,' which is an understanding derived from the natural sciences. Whereas government acts on population through its subjectivity, through what it knows about its beliefs, desires and so on, the milieu implies acting on population through the interfaces connecting people to each other. This can be applied to the present examples: the same data being collected, processed and mediated can be connected to the population through public beliefs and ideas but also these forms of data can also be used to approach the population through the milieu, that is, as a surface, as a series of interfaces and connections.

The varied cultural legitimacy of academics, government departments, private sector companies, and voluntary organisations can be examined in relation to Andrew Abbott's ecological perspective, which is based on the idea that there is contestation between different kinds of experts claiming jurisdiction and engaging in disputes over the diagnosis and treatment of problems. Different expert groups lobby to offer their expertise with different kinds of legitimacy and effectivity. This can be linked to Bourdieu's conception of how different forms of expertise constitute particular kinds of cultural capital that influence the capacity of different actors to command attention and define populations. For example, the experts who dominate the development and deployment of new digital analytics tend to be information technologists and artificial intelligence experts rather than social scientists.

Foucauldian analyses such as those advanced by Nikolas Rose investigate how discourses are implicated in the formation of populations. Rose argues that neo-liberal governance produces the 'person' as a consumer (hence, transactional and administrative data) and the nature of expertise thus changes from being legislative (in Bauman's sense) to 'administrative.' This calls for investigations of how proceduralised forms of expertise are being organised and what is involved in systems based on an economy of audit. Science and Technology Studies (STS) such as the work of Bruno Latour investigate the role of inscription devices, and the ways that certain practices can constitute themselves as 'obligatory points of passage'. For Michel Callon, John Law, and Donald MacKenzie methods are performative, they do not simply describe the world as it is, but also enact it. Nigel Thrift, Scott Lash and Manuel Castells have investigated the way that informationalisation has become embedded into the routine organisation of economic, social and political life. What are the powers of numerical, textual and visual sources of information in this context? In the context of Lash's claim that informationalisation does not allow critique (or 'analysis'), in what ways can informational data be used for research purposes?

In sum, the social sciences need to engage with new forms of data and analytic techniques to undertake rich empirical analysis as well as develop new concepts and theoretical resources for understanding the ontological, epistemological and political consequences of these ways of knowing and governing the social.

Appendix 1: Discussion Leaders and Workshop Participants

Organisers

Evelyn Ruppert, CRESC, The Open University

Mike Savage, Department of Sociology, The University of Manchester

Presenters

Andy Teague, Office for National Statistics

Minda Phillips, Office for National Statistics

Andrew Fearne, Kent Business School

Richard Webber, Kings College London

Discussion leaders

Roundtable 1: Visibilities/Invisibilities

Kirstie Ball, The Open University Business School

Ana Isabel Canhoto, Henley Business School, University of Reading

Roundtable 2: New geographies/mobilities

Roger Burrows, Department of Sociology, University of York

Louise Amoore, Department of Geography, University of Durham

Eleonore Kofman, Health and Social Sciences, Middlesex University

Roundtable 3: Expertise

Mike Savage, Department of Sociology, The University of Manchester

Evelyn Ruppert, CRESC, The Open University

Participants

Tony Bennett, Department of Sociology and CRESC, The Open University

Engin Isin, POLIS and CCIG, The Open University

Jef Huysmans, POLIS, The Open University

John Brennan, CHERI, The Open University

Paul Anand, Department of Economics, The Open University

Francis Dodsworth, CRESC, The Open University

Liz McFall, Department of Sociology, The Open University

Claudia Aradau, POLIS, The Open University

Maureen Meadows, The Open University Business School

Transcriber:

Tamara Shengelia, The Open University

[1] See Thrift, N. (2005) *Knowing Capitalism*. London: Sage Publications Ltd.; and Savage, M., and R. Burrows (2007) 'The Coming Crisis of Empirical Sociology', *Sociology* 41(5): 885-899.

[2] See Appendix 1 for a list of workshop presenters, discussion leaders and participants. The workshop was supported by Theme 3 of CRESC as well as The Pavis Centre for Social and Cultural Research at The Open University.

[3] Surveillance Studies Network (2006) *A Report on the Surveillance Society*, London, Office of the Information Commissioner.

[4] ACORN is a commercial geodemographic tool used to categorise the UK population according to lifestyle groups. It categorises all 1.9 million UK postcodes according to over 125 demographic statistics within England, Scotland, Wales and Northern Ireland, and 287 lifestyle variables. These are used to create a classification map of five categories (e.g., 'urban prosperity'), seventeen groups (e.g. 'educated urbanites') and fifty-six types (e.g. 'multi-ethnic young, converted flats'). See http://www.caci.co.uk/acorn/whatis.asp.

[5] A mash-up is a web page or application that combines data from two or more external sources to create a new representation. Google APIs are Application Programming Interfaces. For example, Google Maps API enables users to embed Google Maps on their own web pages as well as a number of utilities for manipulating maps and adding content to maps.

[6] Dodge, M., and Kitchin, R. (2009) ''Software, Objects, and Home Space', *Environment and Planning A* 41(6), 1344-1365.

[7] Norris, C., and Armstrong, G. (1999) *Maximum Surveillance Society: The Rise of CCTV,* London: Berg Publishers.

[8] MOSAIC a geodemographic segmentation system developed by Experian.

[9] Neo-calculism refers to approaches for managing uncertainty that do not rely on traditional probability calculi but rather create new calculi (e.g., fuzzy logic).

[10] SPSS, Statistical Package for the Social Sciences.

[11] Daston, L., and Galison, P. (2007) *Objectivity*, Cambridge, Mass: Zone Books.

[12] Rose, N. (1999) *Governing the Soul: The Shaping of the Private Self*, 2nd ed., London, Free Association Books.

[13] Watts, D. (2003) *Six Degrees: The Science of a Connected Age*, New York, W. W. Norton & Company.