Session 8: New Data: Sources, Governance, Infrastructure and Analysis

Commentary on Slides

Tim Holt

Slide 1  My purpose is to describe what has been done or is being done in the UK to support research based upon Big Data and to identify the main challenges as we see them. It will be for the audience to judge whether our experience has relevance for Canada.

As background, the UK, like Canada has a long history of making survey data available for secondary analyses by the research community. In addition to anonymized data sources, there are also safe settings where data which is potentially disclosive can be accessed and analysed and the statistical outputs checked to ensure that no breach of confidentiality has occurred. The main organisational difference between Canada and the UK is that in the UK the data are kept in facilities provided by our Economic and Social Research Council (ESRC). Virtually all household surveys conducted for the purpose of producing Official Statistics are deposited in this data service, usually a matter of weeks after collection. In addition all academic research projects supported financially by the ESRC which create new data sets are required to deposit these in the Data Service. In the UK a number of our major surveys such as the British Election studies, the longitudinal household survey ‘Understanding Society’ and the birth cohort studies, are conducted by academic groups supported financially by the ESRC rather than the National Statistical Authority.

Slide 2  This provides an overview of the presentation.

Slides 3-6  These provide a summary of what might be thought of as the main sources of Big Data. These are generally large and often complex data sources that have been collected for purposes other than statistics or scientific research. Slides 3 and 4 cover the various ways in which Government interacts with people and business. Slide 5 covers commercial transactions (e.g., store cards, customer accounts) as well as internet usage such as search behaviour, and social network use. Slide 6 covers other sources such as images and tracking information.

Slide 7  These sources often pose a variety of problems if they are to be used for research purposes:

- Access will be addressed later.
- Provenance: This covers the need to understand the processes that have given rise to the data set; how it is updated and with what frequency; what its coverage properties are and the extent to which it misses cases that one would want to include or included cases (including duplicates) that one would want to exclude. People who work consistently with administrative sources say that understanding these issues is vital to making sensible decisions during the research analysis.
- Permanence: It is a useful scientific principle that data sources have a degree of permanence so that research results can be checked or replicated. Data sets that are continuously updated are in a sense transient and the precise analysis on precisely the same data set may be unachievable. It could be argued that repeating the analysis on whatever is
the latest version of the data set is sufficient but see the following comments on comparability.

- **Comparability**: Much social science research is based upon comparing and contrasting groups or building models of the relationships between variables. One needs to consider whether the data set is suitable for the analytic approach intended.

- **Comparability**: A second important aspect is that social science research may call for repeated analyses at different points in time, in effect to create a time series of results. If the data set is continuously updated then clearly the population of units (e.g., people or businesses) will change and distinguishing between changes in the population structure from changes in the variables of interest may be highly desirable. There is a second important issue here: if this data source is a government data source for example, then changes in policy may have a significant impact on the actual content and coverage of the data source and if these policy changes occur between the successive analytic time points then an apparent change in the time series may be an artefact of the policy change. For example an administrative data base may contain the variable ‘unemployed’ for people who are receiving an unemployment-related benefit. If such a person takes a job then clearly the variable should properly change since it reflects a real change in the circumstances of the individual. However if the Government changes the rules on eligibility for the benefit then this may cause the flag to be changed when there has been no change to the underlying condition of the individual. Such changes can cause time series of results to be incomparable unless there is some way to allow for the policy changes.

- **Legality**: This will depend on the laws that govern the use for research purposes.

- **Ethics**: There may be ethical issues which may call into question whether or how the data may be used for research purposes.

- **Linkage**: Frequently research does not call for the analysis of one data set. Often it is the ability to link different data sets and assemble a range of variables of different categories which is essential to the study. For example combining health, environmental and social variables may provide a much richer source of understanding of an issue. The question is whether these different sources can be linked and, if so, how.

- **Structure**: This is often complex and difficult to work with.

- **Skills and Capacity**: All of the above calls into question whether the social science research community currently have the statistical, data management, and other skills necessary to carry the research through to success. For the UK we will return to this question later.

**Slide 8**
In the UK we have divided our investment into three phases. The first phase, covering administrative data, has been running since late 2014 and first project applications are going through the system established. It is too early to report on any completed projects. The second and third phases (Phase 2: Business and Local Government and Phase 3: Social Media and the Third Sector such as charities etc.) are even more recent and are still being established.

**Slide 9**
The Administrative Data Research Network (ADRN) covers Centres which have been set up in England, Scotland, Wales and Northern Ireland as well as an overarching Research Service at the University of Essex. There are very strong Governance and Access arrangements which have been made as strong as they are in response to access problems which will be covered on a later slide. The Governance arrangements include a Board which operates within our National Statistics
Authority and through which the Board may report directly to Parliament rather than through a Minister. There is a public benefit condition for approved projects and results must be published, including a short, jargon free summary of results to be fed back to the data custodians. Beneath the Board there is an independent Review Committee which considers the ethical framework, the feasibility and public benefit aspects of the proposal. The researcher must be approved and from accredited institutions (e.g. Universities or a small number of independent economic and social research institutions which are eligible for ESRC research funding support).

The actual access arrangements are also elaborate. They involve the researchers, the data custodians and a trusted third party (e.g. in England the Office for National Statistics). The trusted third party is responsible for bringing the various data sets together, managing the linkage processes and making the data accessible to the researchers. It does not analyse the data itself. Even within the trusted third party the ‘identification data’ and the ‘payload data’- the variables of research interest are always kept separate and handled by two quite separate teams so that at no time are the identification variables and the variables of interest held together (see links on last slide for a more detailed explanation).

**Slide 10** For Phase 2: Business and Local Government, Centres have been announced in the Universities of Glasgow, Essex and University College London/ Leeds as detailed on the slide. These Centres are still in the process of being established although each already has a web-site through which more information will steadily emerge. The third Phase is even less well established. A call has gone out for Civil Society Data Partnership projects and a network of people led by David de Roure, Oxford will be advising the ESRC on what arrangements may be needed for research into Social Network Data.

**Slide 11** The term infrastructure is taken to include everything which is not project or data source specific. First there are the various Centres that have been created. Their governance and administrative structures, their mechanisms for publishing and promoting the use of big data are all designed to provide support to would-be researchers. In the early days it is likely that support will be needed through the application process, in gaining access to the data sources and in managing these to produce the required analyses. These Centres will also have a prime role in monitoring how the systems work and in promoting further improvements.

Big Data will throw up methodological issues which are not simply project or data-source specific. These will require research and development and, beyond this, strengthening the skills and capacity of the research community in the use of new methods.

This leads into one of the major infrastructure requirements of the new investments. In the UK the research community will need support in developing new skills and new areas of knowledge to handle projects involving Big Data. Whilst Doctoral training is the prime responsibility of each University, we have a network of Doctoral Training Centres in the UK which provide short courses and other forms of training in specialist topics which are more effectively delivered in this way. In addition some of our research facilities, such as the Centre for Cohort Studies provide specialist training in skills that are particularly needed for cohort analysis. This whole infrastructure will need to take on the additional specialist skill training needed for using Big Data. Our primary targets will be both PhD students and early career researchers who may need skills which they have not yet...
acquired. However, whilst these are our primary targets, all researchers who need additional specialist skills will need opportunities to acquire them.

Finally, infrastructure should be taken to include the whole question of gaining access to the various data sources. In particular, in the case of Government administrative sources, we have spent a number of years building both the case and processes to secure access and also building a collective commitment across data custodians. This is the subject of the next slide.

Slide 12 In the UK access to Government held administrative sources for research purposes has been a significant barrier. Government statements in support of the Big Data initiative have been helpful but, as ever, the devil is in the detail and access to data sources is by no means straightforward. Government is a complex arrangement of semi-independent Departments which are not driven by the same goals or objectives. Indeed the legal arrangements that cover each Government Department and programme vary and the legal barriers to access for researchers will vary across Government. Nor is it just the law, because Departmental lawyers who interpret what a Department may or may not do may take different views on access for research purposes. All of this has created difficulties for researchers seeking to gain access to various data sources. After several years of work we have considerable support from across Government to try to allow access wherever there is not a clear legal impediment. This is one of the reasons that the Governance and access arrangements to administrative data have been so elaborately set out. It is still our view that the whole question could be eased with some overarching legislation that brought some consistency to the question of research access across Government. There is growing support for this but nothing can happen before the upcoming General Election in May and after that it will be for the new Government to decide on its legislative priorities.

Slide 13 As things stand we are hopeful that the investments made will open up access to and use of Big Data for social science research to the benefit of society as a whole. This slide sets out what we see as key aim and challenges.

For reasons given, Access is seen as both a key aim and, given the need for new legislation, a key challenge too. The other aims are clear and, I hope, need no expansion.

The key challenges include exploitation which we see, not simply in terms of high quality research or outputs in the form of publications. Beyond these one would want the research activity to have an impact and here the links between researchers and policy developers and decision makers is vital. We see Capacity building (skills, knowledge and new methods) as a key challenge which will be ongoing for the foreseeable future. Another challenge is to encourage researchers from different disciplines to co-operate since so much potential for improving our understanding is in the gaps between these. The potential to link diverse data sets will, we believe, grow over time and there will be a challenge in finding ways to achieve this. Privacy, ethics and security present an ongoing challenge, not just in ensuring that these are safeguarded but also in building public confidence in our practices and processes and assurance that the data are being used in ways that enhance public policy and improve the well-being of citizens. Finally we have created an infrastructure with help from capital funds but the infrastructure will need continuing recurrent expenditure to maintain it. This becomes yet another competing priority for future budgetary allocations if the infrastructure is to be sustainable.
Slide 14  A small number of links which will be good places to start to find out more about what has been put in place or to follow developments over the coming months and beyond.