ACADEMY OF THE SOCIAL SCIENCES IN AUSTRALIA

AURIN Australian Urban Research Infrastructure Network

RESEARCHER ACCESS TO MICRODATA

CONCLUSIONS FROM 26 MARCH 2015 WORKSHOP

Background

The workshop was hosted by the Academy of Social Sciences (ASSA) and AURIN at Melbourne University on 26 March 2015. It was attended by a mixture of Commonwealth Government and Research Institute data providers, as well as research users from Government and Universities. The list of Attendees is shown at Annex 1. An outline of the Workshop is shown in Annex 2. The focus was on unit record or microdata.

There were four parts to the Workshop.

1. Survey based microdata, including recent and proposed developments.
2. Administrative based and linked microdata, including recent and proposed developments.
3. Emerging Services provided by Research Institutions.
4. Data Provider responses to research concerns.

The Workshop did not provide any great distinction between the issues associated with survey data and administrative data. Consequently, most of the comments below apply to both forms of data. Indeed, one of the areas of increasing interest is the integration of survey and administrative data.

The following summarises the main conclusions reached at the Workshop. There was broad consensus on most points.

Conclusions – Limitations on Researcher Access

Despite the increasing availability of microdata sets, and the variety of means of accessing them, there was universal agreement on the need to increase researcher access but in a way that respects the crucial privacy and confidentiality issues, legislative constraints (but it may be important to make some legislation changes), imperfections in data quality and the costs involved in supporting researcher access.

There are considerable potential benefits to Australia from increased access but only if the main findings of the work of the researchers (based on public data and often using public funding) are to become public. The evidence obtained from this research can be important for policy analysis and other purposes and lead to much improved policy and policy implementation. At present in Australia US data is being used too often by researchers because of the easier access and the findings are of limited value to Australia. Furthermore, improved access avoids researchers trying to conduct their own collections usually of a lower quality and often using government funding. From the data provider's point of view, it increases the value of the data being collected. Direct and indirect users of the data can become advocates of the data collection.

26 Balmain Crescent, Acton ACT 2601 • GPO Box 1956, Canberra ACT 2601 • ABN 59 957 839 703
T +61 2 6249 1788 • F +61 2 6247 4335 • secretariat@assa.edu.au
WWW.ASSA.EDU.AU

The arrangements for the research work can vary and may be (a) work commissioned by policy agencies, (b) work funded by government grants, (c) work based on other sources of funding, or (d) simply 'curiosity' research.

The Workshop showed strong appreciation of the very positive steps undertaken recently by the Australian Bureau of Statistics (ABS) and the Department of Social Services (DSS) to increase access, including the intention to review how legislation is being implemented, the legislation itself, as well as the services provided using NCRIS funding. Other agencies such as AIHW and AIFS have also made important contributions. Researcher access is much better than it was a decade ago. Nevertheless, there were many residual concerns about the current arrangements and the demand has increased particularly for linked data. The most important issues are listed below and discussed in more detail in the following paragraphs.

(1) Lengthy lags from the request for access to a decision on approval.
(2) The cost and complexity of the approval process.
(3) Access to some important data is still not possible.
(4) A lack of consistency in practices across agencies.

Lengthy Lags

The lags can be substantial especially if the data is not already available and ethics and custodian approvals are required. The following are illustrative examples given at the Workshop but feedback suggests significant lags are involved with most researcher requests for access.

The first example is from the Productivity Commission who was seeking access to administrative data from the Department of Human Services (DHS) and State public housing agencies for a research project that examined the effect of housing assistance policies on employment outcomes. Even though this topic was non-controversial it took the best part of half a year to obtain the necessary approvals. Even then, the Commission was only able to receive 1% samples from DHS, so ended up using a population dataset extracted from the Research and Evaluation Database (RED) which contains the Centrelink Payments data but is curated by the Department of Employment. This project, undertaken over 2014, highlighted the challenges, but also the gains that can be made from cooperative approaches.

The second example is data sought from the Population Health Research Network (PHRN). There was one data access request that involved four jurisdictions, five ethics committees, nine data collections, eight data collectors and five linkage units. Not surprisingly, it took some time to obtain all the necessary approvals and the research loses its impact as a consequence.

Even requests to a single custodian can take a long time to approve. For example, the ABS agreed there was a need for much improved performance in this respect and is taking steps to do this.

Access to purpose built survey longitudinal data, such as HILDA, is less problematic with well-developed protocols for researchers to follow.

Cost and Complexity of the Approval Process

This is related to the lags issue. It can be seen from the above examples that multiple approvals are often required. This can be expensive to both the researcher and the data providers. The multiple approvals in the PHRN example would involve considerable cost in organising the various submissions. A challenge for researchers is the decision about whether to apply for grants before they have the approvals for accessing the data, as lack of certainty about access to data may affect the provision of the grant. But going through a long and expensive process to access data, without the grant funding to support this process and the work, is also problematic.

Non-accessibility to some data

Although governments have signed an 'open data' policy which is supported by senior officials in government agencies, there is still a lot of resistance within agencies. This makes the agency approval process more difficult and more complex.

Why are the data sets not made available even though there is an open data policy? Apart from concerns about privacy protection, there are multiple reasons such as competing priorities and lack of skills to do the data management work. Also, political sensitivity should not be underestimated. There is an underlying belief by some staff (often middle level) that they may be placing Ministers at risk if more data is released. It is very different to the culture around the release of macroeconomic statistics where 'open government', efficient policy and efficient markets all dictate the release of a wide range of information around the economy. Some statistical releases will be 'good' news for Ministers, others will be 'bad' news but it is not suggested that only good news be released.

Community Views and Consent

There was discussion about community views on the access to microdata for research purposes. Studies have been taken in Australia and elsewhere. They universally show widespread approval especially if it is to be used for health research. The request of the consent of individuals was discussed but was generally not supported especially if respondent co-operation was protected in other ways. It can be expensive and those that do not provide consent may not be typical of the rest of the population thereby distorting the analysis.

Conclusions – Increasing Access

How do you address these issues which are impeding increased access?

(a) Work to overcome the barriers that limit the amount of data that data custodians will provide.
(b) More trust needs to be placed in researchers whilst clearly specifying the conditions of access. There is a mutual obligation on researchers to behave in a way that is expected of them.
(c) Increase the availability of ready to release data sets including already integrated data sets such as those hosted by PHRN and AURIN.
(d) Acknowledge that different models are appropriate for different users and different data sets. Try to reach consensus on the models and their applicability.

(e) Provide a common law legal environment which could act as a default when legislation does not already exist.
(f) Rethink the role of ethics committees and increased mutual recognition of the findings of the ethics committees.
(g) Find ways of increasing the amount of data that can be released within individual data sets.
(h) Continue to research technological and methodological solutions that increase access.

Each of these is discussed in turn in the following paragraphs.

Trust in Researchers

By and large researchers want to do the right thing and would accept and comply with reasonable conditions and constraints. A risk management rather than a risk avoidance approach can be justified. The starting point should be that researchers can be trusted but how do you avoid deliberate (very unlikely) or accidental (more likely) breaches of conditions. Some steps that might be taken are:

(i) A statement on the respective responsibilities of the researchers and the data providers. One important requirement is for researchers to provide applications for data access that provide data custodians with confidence that the data will not be misused and the research has a net benefit. Templates might be developed to assist this.
(ii) The development of standard protocols for the release of micro data sets, including the licensing arrangements. Individual releases could be based on these protocols as could undertakings to be signed by the researcher. Ideally, these undertakings should be legally enforceable if there are breaches. Model documents could be prepared.
(iii) Guidelines on how breaches might be managed. These might vary depending on the seriousness of the breach. For example, legal action should be taken where the breach was deliberate and significant. In other cases, the actions might vary from banning future access to the researcher and their institution and a warning.

There was some discussion of a 'Trusted User Model' where the Probability (Disclosure) = Probability (Disclosure|Attack)*Pr(Attack). For access by public servants, there is the Crimes Act, other legislation and the Code of Conduct to ensure public servants are doing the right thing ie make P(A) small. Under the model P(D) = P(D|A)P(A), P(D) will be small if we make P(A) small; and if this can be made to happen for researchers, custodians don't need to do as much on P(D|A), which is what they have been concentrating on hitherto (eg removing matching risks) often limiting the usefulness of data. For academic researchers, the implementation of the three measures listed above should make P(A) small but it requires the co-operation of the research institutions.

Increased Availability of Ready to Release Data Sets

Lags are frustrating and these will inevitably be longer when the requested data sets have to be created. It is much easier if the data sets are already available that can be accessed by any researcher who meets a predetermined set of criteria. This is already happening to some extent in organisations like the ABS, AIFS and AIHW although it is conceded that administrative processes might be improved. DSS has chosen to amalgamate the four longitudinal studies it manages under the banner of the 'National Centre of Longitudinal Data'. It is proposed that within this entity,

custodians of the national longitudinal collections will be able to share a range of data management techniques, including those which have the effect of improving data access to researchers'.

More generally, for administrative data, if departments had curated databases of the data that were well documented, specific datasets would be easy to extract. RED is an example of this approach. Having been extracted, datasets should be hosted for reuse wherever possible. This is particularly important for where data from different sources has been linked, as such linkage involves extra effort to create and check the dataset.

AURIN is another example of ready to use data sets and works well from a researcher access perspective. (Note that AURIN doesn't actually provide access to the microdata itself but enables analysis based on the microdata.) The PHRN is not quite the same. Although the data sets are available, many approvals are required before data can be released and this causes inevitable delays.

Different Models for Different Users

The UK has a framework based on the 'five safes' and the nature of the data sets (discussed below). Access arrangements are determined using this framework. This could be adapted for the Australian situation. It already has been by New Zealand using Statistics New Zealand as the co-ordinating authority for the Integrated Data Infrastructure.

Common Law

Some organisations like the ABS and AIHW have their own legislation, although it is being reviewed in the case of the ABS. Most other organisations do not have relevant legislation. There would be benefits in having a common law approach based on best practice. This could underpin the release practices for these organisations. It should also enable greater consistency of practice across organisations.

Ethics Committees

Ethics Committees play an important role but the requirements are onerous particularly for projects that involve multiple jurisdictions. This increases costs and can cause significant delays in the approval process. There were two main concerns. First, ethics committees were sometimes used when ethical clearance may not be necessary given the nature of the request. Second, several ethics committee approvals were sometimes necessary on the same request. Surely, there can be some rationalisation. If an appropriate ethics committee is established, couldn't all jurisdictions act on the advice provided by that Ethics Committee?

Increasing Available Data within a data set

Researchers want to have as many data items as possible available especially if their analysis is exploratory in nature (ie curiosity research). Custodians want to limit the data available and ask researchers to be quite specific about their requirements. One solution may to provide a sample of the full data set to enable sufficient analysis to be undertaken to allow researchers to be more specific about their requirements. This would also circumvent the significant resources and costs on

behalf of researchers, ethics committees and data custodians associated with the provision of additional data that was not specified in the initial approvals.

Technological and Methodological Solutions

These are important and are one of the reasons for increased access over the last decade. There may be further opportunities especially around the linkage of data sets. The ABS is well placed to provide leadership here through its technical and methodological strengths and networks with those doing similar work in other countries.

Linked Data

The Workshop supported the establishment of the Integrating Authorities, namely the ABS, AIHW and AIFS. They are still relatively new so their procedures would not yet be mature. A recent review has recently been completed. It showed that the focus to date had been on protecting the integrity of the data rather than helping researchers to integrate data. Whilst the former is important, there is a need to better balance these two streams of work.

Conceptual Framework for Addressing Researcher Access

Several countries, including the UK and New Zealand , used a conceptual framework based on the 'five safes'. These are:

- Safe people – researchers can be trusted to use the data appropriately and follow procedures
- Safe projects – the project has a statistical purpose and is in the public interest
- Safe settings – security arrangements prevent unauthorised access to the data
- Safe data – the data itself inherently limits the risks of disclosure
- Safe output – the statistical results produced do not contain any results that disclose details about individuals.

This framework could be adopted by Australia. In its presentation, the ABS said it was only currently looking at the first, fourth and fifth 'safes'. Perhaps others should be looked at as well when making decisions on researcher access.

The UK goes further and considers the nature of the data sets when making decisions. For example, is it a sensitive or non-sensitive data set?

Where to from here?

The desired future situation was that there would be improved researcher access to integrated data sets in a way that respects the confidentiality and privacy of the subjects of the data. Unless otherwise agreed, the outputs of this research should be in the public domain for the benefit of those involved in policy development and monitoring, planning of services, and so forth.

It was agreed the main focus should be at the Commonwealth level. Their data sets were more extensive and generally were of greater interest. Procedures at the State level could be adapted from the Commonwealth arrangements. Therefore, it is important that the States have an opportunity to influence the Commonwealth arrangements.

An exception is the health sector where the States have already established good processes. Regardless, harmonisation is desirable with jurisdictions learning from each other's experiences.

The proposed changes will require some modifications to the current arrangements of data custodians even though there have been important steps in the right direction. Importantly, it places an obligation on researchers to behave in accordance with any agreements reached with data providers. If there are breaches there will be consequences for them.

The Workshop agreed that it will not happen without leadership. Where will that leadership come from? The Workshop did not discuss this specifically but the ABS is well placed. It has the technical knowhow and legislated responsibility for the National Statistical Service which covers the States as well as the Commonwealth. In the UK and New Zealand, the leadership has come from their National Statistical Services.

The DSS also has an important role as custodian of the many administrative based data sets and longitudinal surveys. They should also be part of the leadership especially given the changes in policies they have recently been trying to implement. A starting point might be to establish a Working Group to address the issues associated with increased access. The ABS and DSS would be key players.

Infrastructure funding through NCRIS has been important but it is suggested future funding priorities should be based on research projects using integrated data sets.

In summary, the keywords for improvement are trust, leadership and culture.

Prepared by Dennis Trewin
May 2015

Annex 1

WORKSHOP ON RESEARCHER ACCESS TO MICRODATA

Participants

Dennis Trewin FASSA – ASSA Convener

Bob Stimson FASSA – AURIN Co-convener

Kevin Fox FASSA – University of New South Wales

Mark Wooden FASSA – University of Melbourne

Richard Sinnott – AURIN

Andrew Dingjan – AURIN

Alan Hayes – Australian Institute of Family Studies

Ben Edwards – Australian Institute of Family Studies

Siu-Ming Tam – Australian Bureau of Statistics

Sean Innis – Department of Social Services

David Dennis – Department of Social Services

Merran Smith – Population Health Research Network

Jenny Gordon – Productivity Commission

Diane Watson – National Health Performance Authority

Sallie Pearson – Sydney University

Shane McWhinney – Victorian Government

Steve Zubrick – University of Western Australia

Susie Kluth – Treasury

Steven McEachern – Australian Data Archive

Rob Tanton – National Centre for Social and Economic Modelling

Murray Radcliffe – Academy of the Social Sciences in Australia

Background

In its feature article in their 2013 Annual Report, the Productivity Commission said,

"Academics, researchers, data custodian agencies, consumers and some Ministers are eager to harness the evidentiary power of administrative data, but this enthusiasm generally is not matched

by policy departments. Despite tentative steps, overall progress has been inadequate. Leadership and commitment is required to promote the evidence-based policies needed to meet Australia's economic and social objectives within budget constraints that will become more acute given the demographic outlook."

This ASSA/AURIN Workshop is intended to consider and discuss the steps that should be taken to provide improved researcher access to microdata, including linked data, in the direction proposed by the Productivity Commission. It should also consider the authorising environment and protocols to enable improved access.

It will cover publically funded microdata derived from surveys and microdata derived from administrative systems.

The rapid innovations occurring in IT capabilities for securitised protocols for accessing unit record data, linking different datasets, and facilitating on-line interrogation of those data without violating confidentiality requirements now provides potential mechanisms to achieve these objectives. The workshop will, as a case study, explore mechanisms to link datasets and to integrate unit record data with spatial objective data to facilitate innovation in research and policy analysis.

Objective

There have been a number of recent technology developments which have increased researcher access or have the potential to increase researcher access in specific sectors including:

- Australian Urban Research Infrastructure Network (AURIN); and
- Population Health Research Network (PHRN).

Furthermore, as an alternative, many countries are moving to a providing a 'safe environment' to provide access to approved researchers. This recognises that it is extremely difficult to fully confidentialise microdata without destroying its utility for research.

The objective of the workshop is to try to reach a meeting of the minds between the research community and the microdata providers especially around appropriate protocols for access to microdata sets (ie the authorising environment). Linked microdata will be a topic of specific interest.

The scope of the Workshop would include both Commonwealth and State Government providers of microdata.

Participants

The participants would be senior microdata users especially those who are Fellows of the Academy, research users within Government, representatives from Commonwealth and State data providers (ABS, AIHW, DSS, Health) and the Productivity Commission, plus the involvement of AURIN, PHRN and NATSEM. In total, we are looking at 20-25 participants.

Venue/Date

The Workshop is be held at Melbourne University hosted by AURIN on 26 March 2015.

Woodward Conference Centre
10th Floor, 185 Pelham Street, Carlton.

http://maps.unimelb.edu.au/parkville/building/106/woodward_conference_centre

Modalities

There would be four sessions during the day. This would include a concluding session to summarise the main findings. Each Session would be opened by nominated speakers followed by open discussion. The Chair would endeavour to summarise the main points emerging from the discussion. Dennis Trewin of ASSA and ex-ABS is the proposed Chair.

An Issues paper would be prepared by ASSA and circulated to participants beforehand. It will summarise recent developments in the provision of access to microdata in Australia and elsewhere. The Productivity Commission feature article, *Using Administrative Data to Achieve Better Policy Outcomes* would also be circulated.

A draft agenda might be as follows. Session 3 might continue after the afternoon tea break.

Session 1 –Introduction

Introduction of the current arrangements and plans for improved researcher access to microdata

Followed by Open Discussion

Session 2 – Developments in accessing Administrative Data including linked Administrative Data

The session would be opened by a presentation by the Productivity Commission on the importance of utilising administrative data, including linked data, for statistical and analytical purposes. The Department of Social Services would talk about the latest developments in accessing data for which they are custodians. This would be followed by open discussion.

Session 3 – Issues from a Researcher Perspective

What are the key issues from a researcher perspective? What would they like changed? This session will include brief presentations from AURIN and PHRN outlining their approaches for achieving securitised protocols for accessing unit record data, linking different datasets, and facilitating on-line interrogation of those data without re violating confidentiality requirements and how to integrate unit record data and spatial objective data and the interrogation of those data using advanced modelling approaches. This would be followed by Open Discussion. Participants from the microdata providers would be given the first opportunity to comment on the remarks made by the speakers.

Session 3 (continued) – What changes might data providers consider to current arrangements?

This would be a more formal response by the data providers on what changes they might consider and associated conditions. Each of the main data providers would be given an opportunity to speak. This would be followed by open discussion.

Session 4 - Concluding Session

The focus of this session would be to summarise the main outcome of the day's proceedings. A person will be selected to lead this discussion by summarising what they think were the main points. A facilitator for this discussion would be identified.

The Steering Committee would document the agreements reached at the Workshop. ASSA would provide the rapporteur and prepare the first draft of the documented agreements. Dennis Trewin, as Chair, would be the lead author.

Organisational Arrangements

Dennis Trewin will chair the Steering Committee. Other members of the Committee are Peter Harper (ABS), Bob Stimson (ASSA and AURIN), Kevin Fox (ASSA and University of New South Wales) and Michael Bittman (ASSA and University of New England). Murray Radcliffe of ASSA will provide the secretariat services.