Submission to the Productivity Commission Inquiry into Data Availability and Use

Never Stand Still

Friday 29 July 2016

UNSW Contact: **Associate Professor Grainne Moran, Pro-Vice-Chancellor (Research Infrastructure)**

UNSW Australia welcomes the opportunity to provide a submission to the Productivity Commission on data availability and use.

UNSW agrees with the view expressed in the Productivity Commission's Issue Paper that the non-rivalrous nature of data makes them an especially valuable resource.

Data are a new commodity unlike any other in that they are not only not consumed by use, but they can become more valuable the more they are used and shared. As such, data collection and management to ensure quality, durability, security, interoperability and accessibility requires innovative thinking. Data cannot be considered in isolation from the infrastructure, including computational infrastructure, required to deliver the benefits addressed in this paper.

UNSW recommends that:

- *prior to any new data generation, consideration be given to the mechanisms and infrastructure required for the management, curation and re-use of these data;*
- *linking of Commonwealth and State-based data collections be improved;*
- *special consideration be given to the management of "non-traditional data", such as simulation data;*
- *data be more efficiently used for evidence-based decision making, especially in the healthcare and environmental sectors;*
- *government and private sector data owners should fund the infrastructure required for data sharing; and*
- *models to gauge potential risks associated with data usage and mitigation of those risks should be implemented, for example a Principled Proportionate Governance Model.*
- *early engagement with the research community is essential and mutually beneficial, both in planning and in implementing processes for making data available and accessible*

UNSW would be pleased to provide any clarification or further information relating to this submission.

Yours sincerely,

Associate Professor Grainne Moran

Public sector data

What constitutes high-value public sector datasets – characteristics and potential benefits?

High-value public sector datasets are those that present potential for use and re-use for public benefit. In many areas (e.g. health, transport, housing, environment, compliance registries), these include datasets that are collected routinely by governments, and third party providers.

Many of these datasets have high levels of completeness and accuracy, having rigorously applied standardised data definitions. Other potentially valuable datasets may be less standardised and require significant 'cleaning' before being valuable as a resource.

Potential benefits of these data for research and policymaking include[1]:

- **Population reach:** many public sector data have whole-of-population coverage, and can be used to study rare outcomes (e.g. adverse events) and population subgroups.
- **Longitudinal:** when linked internally or across datasets, these data have a longitudinal structure that supports long term monitoring and tracking.
- **Avoid nonresponse, attrition and reporting bias:** routine data collections are not subject to the challenges of surveys such as high and rising rates of socially and health-patterned nonresponse and attrition, as well as social desirability, reporting and recall biases.
- **Cost-effective:** the use of routine data for research and evaluation increases return on investment for the public resources expended in collecting them, and studies over many decades can be undertaken time-efficiently and cost-efficiently compared with prospective data collection. However, additional upfront costs may be incurred to ensure future re-use is enabled.
- **Real world:** routine data often present the only way to evaluate outcomes in population groups and to evaluate the impacts of policies or services that have been rolled out in a nonrandomised manner.
- **Privacy protection:** proven methods to protect the privacy of individuals and organisations have been successfully implemented in Australia – this experience needs to inform data governance in new areas of data collection and sharing.

We also note the existence of simulation or derived data (as opposed to collected or experimentally generated data). This type of data also has value, but has very different needs in terms of how it should best be saved and made accessible for future re-use.

Collection and release of public sector data:

- Should collection, sharing and release of public sector data be standardised?
- What criteria and decision-making tools should be used to decide which datasets to make publicly available?

Public availability of taxpayer-funded data collections should be considered immensely valuable for Australia's future. The need to make data available and re-useable should drive the data collection or generation process. The development of standards is essential, but can be time-consuming and expensive where these do not exist. Consideration should be given to adopting standards compatible with appropriate international standards, where these exist and can add value by opening up data sharing even more broadly.

[1] Jorm L. Routinely collected data as a strategic resource for research: priorities for methods and workforce. *Public Health Res Pract*. 2015;25(4):e2541540.

How and where the data is to be hosted, and the availability of compute infrastructure to process 'big data' is also a factor.

Thirdly, the likely levels of data sharing need to be considered. Will data be made available to researchers under tightly controlled conditions? More freely, including to business and industry? Open (at least for aggregated or summary data) to the general public?

These considerations should be factored in to the design of any new data generation or collection process. It will be much more cost effective to build in data management measures upfront where possible. In many instances engagement with the relevant research sector(s) in designing such processes, both for new and existing datasets, would be mutually beneficial and is strongly recommended.

Which specific government initiatives have been effective in improving data access and use?

In the health sector, the Population Health Research Network has contributed to building further infrastructure for linkage of population health data including a data linkage (integration) facility within the Australian Institute of Health and Welfare (AIHW) that is accredited to link Commonwealth data, and a secure remote access facility for analysis of linked data (SURE, see https://www.saxinstitute.org.au/our-work/sure/). However, real improvements in access to and use of linked Commonwealth health data have yet to be realised, with researchers typically waiting for around two years to receive these data for their grant-funded research.

Another example for increased data accessibility and use is Australia's Integrated Marine Observing System (IMOS) funded under the National Collaborative Research Infrastructure Scheme. IMOS has been instrumental in collating and disseminating oceanographic research data and meta data. The IMOS online data portal has now been merged with the Australian Ocean Data Network (aodn.org.au), which is the "one stop shop" for research quality oceanographic data collected by universities and publically funded research agencies.

Yet another example is the increased data and analytics capability provided to the urban research community through the Australian Urban Research Infrastructure Network which provides free access to academics, and governments to access the AURIN portal. This portal makes available over 1,500 data products and 100 spatial-statistical and visualisation tools. This is also a "one stop shop" e-infrastructure supporting an evidenced based approach to city planning.

Which datasets should be linked across public sector agencies and what are the impediments to linking datasets?

In general terms, the impediments include lack of metadata, data ontologies cost, understanding of data archiving practices, standardisation, quality assurance and quality control issues. Linking of datasets requires significant IT resources - and human capital to ensure meta data and data formats are interoperable.

Taking health data as the example, the foremost impediment is the challenge of bringing together Commonwealth and State-based data collections to generate the most up-to-date evidence about the full-spectrum of health care. For example, it is not possible to quantify the outcomes derived from the Commonwealth's investment in medical and pharmaceutical interventions without linking data across Commonwealth and State boundaries. At a minimum, we require timely linkage of Commonwealth health data (MBS, PBS) with national hospital inpatient and mortality data. This is entirely feasible now. Other high priority datasets for

inclusion in a national linked data resource include aged care data, emergency department data, perinatal data and disease and device registries.

Furthermore, the ethics and governance of data linkage activities need to be simplified to allow agile responses to emerging and contemporary research and policy questions. The current system is underpinned by 'one-off' linkage and integration of data for specific purposes and slow, confusing and cumbersome processes to obtain data custodian and ethical approvals.

Linkage of public and private sector datasets poses additional challenges but these also need to be addressed for other important areas such as environmental, water, transport, housing and economic and social data.

What are the benefits of greater availability of data to research and what benefits could researchers give back to government?

Whilst research is benefitting from greater availability of public data, researchers in turn enable greater availability and use of data. For example, data collected by the Integrated Marine Observing System (IMOS) and AURIN infrastructures are provided by diligent technical support yet still require significant cleaning of the data post-collection. Research use itself represents a valuable addition to data collection infrastructure if the outcomes are fed back into the data repository in a coordinated manner.

In addition, use of data at Universities for teaching also has the potential to improve availability and use. The use of IMOS data in teaching resulted in investigations on how to apply the Ocean Data View software to the IMOS data, driving better methods of data display. The use of AURIN data in teaching is resulting in an increased understanding of how data can be used to support policy and decision-making in dealing with issues facing our cities including congestion, housing affordability, population health and an ageing population.

Australian governments invest more than $100 billion annually on healthcare, yet we have a relatively limited understanding of Australia's return on this investment. Even when medical treatments have undergone extensive pre-market evaluation in randomised controlled trials, they are most often tested over relatively limited time frames. Real-world studies underpinned by health data linkage are the only means of establishing the population benefit and risk profiles of these interventions long-term.

Additionally, research using these data will help us to understand the determinants of disease risk, target therapies to those who will benefit most, compare the effectiveness of alternative interventions, and model the health and economic impacts of interventions and policies. It is estimated that effective use of big data could also deliver reductions to national health care expenditure of around 8%, which would translate to more than $11 billion annually in Australia.[2]

Private sector data

What constitutes high-value private sector datasets – characteristics and potential benefits?

Access to private sector data:

- Are there any legislative or other impediments restricting access to private sector data?
- Should voluntary data sharing arrangements (e.g. between businesses, consumers and third party intermediaries) be mandated?

[2] Groves P, Kayyali B, Knott D, Van Kuiken S. *The 'big data' revolution in health care*. New York: McKinsey and Company Center for US Health System Reform, 2013.

- What role can governments play in promoting the accessibility of private sector data?

High-value private sector data present similar potential benefits to high-value public-sector data. To be considered high-value datasets must also be complete and accurate, including with free-text fields capable of being mined to extract the relevant information.

The ability to link public and private sector data is essential and government has a role to play in promoting appropriate adoption of standards and protocols to enable this to happen.

An example is oceanographic data, which the private sector (oil and gas industry) collects extensively in specific regions of interest, for example the Northwest Shelf, Great Australian Bight, Gladstone Harbour, etc. This data might not hold a commercial advantage to the owner after a period of time, but is of great value in improving oceanic model validation, assimilation and calibration. Access to such privately collected data would make a significant contribution to Australia's ability to model and predict coastal ocean circulation, which is currently limited by data availability.

Management of other forms of data

UNSW recommends that considerations be given to managing data that does not fit within the classical concept of public or private sector data being created and then saved. For example, data generated *via* simulation is different from data collected *via* a field survey or questionnaire in that it can be updated and replaced and as such evolves over time.

Vast arrays of simulation data exist in certain research areas, for example climate system science. Archival and saving of these data can be deeply wasteful, yet is a requirement of the Australian Code for the Responsible Conduct of Research[3].

UNSW recommends that rules around holding data for periods of time for governance, compliance etc. be amended to give special consideration to data that can be recreated, such as simulation data. For such data, it might be more appropriate to archive the means (e.g. the code and source data) to recreate the data rather than the dataset itself.

Managing the costs

How should the costs associated with making public sector data more widely available be funded?

Infrastructure for sharing public sector data should be funded by governments, consistent with their current and emerging open data policies that recognise the value of sharing these data. For example, the Australian Government's Public Data Policy Statement[4] commits it to optimising the use and reuse of public data, releasing non-sensitive data as open by default and collaborating with researchers to extend the value of data. Similarly, the NSW Government Open Data Policy[5] requires agencies to manage data as a strategic asset to be open by default, discoverable and usable and free where appropriate.

[3] National Health and Medical Research Council, Australian Research Council and Universities Australia 2007, *Australian Code for the Responsible Conduct of Research*. Commonwealth of Australia, Canberra. Available at: https://www.nhmrc.gov.au/guidelines-publications/r39.
[4] Australian Government Public Data Policy Statement. Available at: https://www.dpmc.gov.au/sites/default/files/publications/aust_govt_public_data_policy_statement_1.pdf
[5] NSW Government Open Data Policy. Available at: https://www.finance.nsw.gov.au/ict/sites/default/files/resources/NSW_Government_Open_Data_Policy_2016.pdf

Making data freely available to public sector researchers will enable benefits to be returned to government by way of new insights and findings.

Similarly, infrastructure for sharing private sector data should be funded by the data owners, recognising the potential value to their organisations of re-use of these data. Where private sector data owners are funded or regulated by the public sector, requirements should be put in place for the provision of relevant data so it can be shared through the same infrastructure as public sector data. For example, as part of their licensing requirements, private hospitals are currently required to provide monthly data to the relevant State Health Department, which is then included in the admitted patient data submitted to the AIHW for inclusion in the national inpatient data collection.

Marginal costs relating to specific research projects should be funded through research grants or the funders of commissioned or sponsored research. However, pricing models need to recognise the 'hand to mouth' nature of grant-funded research. Costs of data linkage, data supply and data access are currently unpredictable, vary widely between different agencies that are providing similar services, and may not be covered by the funds initially budgeted in research grants (e.g. costs for using the SURE laboratory have risen steeply since its establishment).

Data governance and procurement proceedings should rigorously ensure data acquired using government funding is re-useable for research purposes and where possible is made available *via* Government open data portals and other such initiatives.

Is availability of skilled labour (e.g. data scientists) an issue and is there a role for government in improving the skills base?

There is an acute shortage of data scientists. Education and training in this area needs to be developed and supported in order to ensure a skills base not only in the data management and technical fields, but also to ensure that the government, the economy and society more broadly is able to take full advantage of the opportunities provided by the current and future 'data ecosystem'.

There is an ongoing research element required to ensure that Australia continues to be competitive internationally.

How should the risks and consequences of data breaches be assessed and managed?

Adoption of a Principled Proportionate Governance Model (PPGM) based on clear guiding principles would help to overcome key impediments to using health records for research. Such models gauge potential risks associated with data uses and mitigations to those risks, including the potential public interest that is served by enabling research. They require a clear articulation of roles and responsibilities at all levels of decision-making and effective training for researchers and data custodians. Examples of PPGMs include those developed by the Scottish Health Informatics Program[6] and PopData BC[7].

Approved researcher accreditation could be considered as a risk mitigation strategy for research use of potentially sensitive microdata, such as linked health records. This would be preferable to crude 'one size fits all' approaches to disclosure control such as restricting the records able to be

[6] Sethi N, Laurie GT. Delivering proportionate governance in the era of eHealth. Making linkage and privacy work together. *Med Law Int*. 2013 Jun; 13(2-3): 168–204.

[7] McGrail KM, Gutteridge K, Meagher NL. Building on Principles: The Case for Comprehensive, Proportionate Governance of Data Access. In Gkoulalas-Divanis A, Loukides G (eds.) *Medical Data Privacy Handbook*. Springer: Switzerland 2015, pp 737-764.

accessed (e.g. to a 10% sample) or recoding to limit the level of detail available (such as aggregating geographic areas), which reduce and sometimes negate the value of the data for research and discovery. Examples of approved researcher accreditation processes include those operated by the UK Office of National Statistics (https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/approvedresearcheraccreditation) and UK Economic and Social Research Council (https://www.ukdataservice.ac.uk/get-data/how-to-access/accesssecurelab).

Data infrastructure

What are Australia's infrastructure needs to improve access and usability of public and private sector data?

The ability to store and process digital data is an integral requirement for data availability, use and re-use. IT infrastructure needs to be provided to host growing datasets, ideally in an environment that enables not only storage, but also machine to machine access and cloud computing.

As outlined above, simple collection of data may not be sufficient to enable future use. UNSW recommends that infrastructure should not only be considered as the hardware for data collection and storage, but should also include human capital in the form of data scientists or researchers evaluating and implementing the standards and processing required to enable the usefulness of the data in the future.

A national strategic approach to investment in technical infrastructure is required.

From a health perspective, creation of a new national enduring and accessible resource of linked health data ('Health Big Data virtual laboratory') is the top priority, along with the infrastructure (including human capital) to support large-scale partnership research programs focused on using big data to tackle major challenges for Australia. Existing research groups rely on unpredictable competitive grant funding and short-term contract research, and therefore struggle to achieve and maintain critical mass. Large-scale, longer-term partnership research programs with government and industry would achieve better and outcomes.

The Centre for Big Data Research in Health at UNSW (https://cbdrh.med.unsw.edu.au/) is a world-first research centre that aims to maximise the use of all possible sources of health big data in order to enhance the health and well being of Australians and the global community. All of its projects use public sector data, most use linked data and some involve linkage of public sector data with research data.

The City Futures Research Centre at UNSW (https://cityfutures.be.unsw.edu.au/) is undertaking world-leading scholarly urban research. City Futures is dealing with many of the issues associated with the grand challenge of an increasingly urbanized world. It is doing so taking a data-driven approach, which utilises public sector data to create urban big data platforms, such as CityViz (https://cityfutures.be.unsw.edu.au/cityviz/), and open data dashboards, such as the Sydney Dashboard (http://citydashboard.be.unsw.edu.au/).