

July 2016

Better access to data will lead to better decisions

Submission to the Productivity Commission's Inquiry into Data Availability and Use

By Stephen Duckett, Peter Goss and Marion Terrill

Grattan Institute Support

Founding Members



Australian Government



Program Support

Higher Education Program



THE MYER
FOUNDATION

Affiliate Partners

Google
Origin Foundation
Medibank Private

Senior Affiliates

EY
PwC
The Scanlon Foundation
Wesfarmers

Affiliates

Ashurst
Corrs
Deloitte
GE ANZ
Urbis
Westpac

This report was written by Stephen Duckett, Pete Goss and Marion Terrill, Directors of Grattan Institute's Health, School Education and Transport Programs respectively.

The opinions in this report are those of the authors and do not necessarily represent the views of Grattan Institute's founding members, affiliates, individual board members reference group members or reviewers. Any remaining errors or omissions are the responsibility of the authors.

Grattan Institute is an independent think-tank focused on Australian public policy. Our work is independent, practical and rigorous. We aim to improve policy outcomes by engaging with both decision-makers and the community.

For further information on the Institute's programs, or to join our mailing list, please go to: <http://www.grattan.edu.au/>

This report may be cited as:
Duckett, S., P. Goss & M. Terrill, 2016, *Better access to data will lead to better decisions*, Grattan Institute

All material published or otherwise created by Grattan Institute is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License

Overview

- Data sets held by governments and businesses can now be used for multiple purposes, not just the initial transactional record. They have the potential to transform policy development and evaluation and economic analysis.
- Payment systems for government benefits and services lead to the accumulation of large data sets, containing millions of data items. The primary purpose of the collection of the data is for payment and accountability. But an important secondary purpose is to use the data to understand patterns and trends. Secondary analysis of data sets collected for 'routine' or 'administrative' purposes is now a well-accepted type of research.
- Secondary analysis of government datasets has a number of additional benefits:
 - Because the analysis is conducted on government data, the research is almost always policy-relevant. It can shed insight into patterns of spending or service use that would otherwise not come to the attention of policy makers; and
 - Secondary analysis leverages investments that have already been made in data collection and so is generally a less expensive form of research. Secondary analysis of Australian data sets has addressed a range of issues including quality and efficiency of, and access to, health care.
- Government data holdings, derived from claims processing, should be seen as an important public resource to assist in policy-relevant research which will benefit the Australian community. Failure to harness fully the potential of these data sets represents a significant lost opportunity both for policy development and research.
- Privacy risks can be mitigated by controlling data outputs and/or controlling the data released. The New Zealand approach to controlling data outputs should be explored as a potential additional data release strategy in Australia. Data custodians also limit release to approved data users. This is appropriate and should be continued.
- Data release should be expanded and facilitated by
 - Releasing metadata
 - Developing streamlined and standardised release approval processes; and
 - Developing common-use data sets.

Table of contents

Overview	1
1 Data for decision making	3
2 Controlling the privacy risk	6
3 Moving forward on data release	9
References	12

1 Data for decision making

Businesses, not-for-profit organisations and governments are increasingly recognising the power of data to improve their decision making and service to their customers or clients.¹ Data sets held by governments and businesses can now be used for multiple purposes, not just the initial transactional record. They have the potential to transform policy development and evaluation and economic analysis.²

The potential uses of data are now only limited by our imaginations:

Publishing, linking and sharing data can create opportunities that neither government nor business can currently envisage.³

Daily additions to world-wide data holdings is now measured in the quintillions.⁴ Increased computer power makes linking of data sets (e.g. social security and tax records) more feasible and increases the utility of data holdings.

In contrast to what has been the past history, where government data bowerbirds have collected but obstructed release of data, the Australian government is now committed

to optimise the use and reuse of public data; to release non sensitive data as open by default; and to collaborate with the private and research sectors to extend the value of public data for the benefit of the Australian public.⁵

1.1 Data about government services

Payment systems for government benefits and services lead to the accumulation of large data sets, containing millions of data items. The primary purpose of the collection of the data is for payment and accountability. But an important secondary purpose is to use the data to understand patterns and trends. Secondary analysis of data sets collected for 'routine' or 'administrative' purposes is now a well-accepted type of research.⁶

Despite this treasure trove of data, researchers typically rely on survey data to gain insight. Yet surveys are a poor source of information. The reasons for this are well-documented but worth repeating. Non-response rates are high and increasing. Longitudinal surveys are rare – and suffer from high attrition. Although the largest surveys conducted in Australia are sufficient for calculating basic statistics with confidence, they are well below the amount required for other important research questions. The closure of a major factory would likely not even be detectable in any survey currently available to Australian researchers.

¹ McAfee and Brynjolfsson (2012)

² Einav and Levin (2013)

³ Turnbull (2015).

⁴ See Kim, *et al.* (2014); a quintillion has 18 zeros.

⁵ Turnbull (2015)

⁶ This and other sections of this Submission draw on a previous paper on data access, the first draft of which was written by Dr Duckett. See Academy of the Social Sciences in Australia (2013)

In an era where some Australians are feeling increasingly alienated and isolated from policy decision-making, the ability to conduct detailed public policy research is urgent.

Although survey data is typically less sensitive than administrative data, the barriers to access are nonetheless onerous. Linking and matching is often forbidden. And confidentiality requirements often make good-practice research methods (such as obtaining review from external colleagues) practically impossible.

Secondary analysis of government datasets has a number of additional benefits:

- Because the analysis is conducted on government data, the research is almost always policy-relevant. It can shed insight into patterns of spending or service use that would otherwise not come to the attention of policy makers; and
- Secondary analysis leverages investments that have already been made in data collection and so is generally a less expensive form of research. Secondary analysis of Australian data sets has addressed a range of issues including quality and efficiency of, and access to, health care.

Secondary analysis is increasingly common and procedures have been established by many data agencies to ensure confidentiality of data to ensure that individual privacy is not infringed by allowing further (secondary) use of the routinely collected data. Data released is de-identified (to protect privacy) and small cells are also suppressed to prevent any possible identification. Data release agreements generally require researchers to certify that

they will not attempt to identify individuals and/or specific services.

Government data holdings, derived from claims processing, should be seen as an important public resource to assist in policy-relevant research which will benefit the Australian community. Failure to harness fully the potential of these data sets represents a significant lost opportunity both for policy development and research.

Grattan has been fortunate to use such resources to enhance its research.

A report released in March 2016, made use of linked student data from Victoria recorded across four NAPLAN test years (school years 3, 5, 7, and 9, recorded from 2009 to 2015).⁷ Having student records that were linked enabled the analysis to focus on student progress rather than simply outcomes. Unfortunately, the same linked NAPLAN data were not available for other states – in some cases student identifiers were not properly recorded, and in other cases the education departments were not allowed to share the data with us. Having data linked nationally would have enabled a comparison across states, a valuable input into analysis of the effectiveness of different policies.

A report released in August 2015, used routinely collected data to analyse the use of ineffective treatments in Australian hospitals. International research shows that ineffective and inappropriate care is a major source of waste in hospitals.⁸ Further, it poses a

⁷ Goss, *et al.* (2016)

⁸ Duckett, *et al.* (2015)

risk of harm to patients who could otherwise have avoided the stress, cost, inconvenience and risk of a hospital stay altogether.

Australia currently has no system for identifying and addressing provision of inappropriate care in hospitals. Duckett *et al.* (2015) developed a model for doing so; however, because our data sets were not linked, and what is considered 'inappropriate' depends on the patients' history of treatment, we could only measure use of about one eighth of the treatments listed as ineffective in the United Kingdom. If we were able to access this data, we would be able to observe four times as much ineffective care.

1.2 Data about government assets

Australian governments spend billions every year on public infrastructure, but the data that they collect and generate in the process is not gathered in a standardised way and mostly not made publicly available.

There are two kinds of data that would support better infrastructure investment by governments and make a substantial difference to public accountability and transparency.

One kind of data is standardised information on key characteristics of major infrastructure projects. Key characteristics include the costs, benefits, timing, adherence to appraisal processes, funding sources, financing, risk allocation, contractual costs, procurement approach and scope changes to major public infrastructure projects through to the complete conclusion of the project.

Standardised information of this kind could substantially improve

decision-making. It would facilitate government purchasers' learning from past experience of both major problems and success stories. It would support better benchmarking of project costs, by facilitating up-to-date estimates of typical project elements such as a kilometre of road or track. If project costs and benefits can be estimated with greater precision, there can be a corresponding improvement selecting projects with the highest net benefits to the community.

Standardised project data would also make interstate and international comparisons more feasible and support Australia to strive for best practice. It would improve the federal government's capacity to meet its stated intention of using monitoring and review to improve risk assessment and manage program risk.⁹

A second kind of data that should be available is the business cases for individual projects. These business cases detail the benefits and the costs to the community of a particular project, and form the basis for spending public money. In most cases, it is impossible for researchers or members of the public to access these business cases. Some people claim that business cases are commercially sensitive; however, given that the business case is typically conducted before the tender process, such arguments are not compelling.

Publication of business cases would improve the transparency of government spending, and would increase accountability for decision-making.

⁹ [Australia/New Zealand Standard Risk Management – principles and guidelines.](#)

2 Controlling the privacy risk

Big data comes with risks, including risks to privacy.¹⁰ Privacy risks can be to individual consumers and to providers, but risks to the two different types of stakeholders are quite different

But custodians should not unnecessarily regard all data as private. For instance, currently institutions with the best survey data impose strict confidentiality requirements on researchers. We feel that these requirements, as well as being onerous, are largely self-imposed.

Surveys by institutions like the ABS are clearly intended to be used for research. Survey respondents volunteering this information do so with full knowledge of this purpose. It is therefore unlikely that the assurance of confidentiality has any effect on the information offered. Yet once this assurance is given, it must be honoured.

Before placing restrictions on data release, or issuing assurances about confidentiality, custodians should think critically about whether release would in fact breach someone's privacy or substantially damage the integrity of data. Restrictions should be withdrawn or not imposed unless there is clear evidence of their benefit.

In the past, data have only been released in broad, aggregated, tabular form either in standard published form or through user-specified tabulations. An evolution of this has been to provide

front-ends to data sets to allow users to generate tables on-line from a data extract. Typically the data access tools are structured not to allow personally identifiable data to be generated (e.g. by using age-ranges in drop-down menus).

Unit record data sets have been released by statistics agencies and other data custodians for about 30 years and two broad approaches to privacy risk mitigation strategies have been adopted by data custodians.

First, is *controlling outputs*. Data custodians can allow data analysis directly on data sets in controlled premises and vet all outputs (such as computer-generated cross-tabulations taken off the premises) to ensure none allow person identification.

The virtue of the output control approach is that it provides access (to approved people) for analysis purposes to extremely detailed linked data but minimises privacy risks. The downside is that it limits access to people able to visit specific locations who have prepared specific data analysis requests. This is one of the approaches adopted in New Zealand (see box).

¹⁰ Richards and King (2013)

Statistics New Zealand has carefully reviewed the privacy issues associated with linking data sets and releasing data in this way and shown that release of output from these linked data sets can occur in the context of stringent safeguards on privacy.¹¹

The second approach to mitigating privacy risks is to *control the data released*. This approach includes data modification strategies (e.g. introducing perturbations in small cells¹²) as well as limiting the variables released in one data set (e.g. not releasing detailed age and detailed geography as it may allow re-identification).

In both cases, data custodians may *control users to whom data are released*. This may include police checks on data users, only releasing data to reputable organisations or organisations with ability to guarantee security of access and so on.

Most data custodians treat different classes of data users differently, and this is appropriate. Data sets are often not released directly to doctoral students, for example, their supervisors normally apply on their behalf and act as guarantors for appropriate use.

Similarly, it is appropriate to treat requests from individual, independent researchers, not affiliated with established research institutions, differently from researchers from recognised research institutes or universities. Researchers from established institutions have significant incentives to ensure data are appropriately handled and there are no privacy breaches, and

¹¹ Statistics New Zealand (2014)

¹² Mehmood, *et al.* (2016); Polonetsky, *et al.* (2016)

Box 1: New Zealand Data Labs

New Zealand provides access to rich, linked, long-term data about (de-identified) individuals. The data tell you about people's income, education, tax, health, employment status (and many other things), how all this changes over time, and how it all interacts.

The data can help the government understand the long-term budget impact of, for example, crime or education outcomes. These costs can come in a range of forms, like healthcare or social security spending, and can happen years down the track. That is why linking data over time is so important. The data can allow smarter ways to target services and spending where they are most needed, particularly preventive interventions. The data can also be used by external researchers and analysts to help improve the public debate about policy issues and options.

For researchers, the data are available in secure data labs. Once they are signed up and inducted, there are very few restrictions on the data that researchers can access. Instead, the outputs they generate are checked before they are allowed to take them out of the data lab.

consequently data provided to them requires less de-identification than data that is on open public access.

Whose privacy?

Controls over data release are primarily designed to protect privacy. Data custodians often treat privacy of the provider on the same footing as consumer privacy. This should not be the case, especially where providers are in receipt of public funds.

There are very different issues involved in release of information about providers, where the risks are generally associated with commercial issues, compared to the privacy risks about individuals, which are about sensitive personal information. These should not be treated as equivalent.

Typically, public hospital information is released with public hospital names suppressed. This is appropriate for almost all types of analyses. However, private hospital information is often grouped into a single 'all private hospitals in a state' group. This means that certain analyses (e.g. comparing public and private hospital behaviour) cannot be undertaken.

A significant proportion of private hospital activity is subsidised by the taxpayer through the private health insurance rebate. The public also has an interest in comparing attributes (e.g. quality or efficiency) of the two – public and private - hospital sectors.

Data custodians should revise their approach to release of private hospital data so that it is treated the same as public hospital data i.e. individual names should be removed and data users should not be allowed to publish anything which identifies the individual hospitals.

Medicare data which may identify medical practitioners in terms of

their volume and type of activity should also be able to be released into the public domain. Again data users should not be allowed to publish anything which identifies individual medical practitioners.¹³

In the education space, school profile data (such as socio-economic mix, funding levels, and school sector) is often withheld or highly restricted in system-wide datasets provided to researchers, even while the information about individual schools is freely available on the My School website.¹⁴ For example, in the analysis of student progress discussed earlier, the national dataset included data the ICSEA¹⁵ status of the school that each student attended, but not its location or sector. The Victorian dataset had even less information about the school that each student attended.

Data users should not be allowed to identify individual schools (or students) based on use of datasets provided for research purposes. However, data users should be able to access the more detailed data (potentially with school name de-identified) for the purpose of analysing patterns and trends. Making more information available would strengthen policy research and lead to policy recommendations based on better evidence.

¹³ Government should publish identifiable information about bulk-billing rates and average out-of-pocket costs for general practices and specialists.

¹⁴ See <http://www.myschool.edu.au>

¹⁵ Index of community socio-economic advantage

3 Moving forward on data release

Moving from rhetoric to action about data release would be facilitated by a number of actions.

Release metadata

Routine or 'administrative' data holdings are complex, with vast numbers of data elements collected. Some of these will be obvious and known to researchers e.g. Medicare Benefits Schedule (MBS) item numbers, but others may be system created e.g. information could be extracted about medical services based on year the service was delivered or year a claim is processed.

Information about the variables held and their definitions ('metadata') could be standardised and made publicly available, and the necessary resources must be provided to maintain and update the metadata.

Publication of information on data elements should be accompanied by standardisation of de-identification methods and publication of information about the rules used to protect privacy (e.g. suppression or randomisation of low frequency events).

Both publication of metadata and information about de-identification methods will help to make researchers aware of the data available. This will stimulate use of the data.

Better information about data characteristics will also improve the quality of requests for data extraction as researchers would be better placed to specify requests precisely in terms of the data

elements held.

Develop standard approval mechanisms

Current administrative arrangements provide that claims processing for major government functions (e.g. claims against the Medicare Benefits Schedule) be undertaken by the Department of Human Services whilst policy responsibility for these functions is vested in the relevant policy line agency, in the MBS case, the Department of Health and Ageing. Similar arrangements apply for social services data.

The current arrangements for separation of payment and policy means that although claims data is held by the Department of Human Services, decisions about data release are made by the Department of Health.

Data release decisions require co-ordination between the two Departments and at present this process tends to be slow and cumbersome. This problem is exacerbated when data from more than one Department is requested. A more streamlined management process should be implemented for cross-agency data release decisions.

A list of precedents of previous data approvals – and non-approvals - should be published to streamline data requests and approvals.

Developing common use datasets

Common or public use data sets are standardised extracts of data or reports from surveys which are made available for external use. They can exist at various levels: detailed published tables, data cubes (interactive tables which allow for personalised extraction of aggregate data), confidentialised unit record files based on a sample of records, and researcher access to unit record data under secure conditions.

The Australian Bureau of Statistics has an extensive program of releasing Confidentialised Unit Record Files (CURFs).¹⁶

Although special data extraction requests will continue to be required, secondary analysis would be facilitated by developing 'common use' data sets – data sets which contain core data elements which are made available to all approved researchers and research institutions. Such data sets may meet a significant proportion of common data extraction requests.

Development and release of common use data sets will help to streamline access and provide an important resource for research and policy analysis. In addition to facilitating research which would answer defined questions, a common use data set would also facilitate exploratory research, allowing researchers to undertake preliminary analyses to test whether specific research questions may be worth pursuing and whether special data extraction requests are warranted.

Although many common use data sets could be considered for early release under a streamlined access proposal, it is recommended that a data set consisting of confidentialised information of MBS use of a random sample of 10% of the population for a two year period be the first released. A new version of the data set should be issued annually.

This should be seen as a first stage with similar common use data sets created for PBS utilisation and, over time, data sets linking MBS, PBS and hospital use. Additional common data sets (e.g. provider based) should also be developed over time.

We also wish to note and praise the ATO's current efforts to construct 'longitudinal' sample files of individual taxpayers for future release. Grattan Institute has found the cross-sectional detailed sample files enormously useful for understanding distributional impacts of proposals, as well as to estimate their cost to the budget. However, we still lack reliable, detailed sources for lifetime effects, which would be illuminated by longitudinal files.

The Australia Bureau of Statistics now has extensive experience with releasing data sets and has routinised its processes, making a customer commitment to handle data request speedily, generally within a fortnight. It is able to meet standards such as this because it has developed common use data sets for release from some of its key collections/surveys. These data sets are already 'confidentialised' (the data sets are called Confidentialised Unit Record Files).

Conditions for release of common data sets should be modelled

¹⁶ Australian Bureau of Statistics (2016)

on existing ABS procedures including:

- Approval of the body to which data is released;
- Conditions on secure storage;
- Conditions on not identifying individuals in any publications;
- Prohibitions on on-passing data to third parties; and
- The payment of a nominal charge.

Each application for access to the common use data set should be required to provide a short outline of proposed use of the data to allow a determination that the proposed use meets a 'public interest' test as required by existing legislation.

The international precedent

Many overseas jurisdictions also release common use data sets. The United States Department of Health and Human Services has a very open policy about data availability. Research using important data sets, such as the Medicare claims data and hospital utilisation data has provided the basis for thousands of policy evaluations and research papers.

The U.S. Department has taken great strides in recent years to facilitate data access. It has established a website healthdata.gov

dedicated to making high value health data more accessible to entrepreneurs, researchers, and policy

makers in the hopes of better health outcomes for all.

This website refers to 'liberating' data in the description of the purpose of the website.

Australia needs a similar data access revolution.

References

- Academy of the Social Sciences in Australia (2013) *Facilitating access to routine data for research benefiting the Australian people*, The Academy
- Australian Bureau of Statistics (2016) 'About CURF Microdata', from <http://abs.gov.au/websitedbs/D3310114.nsf/home/About+CURF+Microdata>
- Duckett, S., Breadon, P., Romanes, D., Fennessy, P. and Nolan, J. (2015) *Questionable care: avoiding ineffective treatment*, Grattan Institute
- Einav, L. and Levin, J. D. (2013) 'The Data Revolution and Economic Analysis', *National Bureau of Economic Research Working Paper Series*, No. 19035,
- Goss, P., Sonnemann, J., Chisholm, C. and Nelson, L. (2016) *Widening gaps: what NAPLAN tells us about student progress*, Grattan Institute
- Kim, G.-H., Trimi, S. and Chung, J.-H. (2014) 'Big-data applications in the government sector', *Communications of the ACM*, 57(3), p 78-85
- McAfee, A. and Brynjolfsson, E. (2012) 'Big data: the management revolution', *Harvard business review*, 90(10), p 60-66
- Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G. and Guo, S. (2016) 'Protection of Big Data Privacy', *IEEE Access*, 4, p 1821-1834
- Polonetsky, J., Tene, O. and Finch, K. (2016) 'Shades of gray: Seeing the full spectrum of practical data de-identification', *Santa Clara Law Review*, 56(3), p 593-629
- Richards, N. M. and King, J. H. (2013) 'Three Paradoxes of Big Data', *Stanford Law Review Online*, 66, p 41-46
- Statistics New Zealand (2014) *Integrated Data Infrastructure extension. Privacy impact assessment - second edition*, Statistics New Zealand from http://www.stats.govt.nz/browse_for_stats/snapshots-of-nz/integrated-data-infrastructure/privacy-impact-assessment-extension-idi.aspx
- Turnbull, M. (2015) *Australian Government Public Data Policy Statement (released 7 December 2015)*, from https://www.dpmc.gov.au/sites/default/files/publications/aust_govt_public_data_policy_statement_1.pdf