Australian Government
Australian Institute of Health and Welfare

Authoritative information and statistics to promote better health and wellbeing

Mr Peter Harris AO
Chairman
Productivity Commission, Locked Bag 2, Collins St East
Melbourne VIC 8003

Dear Mr Harris,

In response to draft report on Inquiry into Data Availability and Use

The Australian Institute of Health and Welfare (AIHW) welcomes the opportunity to make a submission in response to the draft report of the Productivity Commission's Inquiry into Data Availability and Use.

As with the AIHW's submission to the Issues paper, this submission draws on AIHW's extensive health and welfare data experience, including our strategic data linkage relationships, to provide feedback and information to assist the Commission in finalising the report. In particular, we have responded to the Commission's information requests and have outlined a number of priority areas for attention. These include:

- Functions of the National Data Custodian and Accredited Release agencies
- Streamlining approval processes
- An integrated framework for data sharing and improving how linked datasets are available for subsequent use.

Should the Commission have any queries about the information I have provided or wish to seek additional data from the AIHW, please contact Mr Geoff Neideck, Head, Data Strategies and Information Technology Group, on (02) 6244 1163.

Yours sincerely,

for
Barry Sandison
Director (CEO)
Australian Institute of Health and Welfare

15 December 2016

AIHW response to the Productivity Commission's draft report of the inquiry into data availability and use

The AIHW welcomes the forward-thinking initiatives presented in the draft report of the Productivity Commission into Data Availability and Use. We also welcome the opportunity to provide feedback on how these initiatives might be applied in practice, to assist in shaping and implementing the recommendations of the Productivity Commission inquiry.

An integrated framework for data sharing and availability

The comments below are in response to issues raised in *Chapter 9: A framework for Australia's data future*.

Key recommendations of the draft report deal with umbrella measures that will aid in breaking down barriers to data sharing. To support administrative and technological solutions, a systematised approach to best practice in the use of data, including understanding the mix of skills required to achieve best outcomes, is warranted.

There are currently two main themes regarding access to public data:

- open data, that is, data that is available to anyone without restriction, sometimes described synonymously by the term 'public' data noting that public data can refer to the source rather than the mode of access
- data that is potentially re-identifiable or otherwise sensitive such that it cannot be released openly.

Ideally, from the perspective of the data user, discoverability and access procedures would cover both types of data in an integrated system. At this stage however, these systems have been developed separately. Aligning these systems into a common set of procedures would increase the effectiveness of the system as end users would only need to learn to use a single system for data access. This would reduce potential confusion and red tape associated with managing two distinct systems.

To realise the vision for an integrated data system that enhances data availability and use, end users need to know what data exists and how to access it. The first step in making data available for use is comprehensive discoverability. The second step is accessibility which includes easy to use access points and streamlined approval processes. Underpinning this is strong infrastructure and clear, comprehensive and consistent metadata. The third and final step is to enable appropriate usage of the data in an efficient way that realises as much value as possible from the data resource.

Strengthening the skills and capabilities of both end users and data custodians is critical. The AIHW's experience indicates that a collaborative approach between the data users and those with a deep understanding of the source data leads to best practice data use.

A simple model of an analysis team producing valued output has been presented in Figure 1 below. **Subject matter expertise** implies a strong understanding of the data source, its purpose and context, changes to data scope and coverage over time, together with the associated metadata and data transformations. This contextual information helps guide analysis and avoids inappropriate conclusions resulting from identification of spurious associations and results. **Analysis expertise** covers an understanding of robust techniques for producing and presenting information, an ability to mine and identify new associations and produce predictive models, but also an ability to apply statistical rigour, when required, to produce robust results applying appropriate scientific principles. **Data handling expertise** on the other hand implies a strong ability to handle, manipulate and manage data. This is of

particular importance in big data that may not be easy to analyse in full or may need additional expertise in order to prepare in a form that can be handled and managed by an analytics platform. Data scientists will often have skills encompassing both analysis and data handling and there are efficiencies which can be realised when an individual has skills in more than one arena. Finally the inclusion of **Policy expertise** recognises that results from analysis will return more value when placed into an appropriate context. In bringing each of these elements together judgement will ultimately need to be exercised to ensure that the data release is appropriate and does not involve undue risk.

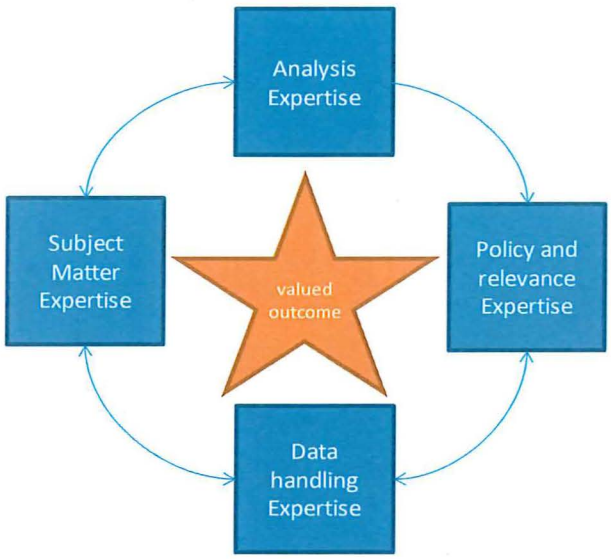

Figure 1: Desirable analysis team model

Key institutional roles in the proposed framework

The AIHW welcomes the proposed roles of the National Data Custodian (NDC) and Accredited Release Authorities (ARAs), together with the proposed nomination of National Interest Datasets. In relation to these recommendations we provide the following comments in response to the institutions and roles outlined in Box 2 of the draft report (page 16).

National Data Custodian

The role of the NDC could include the development and promotion of national standards and guidelines for data release, data sharing and data linkage. The AIHW has observed that for many data integration projects data custodians develop their own approaches and processes for extracting, preparing and transferring data. In approving the release of data the assessment and approval processes are developed in-house and different processes occur from project to project.

A particular issue that has been emerging is the minimum security requirements for researcher access to data. The Sax Institute's SURE environment, the Population Health Research Network (PHRN) data linkage nodes, the NSW Data Analytics Centre, and the ABS's and AIHW's internal environments are all able to provide high levels of security and data governance. However, these sites and others proposed to host sensitive data are often rejected in the absence of nationally-agreed standards on acceptable security and data governance arrangements. Approval processes can be delayed as data custodians consider what hosting sites are acceptable for particular proposals.

Experience from Europe and North America suggests that where national standards are adopted there is improved public trust in public data sharing regimes. The Information Governance Review conducted in the United Kingdom recommended:

> *The Review Panel concludes that in order to support professionals and staff in making data sharing decisions, and to respond to mistakes when they occur, clear good practice guidance and local policies are a vital resource along with national statements, guidance and standards.* (Caldicott review: information governance in the health and care system, National Data Guardian, 2013, page 47)

The AIHW suggests that national guidelines and standards be considered in the following areas:

1. Data governance
2. Data project risk assessment
3. Trusted researcher assessment
4. Data extract and transfer
5. Data security and secure data environments
6. De-identification
7. Data encryption

The NDC would need to work closely with the Office of the Australian Information Commissioner, state/territory privacy commissioners and with research bodies such as the National Health and Medical Research Council to ensure consistency of approaches. It may be useful to consult with Standards Australia in these developments.

There is also a need for a common lexicon of terms around data use. A common problem in sharing information across sectors is the confusion of terms. The following terms are variously used and can mean different things in different settings—data linking; data matching; data merging; linkage keys, spines/indexes; de-identification, potential re-identification; etc. The NDC could play a useful role in promoting consistency of use of terminology. The PHRN has produced a useful glossary that could assist in this activity (see http://www.phrn.org.au/media/27174/phrn%20glossary.pdf). Further, given some of these terminology issues are legislatively based, the proposed Act could also help provide clarity. For example, de-identification is defined under the Commonwealth *Privacy Act 1988* (and that definition is not reflected in the PHRN glossary for the term de-identification).

It is recommended that a key role for the NDC will be to develop practical best practice guidelines and programs to support data access and release. This could be supported by training for practitioners and more in depth courses, such as the Master of Applied Data Analytics recently developed at the Australian National University.

Accredited Release Authorities

The AIHW recommends that the Productivity Commission consider separate institutional roles for "Accredited Linkage Authorities" (ALAs) and "Accredited Release Authorities" where ARAs are responsible for single source data while ALAs manage the production of multiple source data and may also be accredited as ARAs for the release of data.

It is important to recognise the nature of the public datasets that will be created and released under the proposed arrangements. Many datasets will be from a single administrative or survey source and will be produced and curated for researcher access. Either the full dataset or a derivative of it (for example, extracts or summary tables) may also be made available for public access. There is considerable analytic value and increasing demand in bringing

datasets together. Undertaking the data linkage to produce these datasets necessarily requires access to personal information (names, addresses, unique identification numbers, etc.).

In order for linkage to occur agencies with expertise in data integration processes generally need to be involved. It will not necessarily be the case that all ARAs will want to offer linkage services. It is therefore suggested that a distinction be made between agencies undertaking data linkage, who may also undertake data release, and those only managing the release of data.

At present there are various arrangements to accredit data linkage authorities. The Commonwealth has a process for establishing Accredited Commonwealth Integrating Authorities under the auspices of the Secretaries Data Group. Each state and territory has established data linkage centres which operate as part of the PHRN. These arrangements, in a number of states, are being extended to other agencies to link data from other sectors.

Policy analysis is currently limited because jurisdictions are not able to access datasets held outside their jurisdiction. For states and territories to get a more complete picture of their population, the use of services and outcomes for people and businesses, state governments need access to Commonwealth data, as well as data from other states and territories (for example, to capture instances where people from one state access services in another state). Similarly, for Commonwealth agencies, policy development and delivery can be improved by access to state data.

While cross-border data sharing is increasing, significant barriers and inefficiencies remain. In some sectors, sharing de-identified data is possible. However, application (or interpretation) of various jurisdictional legislation limits the transfer of personal information needed for data linkage. The Commonwealth Secretaries Data Group has recently opened the way for state bodies to be accredited as integrating authorities for Commonwealth data under certain conditions, including that they be covered by the Commonwealth *Privacy Act 1988*. The accreditation process and what might be acceptable privacy coverage are still to be determined.

For a truly effective national system for data sharing, cross-border data sharing needs to operate considerably more freely than at present. It is recommended that these be addressed in arrangements for establishing ARAs (and ALAs) across the nation.

National Interest Datasets

The comments below are in response to the section *Broad access to datasets of national interest* (page 19).

In establishing National Interest Datasets (NIDs) it is worth considering at what point in the data lifecycle a particular source should be deemed to be an NID. For example, national income support data is an important source of data for research and analysis. However, it exists in many forms. The data holding at source in the Department of Human Services (DHS) is vast and is stored and managed in complex ways that make it incompatible for direct access for research and analysis. Both DHS and the Department of Social Services hold various extracts of income support data created and used for different internal purposes. Some datasets are also created for research access, such as the Priority Investment Approach (PIA) dataset created for undertaking social investment modelling. The PIA dataset is now being made more widely available for research use.

From both a data custodian and researcher perspective, there are costs and timeliness issues associated with data in these various formats. The rawer the form of the data the more work researchers need to undertake to manipulate the data to prepare it for analysis. However,

there is considerable cost and effort involved for the data custodian to prepare data for researchers. A decision on the optimal dataset to be released will need to be negotiated between the ARA and data owner (if they are different agencies). It is likely that high value and high use data sets will warrant greater effort in preparation.

In response to issues raised in *D.1. Health data the policy and IT landscape* (page 503), the AIHW considers that the implementation of NIDs provides the opportunity to identify gaps in the data needed for important social and economic policy development. The development and use of NIDs may highlight information needs in other areas. The framework for establishing NIDs should develop in such a way as to assist with identifying data gaps.

Demand for linkage

Comments here are in response to *Sharing public sector data with researchers: The extent of data sharing for linkage and integration* (page 114).

It is stated when describing the number of linkage projects undertaken by Commonwealth accredited integrating authorities:

> *"Indeed, reported data linkages among relevant institutions suggests only a few jurisdictions are undertaking a substantial number of linkage projects (figure 3.3):*
>
> - *Data Linkage Western Australia has been integrating unit record administrative datasets since 1995, averaging 45 integration projects per year.*
> - *Similarly, the Centre for Health Record Linkage (New South Wales) has been integrating datasets since 2007 and has averaged 32 projects per year.*
> - *SA-NT DataLink (South Australia and the Northern Territory), however, has completed only 7 projects per year on average since 2011 (though this has been trending upwards with 18 projects completed from mid-2015 to mid-2016).*
> - *Commonwealth accredited integrating authorities have registered only 3 projects on average per year from 2005 to 2015."*

The AIHW suggests replacing the last dot point with the following:

> *Demand for data linkage services at the AIHW has generally been increasing, along with an increased number of queries for integration services. The complexity and size of completed linkage projects has also notably increased. The AIHW undertook 35 linkage projects a year on average over the last 3 years and expects more than 50 projects to be completed in 2016–17.*

Examples of data linkage projects underway or completed at the AIHW that the Commission could reference include:

- Pathways in Aged Care—this linked dataset covers aged care assessments and use of 7 different Commonwealth aged care service programs from 2002 to 2011, as well as deaths. The study showed that even where people are eligible for residential aged care they prefer to remain in their homes for as long as they can.
- Cancer risk in people exposed to computed tomography scans— Medical Benefits Schedule data on 680,000 computerised tomography (CT) scans were linked to the Australian Cancer Database to examine if there was increased incidence of cancer after CT scan exposure. The study showed exposure to CT scans in childhood increased the incidence of cancer.
- Impact of a population-based human papilloma virus (HPV) vaccination program on cervical abnormalities—this collaborative study between the AIHW and the Victorian Cytology Service linked the National HPV Vaccination Program Register with Victoria's

Pap test register. The study was the first in the world to show a population-based HPV vaccination program resulting in a fall in cervical abnormalities within 5 years of implementation.

- Child protection data and educational achievement—this project involves linking child protection data with educational achievement data as measured by NAPLAN (National Assessment Program—Literacy and Numeracy) testing.

Streamlining approval processes

These comments are in response to Draft Recommendation 5.4 (page 29).

Linkage studies require 'identified' data (either with a linkage key or linkable data items such as names) at the unit record level. However, the process involved in obtaining access to jurisdictional data for linkage purposes has often proved to be complex and time-consuming.

In its draft report, the Commission recommends streamlining access to identifiable data within and between Australian governments, and for the limited range of other trusted users with which such data is shared (Rec. 9.8). It also recommends promoting mutual recognition of Human Research Ethics Committee approvals (Rec. 5.4). Section 5.3 of the report, entitled 'Lengthy approval processes waste time and money', notes that the current processes, while intended to ensure privacy and confidentiality are maintained, create a major obstacle for data access.

As noted in the overview to the draft report:

> *"...access requests... can require separate and duplicative agreement of multiple dataset owners, custodians and stewards, integration units, ethics committees, other advisory bodies, and the individuals about whom the information was collected. Each policy and approval step is intended to ensure privacy and confidentiality are maintained, but in combination they create major obstacles to data access."* (page 21).

The AIHW supports the Commission's recommendation to streamline access. An option to assist in streamlining access could involve standardised data sharing and use agreements, to the extent possible.

Examination of case studies of projects which had particularly slow approvals processes shows that in some cases the experience of successful approval has led to a somewhat streamlined process for future projects (particularly for the same data sources).

An example of this is the Enhanced Mortality Database (EMD) project. The EMD project linked four data sets, three of which are owned by the various jurisdictions. The data sets are the National Death Index (compiled from jurisdictional death registration records), the National Hospital Morbidity Database, the Residential Aged Care Database and jurisdictional perinatal data collections. The project was undertaken in 4 phases, with two separate linkages.

The main lessons learnt from the extended EMD approvals process were the requirement for a large number of ethics approvals, for which there are now various national processes, such as mutual recognition between state committees that are starting to come on line and the widely varying differences in approval times for release and supply of the component datasets.

This is however, in contrast with difficulties created by continuing lengthy timelines for many other projects. The AIHW had a recent case where researchers gave up on project proposals due to delays (primarily associated with obtaining data custodian approvals) and not being able to meet timelines for research funding. The AIHW also has a number of other

projects at risk of being discontinued due to the lengthy timelines for approvals and source data supply.

There are various national processes such as mutual recognition between state committees that are starting to reduce the number of ethics approvals that may be required for particular linkage projects, although progress is slow. In a similar way it would be hoped that the creation of a National Data Custodian and clearer guidance for custodians would reduce the inconsistencies and inefficiencies in applying to multiple data custodians to complete a linkage process. When it comes to data custodian approvals, mutual recognition is less likely to work, due to the differences between datasets, but what may be relevant is the ability for a data custodian to pre-prepare a view or version of their data that meets a pre-specified level of trusted access. In this way mutual recognition of the level of 'trusted-ness' of the person accessing the data, their project and access mechanism could allow multiple data custodians to make an expedited application for data custodian approval.

An additional consideration is providing enduring approvals for projects of ongoing value. Certain current documentation for ethics committees, Public Interest Certificates and data custodians follow standard templates and procedures. It is the AIHW's view that development of similar templates to support applications for the establishment of ongoing use of linked datasets, and ongoing approvals for these, would assist substantially in streamlining approval processes.

Improving how linked datasets are available for subsequent use

The Commission has sought further views *on the most practical ways to ensure improvements to linked datasets are available for subsequent dataset uses* (page 29).

When datasets are linked at the highest level of quality, including a manual clerical review process, the following information will be generated through this process:

- dataset and collection metadata specific to linkage requirements
- dataset specific information about processes for data cleaning and the standardisation of information and results
- variable harmonisation information
- linkage keys
- additional intelligence about the entities in the established linked records that could help future linkages such as nicknames, address changes, maiden name information, etc.
- drivers of false positive linkages specific to the collections.

This intelligence base of information can be particularly useful because there is no single identifying characteristic that can be used to reliably identify records across different databases. For example names can be recorded differently due to nicknames, mis-spellings, maiden names and anglicisations. More objective information, such as gender and date of birth will generally be more reliable, but is often recorded differently in different databases and is also less discriminating when used for linking records.

To ensure this information is retained in such a way as to lead to continual improvement, at a minimum, linkage keys from each linkage project should be retained for re-use. In the case of an appropriately accredited linkage authority, linkage keys could also be retained on a spine file so that updated information could be used to improve the links whenever further information is received. Ideally this information should be fed back through the data system to source, so that the data collection point has the opportunity to re-confirm their data by contacting the original unit.