CENTRE FOR POLICY DEVELOPMENT

CREATE
CONNECT
CONVINCE

CPD Response to the Draft Report by the Productivity Commission's Data Access Inquiry

Centre for Policy Development
December 2016

Author: Geoff Shuetrim, Member of the Research Committee

MELBOURNE
15-31 PELHAM STREET
CARLTON
VIC 3053
(03) 9929 9915

SYDNEY
LEVEL 6, 115 PITT STREET
SYDNEY
NSW 2000
(02) 8199 9407

www.cpd.org.au
ACN: 124 425 896
ABN: 82 124 425 896

About The Centre for Policy Development (CPD)

CPD is an independent, non-partisan public policy institute. Our mission is to foster an Australia that embraces the ‘long-term now’. In doing so we seek a future for Australia based on shared prosperity and sustainable wellbeing. One of our three key research programs - *Effective Government* - is dedicated to understanding the role for an active, capable and effective government in the 21st century.

Table of Contents

1. Overview

This feedback has been prepared in response to the draft report released by the Productivity Commission on November 3, 2016 (referred to hereafter as 'the report') and in response to discussions at the Sydney hearing following that release.

At a high level, adoption of the recommendations put forward in the report would take Australia in the right direction. It is important to make better use of the data resources available in Australia. It is equally important that the data is being stored and exchanged using systems that ensure levels of privacy that are commensurate with the sensitivity of the data. The recommendations in the report recognise both of these imperatives.

Our specific feedback focuses on a small number of issues, largely in response to requests for further input by the Productivity Commission.

2. Retention of linked data following completion of research

The report recommends (recommendation 95.3) that the Australian Government abolish its requirement to destroy linked datasets and statistical linkage keys upon completion of researchers' data integration projects.

The efficiency gains deriving from this recommendation are clear. However, the recommendation, in conjunction with other recommendations that such linked data sets be preserved and made available to other trusted researchers places an excessive burden upon researchers to maintain the data in suitably secure systems over potentially long periods of time.

The Productivity Commission should instead consider alternatives whereby stronger incentives are provided to researchers to preserve and make available the programs that they have used to construct the linked data sets from the original data made available from various sources.

This alternative approach eliminates the risk of security breaches associated with storage of sensitive data by people and organisations that simply are not suitably equipped to protect sensitive data.

It will also have the benefit of preserving the knowledge about how data sets were constructed. This is often a key area of contention between researchers and greater transparency on that front will be valuable. By conserving and sharing the intellectual property required to do linking, the linking process becomes both repeatable and transparent.

Another benefit of preserving the data set construction programs is that it will make it easier for future researchers to produce up-to-date linked data sets rather than relying on aging but linked materials. With a proliferation of preserved accessible and linked data sets, there would soon be researchers looking to link up separate linked data sets. This will generally be tough to do if the separate data sets are different ages and have been generated using different approaches to linking.

Relying upon research organisations to manage long-term storage and dissemination of sensitive data that has been provided in confidence by other data collection agencies will eventually result in the kinds of security breaches that will undermine public confidence in Australia's handling of data resources.

3. The benefits and costs of requiring APIs for data sharing

The report requests information on the use of APIs for data sharing. The distinction between an API and a data format is important and did not appear to be made in the report. A data format is the syntax used to express data. An API is a means of interacting with data at a machine level, getting access to it and potentially modifying it. When data is accessed through an API, it can be made available in various formats.

Not all formats are equally useful and a proliferation of different formats can make it more challenging for users of data to work across different data sources.

Likewise, APIs can be designed and implemented in sufficiently different ways that their proliferation can make it more difficult for data users to tap into data from different sources.

By avoiding these proliferations, there are clear benefits in standardisation of both data formats and APIs. However, there are also significant costs that go beyond those involved in just forcing all parties involved in the exchange of a particular type of data to invest in systems that conform to a specific API and provide data in a specific format.

First, there is the question of design for the data format and the API used to expose that data. Depending upon the complexity of the data, the stability of the data (is the information being extended in various ways over time), and the number of stakeholders that need to participate in the agreement process, this design work can take years. In the meantime, data is often not shared because investments in systems are delayed until agreement is reached because of the option-value of waiting. This delay cost should not be underestimated.

Working with a variety of data formats and a variety of APIs is more complex but not enormously so in many knowledge domains. For example, biology taxonomic information is exposed by many scientific organisations around the world. They have not even agreed on a format for uniquely identifying a taxon. They have not agreed on data formats. They have not agreed on APIs. Work is underway on such issues through the Taxonomic Database Working Group but, in the meantime, data needs to be shared and organisations are doing it in different ways. So long as organisations maintain reasonable stability in how they do the data sharing, it is relatively easy to work across the variety of approaches as a user of the data being provided. Moreover, there is a role for data intermediaries, in accessing data from various systems and homogenising that data to make it easier to work with for end users.

Second, agreed data formats and APIs can limit innovation. If the underlying data becomes richer but that new information cannot be incorporated into in the agreed APIs and data formats, then it may not be shared and there may even be resistance to the data augmentation. It is important that Government-encouraged API agreement does not create such artificial barriers to innovation.

4. Methods of disclosure for consumer data

The report requested views on the effects of providing access to consumer data. In particular, views were sought on the potential creation of incentives for deliberate de-identification of data holdings to avoid providing access.

It is clear that such a response to regulatory costs could be driven by forcing organisations to provide data access to consumers. However, that response would tend to arise only in such situations where the marginal value of keeping the data identified were lower than the cost of providing access to that data. This would rarely be the case, except in those situations where release of the data also communicated the intellectual property of the organisation holding the data. Given the relatively broad definition of consumer data that is being considered in the report, this could well be the case for a number of industries.

For example, in the banking industry, a wide range of data is used in assessing creditworthiness. The factors considered by banks vary somewhat between organisations. If banks were required to provide access to the full set of information on all such factors taken into account by their credit assessments, then they are effectively also communicating the nature of their credit assessment system, something that underpins their competitive advantage in the industry.

These situations, where consumer data and the intellectual property of the organisation are inextricably intertwined, are the ones where perverse incentives can directly affect the efficiency with which a business can be run. Rather than attempting to regulate or influence the perverse incentives, the Productivity Commission should instead consider working with a definition of consumer data that does not cause the perverse incentives in the first place.

5. Data security

The report places little emphasis on data security, taking it as a given that security will be required to be suitable to the sensitivity of the data. However, in the future envisaged by the report, there will be many more copies of sensitive data, in many more systems and being transferred between those systems.

The final report should include recommendations about the security principles that need to be in place, such as those raised in the original CPD submission, in relation to sensitive data storage and transfer. The final report should identify the government body with responsibility for establishing those principles and for ensuring that they are applied by Government agencies, research entities, businesses and perhaps individuals.