Data Availability and Use: Response to the Draft Report

Submission to the Productivity Commission

Dr T'Mir Danger Julius

Head of Data Services

Centre for Transformative Innovation
Swinburne University of Technology

DECEMBER 2016

The Productivity Committee has invited comments on its draft report investigating the availability and use of public and private sector data. In my position as a data scientist with the Centre for Transformative Innovation at Swinburne University of Technology, I understand the value of open data from a research perspective and have experienced, first-hand, the pitfalls of poor data management and policy. It is my opinion that non-sensitive data should be more readily available to researchers and the public in a bulk, machine-readable format, but it is my opinion that a blanket policy of making de-identified data readily and perpetually available to the public, as is the ultimate aim of the recommendations of this report is feasible.

Recommendation 3.2

"Publicly funded entities, including the Australian Research Council, should publish up-to-date registers of data holdings, including metadata, that they fund or hold.

Publication of summary descriptions of datasets held by funded researchers but not released, and an explanation of why these datasets are not available, are also essential and would provide far greater transparency about what is being funded by taxpayers but withheld."

Hosting data in a way that is useful and secure is a costly exercise requiring significant expertise. If data registries are to be maintained by publicly funded institutions and data is to be made available to other researchers, funding will need to be provided to offset these costs.

Additionally, it is infeasible to store datasets indefinitely. An upper limit on the amount of time a dataset is expected to be maintained after the completion of a project should be established.

Recommendation 5.1

"In conjunction with the Australian Bureau of Statistics and other agencies with data de-identification expertise, the Office of the Australian Information Commissioner should develop and publish practical guidance on best practice de-identification processes.

To increase confidence in data de-identification, the Office of the Australian Information Commissioner should be afforded the power to certify, at its discretion, when entities are using best practice de-identification processes."

Data de-identification, much like cyber security, needs to be at least as strong as the technology available to break it. As technological advancements are made, and additional datasets become available (both with and without the blessing of data custodians) the de-identification of datasets will be easier to crack. With this in mind, it is not possible to create a best practice de-identification process that would secure data in perpetuity, and it would not be advisable to release de-identified sensitive datasets to the general public regardless of whether or not up-to-date best practice guidelines are followed. As it is not possible to recall and strengthen de-identified data after release, a certification of having followed best practices at the time of release is not an enduring measure of security, and so the risk will become greater over time.

Giving Data Away

The argument that uninformed users are making data freely available, so any further release of their data is low-risk, is misleading. The report identifies that half of Australians "are not fully aware of what data is collected about them and how", and that 47% of Australians alter personal information in an effort to preserve their privacy. However, rather than taking this to mean that the public is concerned about their privacy but need help in managing such, the report seems to use this as evidence that the public's privacy is not worth protecting.

It is also stated that "breaches due to sharing or release are far fewer in number and reach". It is entirely possible that this is due to the relatively small amount of de-identified data that is (1) officially published for public access and (2) highly sensitive (once re-identified). Since highly sensitive personal information (e.g. medical records) is far more rare and valuable than less sensitive personal information, de-identified sensitive data constitutes a far more attractive target for malicious activity.

Recommendation 6.2

"The private sector is likely to be best placed to determine sector-specific standards for its data sharing between firms, where required by reforms proposed under the new data Framework.

In the event that voluntary approaches to determining standards and data quality do not emerge or adequately enable data access and transfer (including where sought by consumers), governments should facilitate this, when deemed to be in the public interest to do so."

The private sector is not in the best position to determine data sharing standards (which affect the rights of the consumer) when it comes to privacy of sensitive personal data, as this would likely present a conflict of interest when consumer data is such a valuable asset.

Recommendation 9.2

"Individuals should have a Comprehensive Right to access digitally held data about themselves. This access right would give the individual a right to:

• continuing shared access with the data holder

• access the data provided directly by the individual, collected in the course of other actions (and including administrative datasets), or created by others, for example through re-identification

• request edits or corrections for reasons of accuracy

• be informed about the intention to disclose or sell data about them to third parties

• appeal automated decisions

• direct data holders to copy data in machine-readable form, either to the individual or to a nominated third party. Individuals should also have the right, at any time, to opt out of a data collection process, subject to a number of exceptions. Exceptions would include data collected or used as:

• a condition of continued delivery of a product or service to the individual

• necessary to satisfy legal obligations or legal claims

• *necessary for a specific public interest purpose (including archival)*

• *part of a National Interest Dataset (as defined in Draft Recommendation 9.4).*

The right to cease collection would not give individuals the capacity to prevent use of data collected on the individual up to the point of such cessation."

It is unreasonable to deny users the right to request the removal of identifiable data from databases for privacy reasons. It may be the case that users have been misinformed or have misunderstood the data being collected, or have become uncomfortable with the information being held about them by a party. For privacy reasons, users should have the right to request that sensitive information about them be destroyed by third parties.

Recommendation 9.4

"NIDs that contain non-sensitive data should be immediately released. Those NIDs that include data on individuals would be available initially only to trusted users and in a manner that retains the privacy of individuals and/or the confidentiality of individual businesses. The in-principle aim should be for these de-identified datasets to be publicly released in time."

It should not be the default for de-identified sensitive datasets to be publicly released in the course of time. It is not possible for a dataset to be robustly de-identified beyond all possibility of re-identification and remain useful. As discussed earlier, with enough external information and computational resources, de-identification can be reversed. Once a dataset is released, control of the dataset cannot be regained, and so once a sensitive dataset is publicly released the risk of re-identification is significant.