Data Sharing and Release legislative reforms

The opinions expressed in this submission are the authors' own and do not reflect the views of their employer.

Responses to questions

1 Do you think the distinction between data sharing and data release is clear?
Yes.
Release is open access for the world at large. Release is uncontrolled and hence permanent. Once data is available on the internet, it can be assumed that copies – or copies of copies - will remain in existence even if the original version is removed. These copies may exist outside of Australia and hence physically beyond the reach of Australian law.
In contrast, sharing means providing access in a controlled, limited way. This requires that access is temporarily and that those granted access do not receive the right to transfer access, complete copies or derived works to third parties.

2 What are the challenges for open release of public sector data?
Once a public dataset has been downloaded, open release cannot be reverted even if the de-identification procedure is later on demonstrated to be insufficient. A practical problem here is that re-identification attempts are not limited to using what is available within the released data, but can also incorporate every other database ever released or leaked.
A formal ban on re-identification may prove difficult to enforce. This puts a burden on data custodians to attain sufficient knowledge in re-identification and differential privacy. While the technical advise could be sought from consultants – people external to the agency in charge of the data - ideally the same people also have a thorough understanding of the data to be released, usually people employed by the agency.

3 Do you think the Data Sharing and Release legislative framework will achieve more streamlined and safer data sharing?
The safest option is to not collect the data in the first place, followed by not sharing the data. It is impossible for the framework to lead to safer data sharing when the baseline is not sharing the data. At best it can be safer than a hypothetical framework with less attention to safety.
A legislative framework and standardised forms may streamline the generation of requests, however as noted in point 2 above, the case-by-case assessment will require expertise that may not be available in every agency.
For some purposes the minimum dataset that is sufficient to provide a useful answer may have a high risk of re-identification. Such data should only be accessible (shared) in a controlled data lab. Fortunately, data labs are mentioned on the ABS' website examples of the five safes framework. Unfortunately, if data sharing requests are streamlined, the demand to use data labs may grow beyond supply.

4 What do you think about the name, Data Sharing and Release Act?
"Responsible Data Sharing and Release Act" may inspire more trust as it de-emphasises the need to share and release all data and provides a requirement to at least consider the possible negative consequences of sharing and release.

5 Do the purposes for sharing data meet your expectations? What about precluded purposes?
The types of purposes and precluded purposes seem appropriate. However, the purpose test does not explicitly ask the question whether or not the data is truly fit for the purpose. For health research, ethics committees will simultaneously evaluate a research proposal on the questions:

a. is the potential harm to people (or animals) minimised?
b. does the research address a question of public interest?
c. is the research likely to succeed?
Questions b and c will typically be addressed also on a research grant application and often arguments take the form of proposed statistical analysis plans and formal sample size calculations (despite limitations to such calculations). Ideally, the analysis plan is developed with input from people familiar with the data as it may contain caveats that the user is unaware off, in addition to the lack of control and randomisation.

6 What are your expectations for commercial uses? Do we need to preclude a purpose, or do the Data Sharing Principles and existing legislative protections work?
Open release of data will enable commercial use unless restricted by licences akin to the creative-commons non-commercial licence.
In the context of controlled sharing of data, allowing only one enterprise access may provide an unfair advantage to the recipient. As such, it can be expected that, once a sharing agreement is in place with one entity, the competitors will insist on receiving identical access, thus leading to a larger number of people requesting access. The means to acquire accreditation may depend on the size of the commercial entity and by itself favour larger corporations.

7 Do you think the Data Sharing Principles acknowledge and treat risks appropriately? When could they fall short?
Unlike the main documents, the Data Sharing Principles do mention that 'demonstrating feasibility' may be included in an initial data request. This should be upgraded to a key requirement to address as the feasibility question is different from the questions 'is the proposed purpose in the public interest?' and 'what are the risks to privacy?' Again, assessing these questions requires expertise: it requires familiarity with the data and familiarity with the research methods. If it is not possible to assess the feasibility of the research, this should be explicitly admitted and taken into account when evaluating the risks.
The Data Sharing Principles document mentions 'direct identifiers' indicating the existence of other identifiers but does not define or provide examples. The combination of age and area code at a fine resolution may result in unique combinations for a large fraction of the population; the combination can be used to re-identify some, possibly all people in a database that was de-identified.

8 Is the Best Practice Guide to Applying Data Sharing Principles helpful? Are there areas where the guidance could be improved?
The guide is helpful in clarifying that secure locations are considered and that ethical approval may be a prerequisite. Ideally, the ethical aspects are judged by a team that is familiar enough with the data to understand the consequences of a leak. Separating the ethical approval process from the data sharing request is inconvenient if the data custodian mandates changes to the protocol as they would need to be re-evaluated by the ethical committee. Even if an application passes both committees on the first attempt, the double procedure will undoubtedly be perceived as a double hurdle to researchers (triple if funding was requested).
Depending on the level of detail in the requested data, multiple application streams with tailored forms could be imagined.

9 Do the safeguards address key privacy risks?
The guide minimises the risk of a data breach and focusses on accidental or deliberate misuse by people who were granted access: "International and Australian experience in data sharing has shown that the main cause of data breaches is people making mistakes when using data rather than failures of technology or deliberate misuse." No citations are provided to back this claim and it is not clear what constitutes making a mistake. Is storing passwords in plain text instead of encrypted a mistake? It is the root cause of multiple high profile "hacks". Reporting about data breaches in the

news seems to occur more often since 2015, possibly due to requirements to notify users of breaches upon discovery. Without such requirements and good adherence to them, it is impossible to know the true incidence of breaches, whether due to 'technical failures' or targetting by a malicious third party. One step further, it is also required for the institute that was breached to actually notice an intrusion happened. As the original copy of the data need not be destroyed, it is theoretically possible for a breach to remain unnoticed.

10 Are the core principles guiding the development of accreditation criteria comprehensive? How else could we improve and make them fit for the future?
It is not clear what is meant by: "individual users within the organisation will have to undergo training that ensures they have the skills to protect, manage and use data."
Training should include protection against data breaches (encryption, password hygiene) and accidental leaks such as tables with small cells leading to easy re-identification. Training in proper use (statistical or qualitative analysis) of data is desirable but beyond the scope of data security accreditation.

11 Are there adequate transparency and accountability mechanisms built into the framework, including Data Sharing Agreements, public registers and National Data Commissioner review and reporting requirements?
This section is not finalised: "We are currently considering what kind of data breach scheme is necessary to safeguard the Data Sharing and Release legislation."
The requirement to report breaches seems essential to maintaining trust and evaluating the frequency of breaches. Where a breach is noticed by an accredited user working at an accredited organisation, this may provide a conflict of interest and a fear of retaliation. If the accreditation of the organisation is temporarily suspended due to a notification, this will affect the work of all its accredited employees. Hence, employees that notice colleagues are complacent, may not be motivated to report this as it may affect their own work.

13 Have we got our approach to enforcement and penalties right for when things go wrong? Will it deter non-compliance while encouraging greater data sharing?
Figure 7 provides a full spectrum of enforcement from education to litigation, which seems reasonable but also seems too simple.
Will enforcement be proportional to the size (value) of the data?
Who is to blame when data is stolen from an accredited analyst working at an accredited organisation? The data thief, the analyst, the IT department or the data custodian who authorised data usage outside a data lab? What if the thief sells the data to a commercial enterprise and the damage to individuals is a result of the way the enterprise exploits the data?
The liabilities are least clear in the middle ground between open release and tightly controlled sharing within a data lab. This middle ground has more potential types of actors.
Hence, a simple solution would be to eliminate this middle ground as much as possible by defaulting to sharing in a controlled environment. Again, this will likely lead to an increased demand for utilisation of data labs. This demand includes both the use of hardware and the support for software as each analyst will have his or her own preferences regarding software even when they work in the same field. In addition to hardware and software support, data analysts may need - or at least they could benefit from – direct contact with the data custodian to answer queries regarding the nature of the data: when results do not make sense, the primary suspects are errors in the data and/or errors in the analysis; the primary suspicion is not a scientific breakthrough...