Data Sharing and Release Legislative Reform Discussion Paper

Joint submission by the Australian Academy of Science and the Australian Academy of Technology and Engineering

The Australian Academy of Science (AAS) and the Australian Academy of Technology and Engineering (ATSE) welcome the consultation on the Office of the National Data Commissioner (ONDC) Data Sharing and Release framework (DS&R). This submission draws on the expertise of Fellows of both Academies, the AAS National Committee on Information and Communication Sciences (NCICS), the AAS National Committee on Data in Science, and the AAS/ATSE joint group on Digital Futures.

The Academies welcome calls for an increased role for the research sector in evidence-based policy making. By increasing transparency with data about government and public service decision making, the Academies expect the Data Sharing and Release framework (DS&R) to move towards the stated goals of effective policy development, especially science, technology and environmental policy.

We welcome calls for standardisation of access to data of societal importance and scientific significance, where DS&R is coupled with international best practice in privacy safeguards and accurate economic, statistical and scientific modelling.

1. Do you think the distinction between data sharing and data release is clear? How could this distinction be clearer?

The distinction allows a legislative framework to handle the complexities of sharing data under different categories. It would be constructive to develop tests that distinguish between open data (for release) and closed data (for sharing), to ensure this distinction can be made to work in real contexts. These tests would help to establish the boundaries of data between two clearly defined data categories.

2. What are the challenges for open release of public sector data?

The paper states that the legislation will not provide a new legislative authorisation for open data release as the current mechanisms are considered sufficient. However, under the proposed legislative framework, the use of shared data between agencies may respond to new drivers. For example, shared data might become a way for managing potential risks across agencies.

This is again a reason for tests for data release to be established. Agencies should be encouraged to develop a data sharing and release charter towards citizens, to commit to such data being released while addressing privacy concerns. Given the legislation is not proposed to provide compelling powers under legislation or through the commissioner, it will be important for agencies to develop such charters to make them accountable for their use of data.

Clarity regarding ownership of data must also be made explicit. While data curators manage data access and use under DS&R, data cannot be owned in the sense of property rights.

Furthermore, while data collected as legislated or explicitly consented may be used for primary purposes now, DS&R moves to relax the need for consent to secondary uses. The National Statement on Ethical Conduct in Human Research 2007 (updated 2018) [p. 16] states that "The guiding principle for researchers is that a person's decision to participate in research is to be voluntary, and based on sufficient information and adequate understanding of both the proposed research and the implications of participation in it." Obtaining consent requires strong articulation of the value of research, which in turn can help reduce misreporting of information at the time of collection and reduce bias in statistical and economic analyses of societal importance.

3. Do you think the Data Sharing and Release legislative framework will achieve more streamlined and safer data sharing?

The Data Sharing Principles (that are based on Five Safes) is a governance framework. It is important that any use of the framework have requirements for risk assessment, and that concrete trust models be established.

The distinction between sharing and release is a positive step and will create an effective way to manage the complex data landscape. We believe that it is a positive step and has the potential to achieve the objectives, especially if the constructive steps identified above in 1 and 2 are taken on board.

If data sharing agreements are mandatory for all data releases under the legislation (Section 5.3), it will be important to provide templates of such agreements. Without clear guidance from the Office of the National Data Commissioner, such a requirement will present a barrier between the public sector and users if those users do not have sufficient negotiation capacity or legal expertise. Data custodian agencies will be able to adapt these templates to their own purposes, providing agency-specific templates rather than negotiating each agreement from scratch.

4. What do you think about the name, Data Sharing and Release Act?

The proposed legislation appears to relate specifically to sharing and release of publicly held data. The name should reflect this: i.e., "Public Data Sharing and Release Act". This would distinguish the sharing and release of public data from data from private, not for profit and social organisations.

It should be noted that the DS&R would be useful as an exemplar to non-government organisations with respect to sharing of their data, and the templates mentioned above would serve as models for such organisations.

5. Do the purposes for sharing data meet your expectations? What about precluded purposes?

Data sets that promote perceptions of bias based on gender, religion, race, opinion, and ethnicity may warrant an exceptional handling to ensure additional safeguards in place.

While the focus of the discussion paper is on administrative data sets, the inclusion of "research programs funded by the Australian government" in the definition of public sector data raises additional issues of particular interest to the academic community:

- Scientific data collected by Australian public agencies (such as the Department of the Environment and Energy, the Department of Health, Geoscience Australia, the Bureau of Meteorology, and research agencies such as the Australian Nuclear Science and Technology Organisation, the Commonwealth Scientific and Industrial Research Organisation and the Australian Institute of Marine Sciences) should be in scope for DS&R. In particular, data in the climate, environmental and geoscience areas should be available for researchers in raw format with minimal manipulationor conversion as well as high resolution data products to facilitate their broad use in scientific research.
- Government data should also be available in open, non-proprietary standards and formats with clear definition of licensing and access constraints to promote accessibility, interoperability and reusability (see the FAIR principles, Wilkinson et al., 2016[1]). This is a matter of good research practice, equity of access, research quality and value for research funding, as well as of compliance with Government policies and strategies such as the *Archives Act (1983)*[2], and the National Archives of Australia Digital Continuity 2020 Policy,[3] which 'seeks to support efficiency, innovation, interoperability, information re-use and accountability'.
- As elucidated in the FAIR principles, government funded research data and metadata should be available both human readable and machine actionable formats at the level of the individual dataset. Allowing this data to be more readily accessed and reused will promote innovation and foster Australian research into areas of public policy interest.
- The Australian Government has invested in significant computational infrastructure for the research community, including two supercomputers. Large volume Earth Systems and environmental research datasets housed in government research agencies are ideal for these new infrastructures, and enable data analysis at scales and resolutions not hitherto possible, using the latest techniques in machine learning, deep learning and artificial intelligence. However, many of these government reference datasets are currently difficult for the research community to access in ways that can be utilised by these high powered infrastructures due to formats, bandwidths, storage capacity and in some cases, reluctance of government agencies to allow access to the rawer forms of the data.
- By default, full compliance by government agencies with many government policies and strategies – such as the *Archives Act (1983)*, the Digital Continuity 2020 policy, and the *Freedom of Information Act (1982)* – makes government data available in

[1] M. D. Wilkinson *et al*. (2016). "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data* **3**, Article number: 160018

[2] See: https://www.legislation.gov.au/Details/C2019C00005

[3] Policy available at: http://www.naa.gov.au/information-management/digital-transition-and-digital-continuity/index.aspx

formats suitable for greatest uptake in the research community, with little additional work required. Compliance with these instruments should be enforced across government agencies.

- Research data created through programs funded by the Australian Government via the research councils (NHMRC and ARC) and other grant funding mechanisms is theoretically in scope of the DS&R mechanisms, but such data is produced at a remove from government, and is not necessarily suitable for the provisions of the proposed Data Sharing and Release Act. The Australian research community is best served through open and accessible research data, but the Office of the National Data Commissioner (ONDC) should consider the interactions of the DS&R with publicly funded (as opposed to publicly created) research and produce clear guidance for non-government, but government funded, researchers in conjunction with the research councils and other research funding bodies. Data derived from NHMRC and ARC research should be made accessible in an appropriate form for further research, verification, validation and progress of science, but not necessarily be made available for commercial work without proper safeguards.

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?

The DS&R should empower innovation and therefore the commercial uses should be allowed. It is counter-productive to limit commercial use of data, as it denies opportunities for independent innovative solutions to be developed for pressing public issues. Commercial agencies can often increase uptake of government data, because they can afford to tailor third party-products to suit individual client needs in ways that government agencies cannot. Legislation should allow the possibility of such access instead of only limiting such access to open data alone. However, commercial uses require appropriate safeguards. A standard of simply "common good" is insufficient, as it is very difficult to test. Overcoming this requires obtaining consent for primary and secondary data use in the first instance.

A more focused scope for DS&R legislation would not only help put in place more effective governance and technical safeguards, it would also help the ONDC articulate expected value that may be derived from DS&R, and build and maintain public trust towards appropriately regulated data sharing and release. Through obtaining consent for sharing or release, and secondary uses in general, the value of the use is articulated, and public views of the benefits of analyses on public data have the opportunity to improve over time. We recommend the ONDC commission a survey into past benefits of shared/open data and analysis of how decreasing sampling bias effects downstream economic analysis for concrete policy choices.

Although DS&R should, and can, empower innovation in the commercial space, safeguards addressing the potential misuse of government data must be considered. In the example of Cambridge Analytica – where Facebook shared data with an academic who then shared it with the political consulting firm – it cannot be assumed that commercial use of government data leads to public good.

7. Do you think the Data Sharing Principles acknowledge and treat risks appropriately? When could they fall short?

The Data Sharing Principles (derived from the Five Safes) are the key mechanism in place for safeguarding security and privacy in DS&R. For example, they are the primary mechanism for determining whether *"risks of sharing [can] be managed"* in the DS&R process for sharing data about the public (see Fig 4 on page 16). The Data Sharing Principles are a governance framework that by themselves do not prescribe how risk is measured, mitigated or compared. While the reframing of Five Safes around risk is more accurate, the framework remains largely unchanged.

Technical safeguards such as cryptography, secure multi-party computation and differential privacy have undergone extensive scientific peer review. The Australian research and business sectors have significant capabilities to draw on and develop into world leading capacity in these areas. Cryptography is prevalent and protects against untrusted storage media, communication channels, and computing platforms; differential privacy protects against untrusted recipients of released data and will be deployed by the U.S. Census Bureau across all outputs of the 2020 Decennial Census[14]. We recommend the ONDC adopt rigorous safeguards that can be verified and quantified, and that protect privacy and security within well-defined trust models. Such safeguards can sit within the Data Sharing Principles, for example by making concrete which data and outputs are considered less risky to share or release, in *certain* circumstances. These technologies are no panacea but best represent *"international best practice to safely share data".* While best governed by legal and policy instruments, data privacy & security are much like medical science, in that they have a firm grounding in scientific disciplines and cannot be achieved through law alone.

8. Is the Best Practice Guide to Applying Data Sharing Principles helpful? Are there areas where the guidance could be improved?

The purpose test as described in the best practice guide does not adequately address the fitness for purpose of the data being shared. Without knowledge of the data quality profile of the datasets being shared, or at the very least an agreed level of metadata, there is a risk of users investing in lengthy processes to obtain approvals (see fig 4 of the discussion paper) only to find out that the quality of data is not fit for their intended purpose. Additional guidelines or stipulations on the acceptable level of accessible metadata and data quality profiles may be needed to mitigate the risk of unforeseen data quality problems.

For the same reason, there should be a mechanism for explorative access of limited datasets to establish the requisite quality of data, and to determine the capacity of data set to help answer the questions.

9. Do the safeguards address key privacy risks?

In light of recent scientific expert study of privacy protections in public data releases, DS&R should recognise that deidentification of unit record-level data is not possible.[5] The primary approaches to de-identify data are via aggregation, swaps, or other ad-hoc perturbations. All known approaches have been successfully attacked, which include re-identifications

[4] Logan Kugler (2019). "Protecting the 2020 Census." *Communications of the ACM*, **62**(7), pp17-19.

[5] Office of the Victorian Data Commissioner (2018). Protecting unit-level record personal information: The limitations of de-identification and the implications for the Privacy and Data Protection Act 2014. See: https://ovic.vic.gov.au/resource/protecting-unit-record-level-personal-information/

(identifying individuals in datasets) and reconstructions (making coarse attributes more granular or recovering suppressed attributes). The methods used in de-identification do not guarantee any specific security property. Differential privacy (cited in Q7 response) is the only mathematical framework for provably guaranteeing strong security properties when releasing data to potentially untrusted recipients. It was developed in light of the impossibility results on de-identifiability.[6]

10. Are the core principles guiding the development of accreditation criteria comprehensive? How else could we improve and make them fit for the future?

Certain elements of 'working with data' should be expected to be covered in tertiary and TAFE qualifications as well as a basic exposure to digital ethics.[7] In particular, individual accreditation may be linked to professional bodies with certain modules to ensure that there are no effective barriers to this.

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?

Breach or loss of data to unauthorised users or access should be reported to the National Data Commissioner even if this does not amount to personal data. This could be approached in a structured way with different reporting arrangements depending on the severity, size and impact of the breach.

14. What types of guidance and ongoing support from the National Data Commissioner will provide assurance and enable safe sharing of data?

We recommend that where ONDC deliberates on cases for potential data sharing and release, that consideration be given to whether modern survey sampling and accurate statistical estimation can be used in place of collecting more data, with more accurate analysis leading to more cost-effective and calibrated policy decisions from existing data sets. Australian universities, the CSIRO and the ABS are home to many internationally-recognised statisticians and data scientists well versed in such techniques.

"While consent is important in certain situations, the societal outcomes of fair and unbiased government policy, research and programs can outweigh the benefits of consent, provided privacy is protected." Consent cannot be overridden by privacy protections. The Australian research community, through institutional ethics review committees that are freely formed and governed by universities and research institutes, demand that explicit consent be freely given to data collection, analysis and sharing for pre-specified uses. Here, 'freely given' means that no information is withheld from the participant until they sign a data sharing waiver, and 'explicit consent' means that data subjects are shown a clear request for consent that describes the primary use of collected data, and that the data subject is truly consenting. The explicit consent from data subjects to how and what their data are used for

[6] Irit Dinur and Kobbi Nissim (2003). "Revealing information while preserving privacy." Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. ACM, pp. 202-210.

[7] Luciano Floridi and Mariarosaria Taddeo (2016). "What is data ethics?" *Philosophical Transactions of the Royal Society A*, **374**(2083), doi: https://doi.org/10.1098/rsta.2016.0360

must be freely given. We recommend that a default baseline level of consent be adopted by the DS&R upon which specific exemptions may be considered.

We see the National Data Commissioner as a proactive champion for data sharing and release and as the technologies develop and improve, the Commissioner will be able to leverage the National Data Advisory Council to re-evaluate its guiding principles. The Commissioner should engage in awareness raising campaigns with public agencies, as well as seek input from the research sector and data user community on potential roadblocks to accessing data under the new framework. The Commissioner should especially consider the effectiveness of safeguards, the causes and repercussions of breaches, and development of improvements to the principles of data sharing on a continual improvement basis.

To discuss or clarify any aspect of this submission, please contact Dr Stuart Barrow at the Australian Academy of Science (stuart.barrow@science.org.au; 02 6201 9464) or Dr Fern Beavis at the Australian Academy of Technology and Engineering (fern.beavis@atse.org.au; 03 9864 0900).