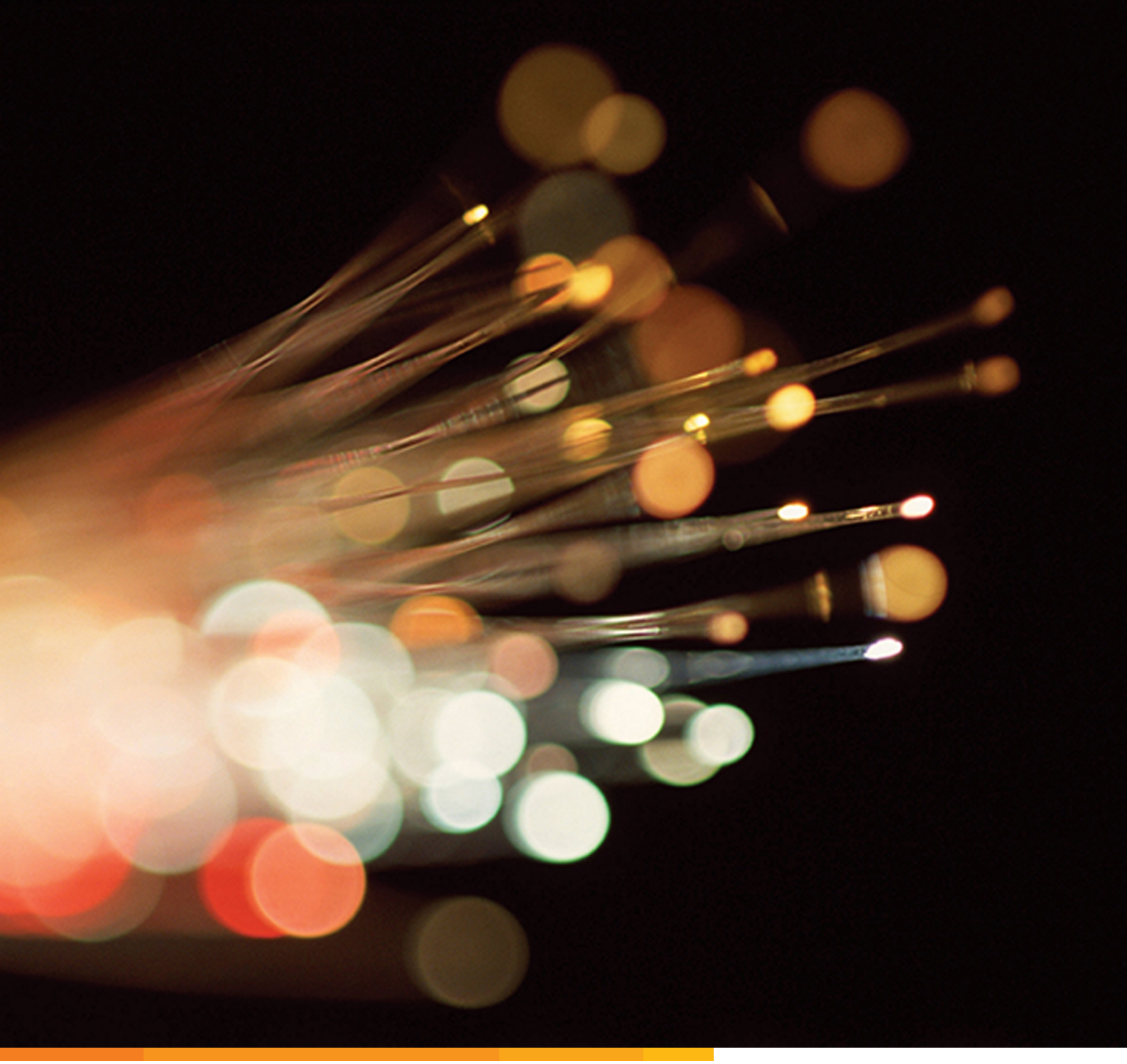

NEW AUSTRALIAN GOVERNMENT DATA SHARING AND RELEASE LEGISLATION ISSUES PAPER – AARNET SUBMISSION

Prepared by Dr Frankie Stevens
AARNet eResearch Engagement Strategist

Contact: frankie.stevens@aarnet.edu.au

EXECUTIVE SUMMARY

AARNet welcomes new data sharing and release arrangements to ensure a growing economy, improved service delivery and transformed policy outcomes for all Australians. As advocates for the Australian Research and Education community, and providers of essential research enabling infrastructure, AARNet values being able to provide input on the proposed Data Sharing and Release bill. AARNet notes that research is increasingly data and technology intensive and the demand for access to government data across all research disciplines is high. Establishing principles for the approach to enabling access to government data - and the means for operationalising this - both need to be considered. The removal of legislative and procedural barriers both aid in unlocking government data.

The suggested use of the five-safes framework by which to determine appropriate data sharing and release is supported. The Australian research community has processes in place to determine safe people, projects, settings, data and outputs, mostly available through the competitive research funding landscape, and the checks and balances established within this, through institutional ethics procedures. It is recommended that these existing research management processes are leveraged to determine appropriateness of data sharing and release to the research sector, such that there is lessened burden for this community to provide this to government anew.

The research sector has research and data infrastructure specialists (for example, the sensitive data linkage experts, PHRN) and service providers (AARNet and AAF) with the expertise to technically broker and enable safe sharing and storage of research data amongst trusted entities. Technical requirements should feature more prominently in any proposed legislative change, and the operationalising of data sharing and release, where that data is exchanged between government and the research sector.

Finally, AARNet supports the cultural change required for government data to be better valued and repurposed. This cultural change is already underway in the research community, and existing expertise and technical advancements in the research sector can assist with this change in government. The cultural change necessitated for successful data reuse is perhaps the most critical piece in the puzzle, and needs to be further developed to ensure success of the proposed approach. The incentives and priorities set through new legislation, and the establishment of managed services on top of connected cyber-infrastructures to support new procedural conventions, need to serve the digital transformation of government and research.

AARNET BACKGROUND

AARNet[1] was established in 1989 by the (then) Australian Vice-Chancellors Committee and CSIRO to reate Australia's National Research and Education Network (NREN). Like other NREN's around the world, AARNet interconnects its members' institutions nationally, and provides international connectivity via other NRENs to the global research and education community. AARNet brought the Internet to Australia and pioneered the use of network data technologies and applications.

Today, AARNet operates as a not-for-profit company limited by shares and owned by the universities and CSIRO. AARNet provides ultra-high speed, very high-quality network and networked services to Australian research and education organisations. As part of its networked service, AARNet provides research data storage, encrypted movement, sharing and access services that leverage the high-quality network to over 55,000 members of the academic community across more than 350 research domains. For example, AARNet video conferencing software provided 244,000 online collaborative meeting environments for the research and education sector in 2017.

AARNet has over 270 directly connected member and customer institutions. Members are the Australian universities and CSIRO, and customers include scientific research organisations, numerous TAFE's and training organisations including health training, more than 700 primary and high schools, and a variety of galleries, libraries, archives and museums (GLAMs). AARNet's operational costs are covered through a membership subscription model.

Australia's NREN, AARNet, is a national asset and is a critical enabling factor in facilitating Australia's research and education excellence. In this submission, potential data sharing and release capability gaps are identified as summarised above and further detailed in the remainder of the response.

KEY PRINCIPLES OF DATA SHARING AND RELEASE BILL

1. *Are these the correct factors to taken into account and to guide the legislative development?*
2. *What else should the Government take into consideration when designing the legislation?*

The key principles listed will be very useful in guiding the legislative development, as they cover aspects such as data safety, data system integrity, data sharing and trust in data. However, it is noted that whilst enhancing the quality of the data system is mentioned, the quality of the actual data itself is not considered, which could have repercussions on building the trust in the use of public sector data. A suggestion here would be to ensure data quality before sharing and release. Whilst quality controls are very briefly proposed later in the issues paper as optional advanced data services provided by Accredited Data Authorities (ADAs), the data wrangling would be more efficiently performed by the data custodian, whose intricate familiarity with the data makes them more qualified to carry out the task than an ADA, who would have to gain an understanding of the data first. For data to be reusable, visibility of its provenance, to engender trust in the data, and detailed metadata records to enable understanding of the data must accompany the data itself. Responsibilities around documenting the provenance and metadata of a data asset arguably lie with the data custodian, or a similar role that was involved in compilation/generation of the data.

Another aspect that potentially impacts on a number of the overall aims for the DS&R bill is that there are "community expectations that we handle data safely and appropriately". The suggested processes described in the Issues Paper cease once a decision has been made with respect to whether the data should, or shouldn't be shared/released.

[1] https://www.aarnet.edu.au

However, there are subsequent actions on the data that could impact on those expectations - for example, a lost usb stick in the mail does not engender trust in the integrity of the data exchange system. At the very least, reference to best practice in the downstream sharing technologies, and access and transfer controls around data, should be considered. The chain of custody as data is shared and transferred needs to be undertaken more systematically and encompass controls that support open and sensitive data transfers at scale (big data requests and many requests for data). Ideally, expert input should be sought to ensure sharing can be carried out safely and efficiently post sharing/release decision, and to unpick when the responsibility for data safety shifts from data custodian to data user.

There is increasing support both nationally and internationally for the FAIR data [2]principles with respect to research data - acknowledging that data must Findable, Accessible, Interoperable and Reusable in order to be effectively repurposed. To be findable, both metadata and data should be easy to find by both humans and computers. Machine-readable metadata are essential for automatic discovery of data assets, which would be required to more quickly progress the cultural change necessitated for data reuse. Such metadata should comply with international standards such as Dublin Core, and be harvested and held in a recognised repository, so that it itself can be found easily (in the academic research data field, there is now a network of such metadata repositories worldwide, which complies with the OASIS guidelines). Accessibility of the data should also be incorporated into the share and release process - where data has been deemed open by default, this should be accessible openly online, and discoverable in the same way. For data to be interoperable, it must be capable of interoperating with applications or workflows for analysis, storage, and processing. If data is to be genuinely open and reused, then commitments to data standards and technology (DOIs, APIs, etc,...) must be incorporated into the process. It is therefore strongly recommended that government (open) data consider the relevance of FAIR data principles on their own data assets, and this be reflected in the DS&R bill.

SCOPE OF THE DATA SHARING AND RELEASE LEGISLATION

3. *Should the scope be broader or narrower?*
4. *Are there entities that should be included or excluded from scope? How would this be justified?*
5. *Should any specific categories of data be specifically out of scope? How would this be justified?*
6. Should exemptions, for example for national security and law enforcement, occur at the organisational level or for specific data categories?
7. *Are there instances where existing secrecy provisions should prevail?*

The Issues Paper mentions that presently ”a dense web of legislative requirements which lack consistency” inhibits data sharing, and may be further complicated by entanglements with legislation that does not anticipate changes in research requirements for data (e.g. big and linked data), the development of new technologies (e.g. encryption), and social norms (e.g. web and social media data). Adding a new piece of legislation that might bypass some of these existing restrictions doesn’t automatically provide a clear path to effective data sharing. For there to be the required clarity for confident data sharing, conflicting legislation would benefit from being reviewed, and where appropriate updated accordingly. The removal of conflicting legislation would be more effective in encouraging a more open approach to data sharing than introducing a new piece of legislation alone, and whilst this might be daunting, and not currently within the direct scope for the DS&R bill, the success of the bill would benefit from this activity.

Another aspect that arose under the scope banner was in relation to the statement that the bill would not, by default, compel all data to be shared. The absence of this “stick” necessitates that the “carrot” for making data available is compelling, otherwise investment in the process might be limited. If a default shared data option cannot be considered,

[2] https://www.go-fair.org/fair-principles/

provisions will therefore need to be made to ensure the priorities and benefits are widely understood, so that sharing occurs in the absence of requirement. An incentive scheme to make departments want to make data more accessible would be beneficial here. Some incentives are already there, but need to be better communicated to existing data holders - the Australian Government has already acknowledged that the country depends on science and research to increase productivity, achieve sustainable economic growth, create jobs, and improve national well-being.

The research sector requires data to solve national research problems and generate critical insights that inform government policy and enhance societal and cultural wellbeing. Therefore, where possible, data should be made available to the Australian research community as readily as possible.

The NDC is stated to drive a cultural change across the public service towards greater and appropriate use and reuse of data with consistent risk management. The cultural change would most effectively be influenced by all commonwealth entities being involved in the proposed data sharing approaches. As such, opinion is towards national security and law enforcement as participating in suggested practices, and specific data categories being appropriate for exemption. If industry standards are to be adhered to, or established as part of the digital transformation of government, to support secure data sharing and transfer, via connecting cyber-infrastructures between government and research organisations, there are legislative and financial considerations.

STREAMLINING DATA SHARING AND RELEASE

8. *Do you agree with the stated purposes for sharing data?*
9. *Are there any gaps in the purpose test that would limit the benefits of public sector data use and reuse?*
10. *What further detail could be included in the purpose test?*
11. *Should data be shared for other purposes? If so, what are those purposes?*
12. *Should there be scope to share data for broader, system-wide purposes?*
13. *Should the purpose test allow the sharing of data to administer or enforce compliance requirements?*

The inclusion and authorisation of "research and development with clear and direct public benefits" as a driver for government data release as listed in the purpose test is welcomed. The specification that there must be clear and direct benefits to the public has the potential to restrict much innovative data reuse in this space however, where partnerships with industry arise in research. In many instances, the longer term impacts of research, and end beneficiaries, can be vastly different than the research originally conceived, and yet are still very valuable to society. Stifling the research reuse of data with this constraint could inhibit potential innovation arising from repurposing of data. An option here could be to provide data to research that has been demonstrated to be a recipient of a competitive funding outcome, which should provide sufficient confidence that the research purpose is sound. Consideration should also be made to ensuring industry users are enabled access to maximise the reuse outcomes of shared data, and specific mention of the stakeholders should be included if deemed appropriate.

Whilst specific feedback was not requested on the proposed process workflow, there is opportunity for enhancements here. Currently, the process begins with a request for data access - this suggests an existing awareness of the presence of the data, which could be lacking without a discovery service. In order for innovation to occur from shared data, there must be concerted effort to ensure it is findable, as academia has done with their "Research Data Australia[3]" discovery service. Whilst some consideration has gone towards the discoverability aspects in making the data sharing agreements

[3] https://researchdata.ands.org.au

publicly available, more needs to be done to publicise data assets if the benefits of innovative reuse are to be realised.

The process also has potential for some simplification, in due course. At present, the workflow depicts the first decision point as determining whether data should be open by default, as per the Public Data Policy Statement. Presumably, as more and more data assets become publicly available, fewer and fewer access requests will result in this "Open by Default" release. However, it is arguable that by combining processes for default open data with those presently suggested for the considered sharing and release approaches, open data might only being released if a request for this is received. Government default open data should be being made available publicly, irrespective of any serendipitous discovery and request for this, to harness the value of the data.

DATA SAFEGUARDS

14. Is the Five-Safes framework the appropriate mechanism to ensure data is safeguarded?
15. Are there any additional safeguards that should be applied?
16. Are there any instances when the Five-Safes could not be applied?
17. Is the Five-Safes appropriate when data is shared and used for the specific purposes in the purpose test above?
18. How should the responsibility for managing risks be shared in the framework?
19. How would you envisage Five-Safes principles be applied over the life-cycle of data to ensure data safeguards are continually met?
20. Under what circumstances should trusted users be able to access sensitive data?

The Five-Safes framework has been demonstrated to be appropriate in helping make decisions about making effective use of data which is confidential or sensitive. The Five-Safes unpicks the decisions surrounding data access and release, which would simplify the sharing and release process. The suggested use of this framework is supported.

Trust in the end user could be achieved through understanding of the proposed reuse purpose, described in a plan for data reuse and on-sharing, which includes provisions as to how the data will be managed safely. Trust in research end users could be given by default, should the research applicant have been a recipient of a competitive research funding process where the funding body has been demonstrated to be committed to high standards of merit and integrity in the research it supports. Given that funding bodies collect vast amounts of data on both the researcher and the research they propose to carry out, it should be feasible for this data to be reused in determining the status of safe (research) people and safe (research) projects. This approach would considerably facilitate and streamline research reuse of government data.

PUBLIC SECTOR DATA SHARING ARRANGEMENTS

21. *Would this arrangement overcome existing barriers to data sharing and release?*
22. *Would streamlined and template agreements improve the process?*
23. *Do you agree that data sharing agreements should be made public by default?*
24. *What level of detail should be published?*
25. *What else should a data sharing agreement contain?*
26. *What other transparency mechanisms could be mandated?*

As suggested, the National Data Commissioner could usefully provide templates to simplify and streamline the data sharing and release agreements. Making these agreements public by default is also welcomed, and as indicated has the potential to aid discovery of data assets - yet this mechanism for discovery alone is insufficient, as previously discussed. It may also be of benefit to release anonymised outcomes of the five-safes assessments, in order to inform those

involved in undertaking these.

It should also be noted that the more legalese is added to any release procedure, the higher the fraction of profit driven corporates there will be who benefit from the data, vs other use, as these have access to additional resources to work through these. As such, any data sharing agreement needs to be simple, such that supplementary resources and excessive time requirements are not warranted. Streamlining of data sharing agreements could be facilitated for the research sector through incorporation of these into the competitive research sector funding agreements.

ROLES & RESPONSIBILITIES WITHIN THE SYSTEM

27. *How long should accreditation as an ADA or Trusted user last?*
28. *What could the criteria for accreditation be?*
29. *Should there be review rights for accreditation?*
30. *Should fees be payable to become accredited?*
31. *Is the Australian Government Charging Framework fit for purpose in this context?*

Approaches that support the cultural change required to maximise open data value should also be rewarded. As such, should a candidate data user also profess to make their resultant data products available under an open license, they might be exempt from any fees proposed in the data release process.

Consideration on how data will be provided from a data custodian to an intermediary ADA is also important, particularly as the ADA is suggested to perhaps perform the de-identification of the data. This step in the workflow is currently risky, unless performed under strict and secure mechanisms, using appropriate infrastructure and technology approaches.

It is understood that the NDC will determine the criteria and processes for accreditation of ADAs. Whilst building on the expertise of integrating authorities is welcomed, the expertise required for the sharing and release of data is broader than this. Trusted Data Authorities should also be consulted here, particularly those certified with the CoreTrustSeal [4] Data Repository certification which demonstrates reliability and durability of data repositories and hence enables sharing data over a long period of time. Other key roles within the system are those with expertise in access and authorisation around data resources, such as the Australian Access Federation [5](AAF), which provides identity brokering services enabling access to online resources for the education and research sector. The AAF is a vital part of the Australian eResearch infrastructure landscape facilitating trusted electronic communications and collaboration between education and research institutions both locally and internationally. AARNet is also a critical enabling factor in facilitating Australia's research and education excellence, particularly with respect to data sharing and collaboration, elements that are key in enabling cultural change towards data reuse. AARNet would be keen to provide expertise on how to efficiently, and safely, transfer and share data between roles within the proposed Data Sharing and Release process.

Any consideration of charges with respect to data access needs to ensure that this is not inhibitory to re-use, and counter to the open data movement that is being nurtured. Should data access fees be deemed appropriate, then fees should be payable on ADA accreditation to offset the access costs. AARNet, as national experts in the provision of research data services for academia, would be willing to provide ADA support for use cases where data.gov moves into data.edu.

[4] https://www.coretrustseal.org
[5] https://aaf.edu.au

NATIONAL DATA COMMISSIONER

32. *Are these the right functions for the National Data Commissioner?*
33. *What review powers should the National Data Commissioner have?*
34. *Should the NDC have the power to conduct an investigation into system-wide issues?*
35. *What other actions could the NDC be able to take?*
36. *Are there other ways community values and expectations can be captured and addressed?*
37. *What aspects should be taken into consideration when considering consequences for non-compliance with the DS&R Bill?*
38. *Should the consequences differ depending on the type of data involved or the type of misuse, e.g. harsher penalties for intentional misuse?*
39. *Should penalties be strict liabilities?*
40. *What would be an appropriate penalty for intentional misuse of data?*
41. *How would responsibility for misuse of data be shared across the data system?*
42. *To what extent should there be a complaints mechanism and how should it work?*
43. *Should a complaints mechanism provide for complaints by the public?*

It is stated that the NDC will "work to empower public sector agencies to share and release data". The issues paper does not address how this critical activity is proposed to be carried out. Without success here, all other efforts could be wasted. It is therefore critical that the NDC have a well thought out engagement and communications strategy over data holders, trusted service providers, the learned academies, universities and research organisations, so that the frameworks that will be made available are leveraged accordingly.

Data custodians and potential data users would benefit from guidance on how to share, manage, move, and securely store data at the technical level, and this should be provided in general terms by the NDC, so that basic data functions for "open by default" cases, and where access is conditional, do not have to rely on an appropriate, or available ADA. Data exchange where possible should be direct and systematic, and barriers to access be removed e.g. time, cost, technology.

The Australian Research Integrity Committee contributes to quality assurance and public confidence regarding the integrity of Australia's research effort. Any DS&R non compliance involving data reuse for research purposes could usefully be funneled into this established mechanism for determining best course of action on breaches of the Australian Code for the Responsible Conduct of Research.