www.csiro.au

CSIRO Submission 19/685

Data Sharing and Release Legislative Reform Discussion Paper

Office of the National Data Commissioner

October 2019

Enquiries should be addressed to:

Brendan Dalton
CSIRO Chief Information Officer
GPO Box 1700 Canberra 2601
[redacted]

Main Submission Author:

Gavin Walker
Principal Experimental Scientist
CSIRO Data61

Introduction

CSIRO welcomes the opportunity to provide input to the Office of the National Data Commission (ONDC) on the Data Sharing and Release Legislation Discussion Paper. Australia's future success is heavily dependent on our national ability to engage and lead in the use of data for research and innovation in all segments of the economy, including in its use to inform industry and government decision-making.

The availability, accessibility and use of data is essential for the delivery of research and innovation services across all parts of the national innovation system. CSIRO has world class capabilities and delivers rigorous and comprehensive research in data and digital innovation, from data analytics, and data network platforms, to collecting, managing, disseminating and using extensive amounts of data collected by CSIRO and other organisations. CSIRO is a major provider of data, data products, analysis and infrastructure, servicing industry, government, the science community and the public.

CSIRO's submission to the ONDC draws on our broad range of scientific expertise in dealing with data across many sectors of activity and with government departments, agencies, research institutions and with the commercial and non-government sectors. CSIRO has been appointed by the Treasurer as the Data Standards Body to support the delivery of the Consumer Data Right. The Data Standards Body is responsible for assisting the Data Standards Chair, Andrew Stevens in the development of common technical standards to allow Australians to access data held about them by businesses and direct its safe transfer to others. The Data Standards Body draws on the expertise of CSIRO's Data61 team.

The Office of the National Data Commissioner (ONDC) has asked for responses to their discussion paper on Data Sharing and Release Legislative Reforms. CSIRO encourages the ONDC to continue the good work that it has already commenced to reduce unnecessary impediments to data sharing while respecting privacy, consent and the public interest. Our response is in two parts, first some comments on CSIRO and data, and then we address the fourteen questions in the discussion paper.

CSIRO welcomes the opportunity to discuss these matters in more depth with the ONDC.

General Comments: CSIRO and Data

CSIRO is a significant generator and user of data in a variety of domains (agriculture, mining, energy, environment, health and biosecurity, astronomy and earth observation). We are experienced with creating, managing and delivering exceptionally high volumes of data. In the astronomy domain, this is particularly evidenced in the successful development of the Australian Square Kilometre Array Pathfinder (ASKAP) radio telescope and for our aspirations to deliver the Square Kilometre Array (SKA) where the big data processing and analytics capability we have developed with our partners in the Pawsey Supercomputer is essential. The data is made available through CSIRO's Data Access Portal (data.csiro.au) in collections such as CASDA (CSIRO ASKAP Science Data Archive) and the Pulsar Data Archive.

CSIRO's Data61 business unit is founded on the premise that the world is undergoing a digital transformation on par with the socioeconomic shift of moving from agriculture to manufacturing. Every industry is becoming data driven and Governments and corporations need to evolve more quickly to avoid disruption. Our goal through Data61 is to accelerate the nation's scaling of capability and connections to tackle the world's biggest problems and accelerate new sunrise industries creating economic growth and jobs for Australia. Dedicated to a dual mission: pursuing new-to-the-world fundamental and applied research; and by working collaboratively with others in the nation's innovation ecosystem, to reimagine and seed new industries in a data-led world.

CSIRO is delivering a new Challenges and Digital Transformation Program. Amongst other benefits CSIRO is seeking to bring together expertise from across disciplines to focus on data network and data analysis and deliver novel digital solutions. Through our planning for the program CSIRO has identified and established

some domain specific projects to leverage the big data and complex data capability for national and global outcomes such as:

- Autonomous biosecurity
- An Artificial Intelligence/Machine Learning future science platform
- Australia's Energy Oracle - providing the ability to scenario plan energy outcomes
- A national clinical trials data network
- Going global with food provenance
- Specimen digitisation, phenomics and Artificial Intelligence

Underpinning the program CSIRO has established a Managed Data Ecosystem initiative. The initiative will provide a platform to enable greater cross disciplinary consistency in data management and governance. It has been established to support, safe and secure access and use of data, implementation of the FAIR principles in CSIRO, creating an environment where data important to CSIRO and our partners is Findable, Accessible, Interoperable and Reusable. A second underpinning initiative is the CSIRO Digital Academy, recognising the importance of a data and digital literate workforce and aiming to provide staff the opportunity to add to their digital skill sets to take advantage of the opportunities that data reuse presents. Both initiatives are well underway and seen as foundation work to enable CSIRO to continue to excel and undergo a digitally based transformation necessary in the 21st Century.

CSIRO would welcome closer collaboration with the ONDC as it progresses these initiatives and is keen to pilot projects of mutual interest and broader ecosystem benefit.

Response to specific questions in the Data Sharing and Release Legislative Reforms discussion paper

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

The distinction proposed between data sharing, reaching an agreement between parties on how the data sharing principles apply, and data release where data is placed in the public domain with limited further controls, was clear. However, it was less clear how an agency determines what data is suitable for sharing and what data is suitable for release. CSIRO, and in our experience other agencies and institutions we work with, find the decision process difficult. CSIRO as part of its CSIRO Managed Data Ecosystem initiative (mentioned above) will be working through its own methodology for the release of data and is keen to work with other agencies and ONDC to develop easy to follow guidelines and principles for the Research sector. CSIRO looks forward to any guidance the ONDC might develop in this area, especially with released data being subject to data breach legislation.

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

CSIRO has found that one of the impediments to releasing, otherwise safe, data is the risk of inappropriate use of that data. The data sharing principles can be applied to assist with this. For example, a data custodian can release data and provide guidance on how it can be safely used. Good data description is required to support release decisions. ONDC could assist with more guidance on description necessary to support release and make release safer.

The Productivity Commission's *Data Availability and Use* inquiry (Productivity Commission, 2017) found that expertise on preparation of data for release what not widespread in the public sector. Smaller agencies especially "have limited capacity to invest in data management" (Productivity Commission, 2017). Accredited Release Authorities (now Accredited Data Service Providers) were proposed by the Productivity Commission to provide that support and develop expertise in the data alongside the data custodian. This level of support which is also recommended by CSIRO seems to be missing from the ONDC's proposed Accredited Data Service Provider (ADSP) model, although CSIRO anticipates that these support

requirements could be developed in the associated working group activity. Perhaps there is an additional role encompassing data release assessment and data quality improvement. This role could belong to an ADSP, a data custodian or a trusted third party. Such a role should be backed up with detailed, authoritative guidance from the ONDC or an appropriate body.

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

CSIRO commends the ONDC for seeking to publish a register of agreements and participants. Publishing successful agreements will make it easier for participants to create future agreements. If a machine-readable form of agreements were available, value added agreement services could be developed. For example, in assisting users in choosing agreements best suited to their task. CSIRO encourages the ONDC to also consider what it can do to encourage data custodians to publish the existence of data and any timeline or requests for the sharing or release of that data. As part of registering ADSPs the ONDC could publish the data sharing and release services the provider offers such as data deidentification or data integration.

The ONDC's proposed framework appears to involve data sharing agreements between:

1. data custodians and accredited data users
2. data custodians and accredited data service providers
3. data service providers and accredited data users but with the data custodian(s) as part of the negotiation.

The Productivity Commission's *Data Availability and Use* inquiry (Productivity Commission, 2017) found a major bottle neck in health data sharing was the need to negotiate agreements with multiple data custodians. The arrangement proposed by the ONDC and described above, does not appear to make it any easier.

CSIRO suggests ONDC consider two mechanisms to improve the ease of negotiating data sharing agreements: broader or templated data sharing agreements, and empowered ADSPs.

Broad or templated agreements

Previously researchers needed to get approval to use data from custodian and their own ~~and a~~ custodian endorsed ethics committee (Productivity Commission, 2017). Data sharing agreements could provide access to data for a broad range of purposes. For example, an agreement could require:

- a certain level of accreditation for the data user (perhaps with some additional constraints)
- data being used in a range of pre specified projects (such as the discovery of improved disease treatment)
- data can only be processed in a specific environment (such as the Sax Institute's Secure Unified Research Environment (SURE) or an equivalent level of control)
- the results encompass a broad range of outcomes (data aggregated to the Australian Bureau of Statistics Statistical Area 4 (SA4) or above).

These premade agreements based on levels in the data sharing principles, would allow the agreements to be used off the shelf. Researchers would still require ethics approval and non-researchers would still need to pass a public interest test.

Empowered ADSPs

ADSPs can improve data sharing if the data custodian gives the ADSP a well described remit for sharing data under certain conditions. Use of ADSP can reduce the number of parties the data user needs to engage with in order to be part of the data sharing. Such data sharing would not necessarily involve continued authorisation from the data custodian.

This does come with risks as there is never complete communication of the constraints on the data use between the custodian and ADSP. A collaborative and communicative relationship between custodians and the ADSP also needs to exist providing a level of informed governance over data use and release. Data sharing allows for retraction of the data, so a blocking veto power is not required. Data custodians could choose to retain a pre-sharing veto power in their agreements with the ADSP.

Data Integrators

The framework proposed by the ONDC does not seem to support data integrators. For example, if custodian A and custodian C share data with service provider B using separate data sharing agreements. Then if service provider B gets a request to provide data that must be integrated from data from A and C, is B allowed to provide this integrated data set? CSIRO suggests that guidance in the framework on how to include integration allowances or restrictions in data sharing agreements would clarify this.

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

No comment.

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

CSIRO acknowledges that compliance and national security have specifically been precluded in the proposed framework. These purposes are difficult to gain a social licence to operate and, if the data is misused, runs the risk on derailing infrastructure supporting better endorsed sharing purposes. The discussion paper associated "supported by the community" with the purpose test. CSIRO suggests that if this is used to broaden purposes, community support will need to be significant and transparently researched.

If agencies are successful in obtaining independent legislation to support sharing for other purposes, there is a reasonable case that such legislation may also allow the use of data sets packaged under data sharing arrangements to reduce duplication of infrastructure. However, this may also be perceived as breaking public trust in the data sharing system. CSIRO suggests that such a perception should be tested before there is any potential to proceed in that way.

The purposes listed in the discussion paper are very broad. Data sharing could be allowable for those purposes but still not be suitable for sharing for the intended specific project (data principle) because of data quality issues. CSIRO suggests there needs to be a distinction of data sharing purposes and the types of projects which the data custodian sees the data suitable for, so that the data can be deemed fit for purpose. This is the same sort of guidance CSIRO suggested for data release earlier in this submission.

CSIRO suggests that keeping track of the provenance of the data will help the system understand what purposes data is useful for and what purposes data has been applied to. Provenance of analysis output can be a valuable tool to help data custodians audit the use of their data and be confident it is being used appropriately. Because data sharing happens under an agreement this usage information can flow back into the systems to support investment and guard against misuse.

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 data sharing vision in the discussion paper appears to focus solely on government sharing data with itself i.e. across government departments and agencies. However, this vision could be expanded to allow entities to direct government to share the information government holds about them with other non-

government parties in the economy and society. This would provide access to the data but may also increase confidence in the authenticity of the data that is shared. CSIRO suggests that the access and consent rules and mechanisms could be similar to those for the Consumer Data Right (in which CSIRO is playing a key support and development role). Obtaining specific consent in this way could significantly increase the scope of data that could be safely shared. Initial data sets that might be considered for this could include some of the data about business, personal property securities, licenses, and customs.

CSIRO's experience is that commercial use of data between government agencies is complex. Commercial entities can derive great value from public data and are likely to do so from shared data also. Some agencies already share data commercially:

- Government research institutions create data and provide research services commercially from data.
- Government pays commercial parties to provide services on the government's behalf.
- Government procures commercial research services and may reasonably provide data for that purpose.

Many research institutions are conducting commercially funded research. A clear and transparent public interest test is critical in these situations. The public needs to be confident that a commercial interest will not lead to less supported outcomes such as for example payday lending or new treatment advertisements targeted at the individual or community level.

It is CSIRO's view that commercial sector or private sector use of government data could be subject to the same level of accreditation as for public agencies and research organisations. For example this would enable the safety of the project, people and environment to be evaluated and a public interest test or research ethics approval applied. As with other data sharing agreements, those with commercial /private sector parties would be published openly and also rejected requests could be considered to be made publicly available. Public scrutiny helps to maintain community trust.

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

The proposed data sharing principles are an excellent framework for treating risks of data sharing. Each of the five principles can provide a degree for flexibility about what is important in the data sharing agreement. If in the agreement one or more is treated in a restrictive manner, then some of the others might be relaxed a little where appropriate. While the Best Practice Guide to Applying Data Sharing Principles covers that flexibility the examples given in the discussion paper do not. The discussion paper examples assume personal data and apply maximum protect. Those just reading the discussion paper may not see the flexibility on offer.

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

The Best Practice Guide to Applying Data Sharing Principles is helpful. CSIRO commends the description of the different levels of the five principles and how they are applied. Realistic examples using synthetic data, systems and people may further enhance it utility. It may also be helpful to be able to grade or certify data and system to the risk level for each of the principles to give data custodians greater surety as to how to apply the principles.

9. Do the safeguards address key privacy risks?

The discussion paper seems to lean towards not requiring consent for data sharing on the basis that allowing people to opt out will bias data. CSIRO notes that refusal to ask for consent can lead to attempts to sabotage the data by skipping mandatory fields of surveys or injecting false data. It is important to consider that when information is collected without consent for use in research, human research ethics

committees reviewing a project must test the request to use the information against the waiver of consent criteria specified in Chapter 2.3 of the National Statement on Ethical Conduct in Human Research (2007). Where these criteria are not satisfied, researchers must seek consent to use the information and may not be able to proceed with the project if this cannot be done.

CSIRO proposes that the interaction between human research ethics committee review, the sharing of government data for research purposes and consent, should be comprehensively explored as part of the decision-making process around the role of consent in the data sharing system, as failure to do so could undermine the effectiveness of the system to promote the use of shared data for research purposes. For non-research purposes there needs to be a substantive public interest test but that too should be able to require consent be obtained.

Where a person has given consent for the use of their data with constraints, then those constraints should be respected, and data custodians will need to be able to pass on those constraints where necessary. The best place to obtain consent is when the data is collected.

The privacy safeguards will need to be clearly articulated and supported in any policy that is applied to collected data and it will need to be easily understandable to the Australian people. The ONDC could give guidance on where it expects consent will be required. The discussion paper appears to overall increase the privacy obligations on agencies. Perhaps the New Zealand consent model may be informative.

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 good to see accredited users work in tandem with accredited organisations. The organisations will need to provide access to data management infrastructure to adequately safeguard the data, however it is unclear who is responsible in ensuring the infrastructure is up to the task. Different data will need different levels of protection. CSIRO suggests use of a grading system for the protection offered by a data management system to match the levels in the safe environment principle. Developing examples for each level of protection will be useful in helping to develop appropriate accreditation. For example, not all shared data is personal and may not need the rigid protections offered by ABS Data Lab. The ONDC advisory committee may provide insight into such a grading system.

CSIRO suggests that Accredited Data Service Providers (ADSP) will be critical in ensuring the scalability of the data sharing system. The discussion paper expects Integrating Authorities to become ADSPs. These special ADSPs will be experts in the data just as current Integrating Authorities are (alongside the data custodians). One of roles of ADSP envisaged in the Productivity Commission's Data Availability and Use inquiry (Productivity Commission, 2017) is to act as bidirectional links between data custodians and data users. Data users can, and do, improve the quality of the data sets and ADSP have an opportunity to feed higher quality data back to custodians so the overall data set can be improved. Another role would be to hold integrated and value-added data between projects so that the value gained in the project is not lost. Many value-added datasets are destroyed at the end of projects under current arrangements. These roles could all be captured in data sharing agreements, but the Data Sharing and Release Act could make it clear that ADSP are able to perform these roles.

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?

As discussed earlier, transparency is key to maintaining public trust in the data sharing system. Compliance with this transparency should be as simple as possible, which is best achieved by automation. If the

reporting systems are not automated, then the reporting obligations may stifle sharing. A balance must be reached, or the appropriate IT support put in place. Mandatory data breach reporting should be required in line with Privacy Impact Assessment recommendations. Annual reporting on the level of participation in data sharing framework including agencies declining to include data will be provide evidence for the success of the data sharing framework and the support for innovation that the sharing of public sector data is providing.

CSIRO suggests that progress reports on the level of multi-jurisdictional participation (commonwealth, states and local government) and progress toward the creation of National Interest Datasets, become integral elements of the reporting requirements, as these are core parts of the ONDC's mandate. This should be supported with a periodic external review of the data sharing system, to improve transparency.

12. Have we achieved the right balance between complaints, redress options and review rights?

The provision in the framework for the ADSP to remove the need for the data custodian to approve every data share, will make sharing more efficient. CSIRO suggests that regular review and redress by custodians is critical to ensure custodian support and that a process whereby individuals or organisations (accredited or not) are able to raise complaints and concerns would be likely to build trust.

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?

We note that while revocation of accreditation is mentioned in the discussion paper it does not clearly appear in the enforcement pyramid at Figure 7. Being able to revoke user and institution access to data should be considered as it could be a powerful tool. Some level of commercial or personal compensation for deliberate disclosure might also be relevant for consideration. When data is used inappropriately then it should be possible in principle to ask for the deletion of shared data. That is to effectively withdraw the data sharing, and this should also be considered in the framework.

Some data is currently shared because it is perceived to be low risk to do. A significant penalty may heighten total risk and reduce the amount of data shared. However, a balance needs to be struck and for safer data the maximum penalty could be lower.

There is potentially a problem with "gap coverage" where for combined datasets the most severe penalty is used. If two benign data sets, when combined, are highly disclosive then the penalty would still be low. In CSIRO's view it is the risk posed by the combined data set which should be considered to determine the penalty.

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

The Data Sharing and Release Act as proposed will support the work of the ONDC, and CSIRO encourages the ONDC to begin to emphasise more their role in facilitating national data utility. The "scene setting" chapter of the discussion paper eliminated State jurisdiction held data as out of scope and only mentioned national interest data once. However, National Interest Datasets (a key recommendation of the Productivity Commission's Data Availability and Use inquiry) are the catalyst for working with state government, local government and the private sector to obtain data to transform the way Australia operates at all levels. It can empower the use of nationally informed, integrated data by state and local government. It would be potentially of benefit if the Data Sharing and Release Act talked positively about these things even if it cannot yet describe their principles. Likewise, "Improving the way the Australian

Government make public policy" is currently put forward as a minor point, yet there are substantive policy benefits to Australians from data sharing.

It may be useful to contemplate if it is possible for the Data Sharing and Release Act to outline the process for the inclusion of state jurisdictional data and the development of National Interest Datasets at a later date. It would also be useful for the ONDC or the Act to outline a timeline or change process for enhancing the Act and how states might participate in this change.

CSIRO commends the initiative of the ONDC and looks forward to the Data Sharing and Release Act and the subsequent improvements in data sharing and release across Australia.

CSIRO is available to discuss this submission with the ONDC and is keen to consider collaboration opportunities, for example in piloting projects of mutual interest and for broader ecosystem benefit (please see contact details on the cover page).

References

Productivity Commission. (2017). *Data Availability and Use Report No. 82.* Canberra.