SCIENCE AND GOVERNMENT
www.csiro.au

CSIRO response

Data Availability and Use draft report

Productivity Commission

Submission number 16/581

13 December 2016

Gavin Walker (editor)
Paul Box
Simon Cox
David Lemon
Cynthia Love
John Morrissey
Bill Simpson-Young
Bob Williamson

Contents

1 General response

CSIRO has considered the draft report on Data Availability and Use produced by the Productivity Commission in November 2016. It is a well-researched report that promises a major improvement of data flows in the public, private and research sectors if adopted. This section contains a general response, while section two addresses specific recommendations and requests for information.

Ethical challenges

CSIRO acknowledges that the broad area of data availability and use raises numerous significant ethical challenges requiring careful and detailed attention. We do not specify these in this document but, given CSIRO's experience in human research ethics and data use in particular, we are able, if required, to offer assistance in defining these. In particular, before the implementation of any new policies or practices, a thorough assessment should be made of their applicability to, and consequences for, the use of individuals' personal data in research and the primacy of the rights of those individuals in relation to decisions on collection, analysis, storage, sharing, linking, release and destruction of data.

In the research domain, the Australian Code for the Responsible Conduct of Research (https://www.nhmrc.gov.au/guidelines-publications/r39) is the guiding document for decisions in relation to use of personal data for research purposes. It is recommended that extensive consultation occur with NH&MRC, ARC, research organisations and other lead stakeholders to ensure that these issues are fully canvassed and addressed in policy outcomes.

Data origin and intent

A key factor for determining if data can be easily released is the intent behind its creation. There is a significant difference between data created in the course of research where the data was not the primary intended output and data which is the intentional result of a process to create that data. In the latter instance, contractual constraints may limit the ability to release data unless those contracts can be renegotiated.

It is CSIRO's view that, in the first instance, there should be a focus on datasets that arise from publicly funded activities for which the funding was provided with the intent that the data would be made publicly available.

CSIRO believes that it is not feasible at this time to mandate release of *all* research data as this would place significant burdens on agencies, impact upon research efficiency in particular during the first three years of start-up of implementation, and risk not prioritising the most useful datasets.

Authentication

The draft report makes no recommendation for infrastructure around authentication and authorisation. There are at least two classes of trusted users, or at least trusted activities. The less onerous class identified in 9.7 appears to be anyone associated with research. The second classes associated with identifiable data in 9.8 would need to pass ethics approvals. Clear guidelines will be required for what data and what purposes fall into each class and how members are chosen. Once chosen then an authentication and authorisation framework is required to grant access. Authorisation questions are also relevant for personal data rights and how that affects deceased estates.

Process governance

Data is cross cutting. It is important to entities in all levels of government, academia, private sector and the general public. Ideally, bodies with governance responsibilities in this domain should be broadly representative and collegiate in nature, with intrinsic authority and public support from across levels of government and other relevant institutions as required. This approach is more likely to lead to adoption at the national level required to achieve results.

Data custodianship

With clear rules of engagement, the National Data Custodian can support release authorities to consistently assess data for release or sharing. Data does not perish and long lasting curation of especially National Interest Data is required. Machinery of Government (MoG) changes should list the new custodians of National Interest Data sets and ensure that expertise moves with the data. A custodian of last resort is recommended for when an agency is disbanded and curation of the data is no longer of interest to government or the private sector. This is important as some very small agencies hold very significant data such as indigenous data sets.

Economies of scale

Small agencies can often struggle to manage their data. There are economies of scale to be had. Sectors such as galleries, libraries, archives and museums and the health services sector could benefit greatly from economies of scale in making data available. Common types of systems, data models, APIs and risk assessment methodologies would smooth data access and use.

Kick-starting the market

While it is ideal that government acts as a wholesaler of data to private sector for value adding, Australia might not support active private sector use in all data domains. Some stimulation of the market may be of value. Some agencies have fully self-contained commercial entities, such as CSIRO Publishing, which may need to be regarded as private sector parties rather than public sector with regard to their data.

Support for standards development

Standards development is usually a multi-year process and may require close cooperation between government and private sector bodies with appropriate levels of government support.

While shared agreements and standards will make data flow better, we suggest that agencies should not wait for these to be finalized before publishing data. It is preferable to get data out and receive feedback to guide continuous improvement than to hold back until it is ready. Getting the data out enables communities to engage with the diversity of data requirements as a first step towards agreements about data which can in turn lead to the smooth flow of information.

Good data governance within organisations will ensure a pipeline is in place to keep data fresh and improving. It will readily identify what part of best practice can be adopted right away and what requires further systems in place to be sustainable.

Timelines

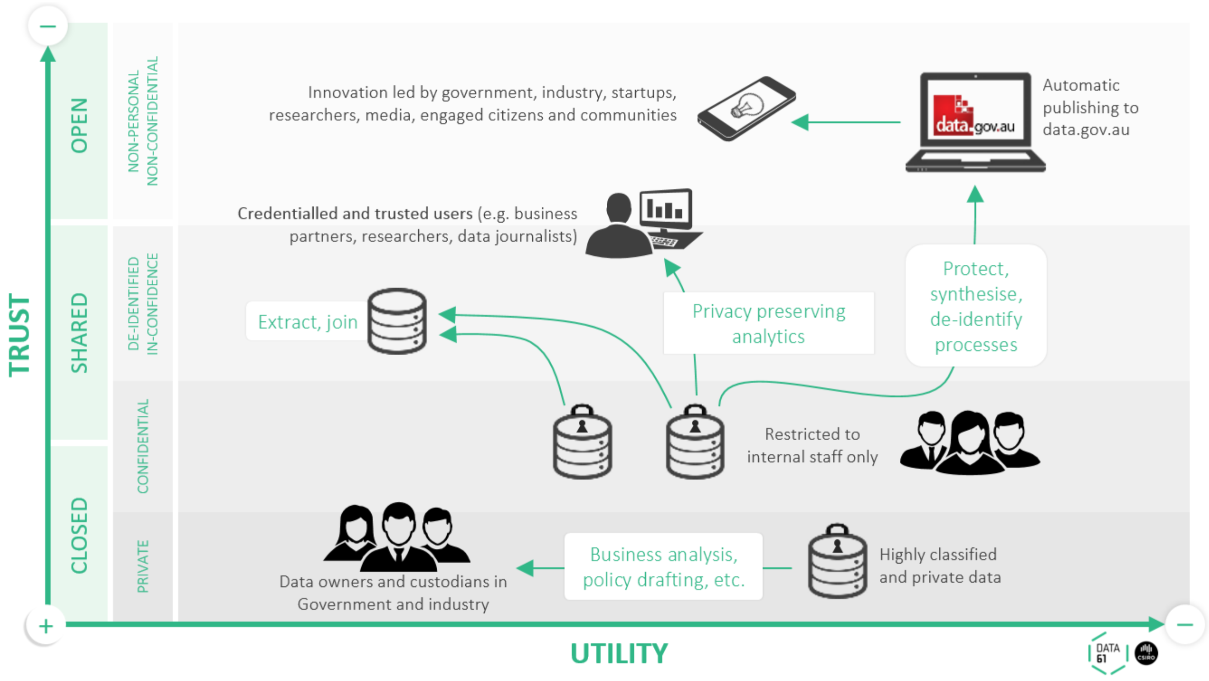
A timeline encompassing all recommendations in the report would be an advantage, with an early focus on high value data release, and sharing and improving metadata content. Publication of metadata is simpler and a listing of the breadth of potential data is usually available well before all the data has been assessed for release.

Data services

Datasets are much more than files transferred between parties. These recommendations should also include possibilities of streaming data, APIs and publish and subscribe models. Data comes in textual forms, binary forms and multimedia.

CSIRO Data61 is developing technologies which use services to access data from a less trusted data layer by applying a dynamic de-identification technique. These technologies (Fig. 1) are not about files but about service level access. Thinking outside the file paradigm opens up more possibilities.

Figure 1 Access across trust layers



TRUST
OPEN
SHARED
CLOSED
NON-PERSONAL NON-CONFIDENTIAL
DE-IDENTIFIED IN-CONFIDENCE
CONFIDENTIAL
PRIVATE
Innovation led by government, industry, startups, researchers, media, engaged citizens and communities
data.gov.au
Automatic publishing to data.gov.au
Credentialled and trusted users (e.g. business partners, researchers, data journalists)
Extract, join
Privacy preserving analytics
Protect, synthesise, de-identify processes
Restricted to internal staff only
Data owners and custodians in Government and industry
Business analysis, policy drafting, etc.
Highly classified and private data
UTILITY
DATA 61
CSIRO

Costs

Whilst the recommendations in this report are positive, it needs to be recognised that there will be a potentially significant cost to agencies, entities and governments that collect data, over and above the cost of establishing the standards. For example verification of de-identification may require specific analysis of reverse engineering. Whilst these costs are inevitable, they need to be recognised as real, because it may lead to agencies etc. not collecting data because of cost.

CSIRO

CSIRO has a track record of helping the public and private sectors make data more available and useable. CSIRO welcomes the opportunity to assist in enabling a new world of fluid data availability and use.

2 Responses to specific recommendations

DRAFT RECOMMENDATION 3.1

> All Australian Government agencies should create comprehensive, easy to access data registers (listing both data that is available and that which is not) by 1 October 2017 and publish these registers on data.gov.au.
>
> States and territories should create an equivalent model where one does not exist and in all cases should make registers comprehensive. These should in turn be linked to data.gov.au.
>
> The central agencies responsible for data should:
>
> - set measurable objectives, consistent with best practice, for ensuring that available data and metadata are catalogued and searchable, in a machine-readable format
> - improve accessibility of data for potential data users.
>
> Limited exceptions for high sensitivity datasets should apply. Where they do, a notice indicating certain unspecified datasets that have been assessed as Not Available should be published by the responsible department of state, on the relevant registry.

It is to be commended that the PC wishes to promote the creation of comprehensive registers of public data holdings. This is consistent with the National Archives of Australia's (NAA) Continuity 2020 program which calls, in principle 3, for business processes to meet minimum metadata standards. NAA’s targets are less ambitious. Alignment of similar activities in the Australian Government seems a positive approach.

However, most agencies will face a very large challenge to undertake due diligence on the origins of data and any restrictions that were put in place at the time of its creation. These checks will need to be particularly thorough in the case of private personal data. It is CSIRO’s position that there will, in some cases, be legitimate reasons for retaining or delaying dataset release including in order to meet CSIRO's other policy mandates and obligations in respect to privacy, ethics, contractual arrangements, and commercialisation and IP requirements.

It is therefore highly likely that the proposed deadline will not be (even close to) achievable for many agencies, including CSIRO. Timelines and deadlines will need to be set that recognise the high level of variability in data types, origins and contractual status, and the attendant complexities in unravelling these.

The minimum metadata requirements set by the NAA focus on archival metadata (identity, authorship and creation date, format, disposal class, security, rights to use and integrity checks). Comprehensive metadata is much richer, making search more effective and helping the likely user base interpret the data correctly. The NAA can accept this metadata but it is not compulsory at this stage.

It is unlikely that all agencies have the skills to discover and adopt metadata best practice. There may be sufficient skills in the Australian information community which could be harnessed through an appropriately supported program. Metadata creation is best automated, being built into work practices and systems. Doing so will make metadata more consistent, reduce the burden of creation and assist with the management of corporate knowledge. There is however an initial cost in understanding, articulating and setting up support systems.

These systems need to be supported by a data governance framework within an agency. Data governance determines form and content of data and metadata. More importantly data governance engages with data users inside and outside an agency and with an agency's data community to continuously improve the data listing.

The defining characteristics of the registers need to be determined. For example does a register hold the metadata for single agency, or are the registers aligned with a domain - such as a register of water related data. Registers may overlap in content, but the metadata record only needs to appear once with other registers referencing the same, complete metadata record.

Some metadata can breach confidentiality. For example providing the upper bound of profit for businesses in Australia may disclose the profit of a specific business that is well known to have record profits. A decision to publish comprehensive metadata needs to assess the disclosure risk.

The characteristics of a register will disclose a limited amount of information. For example Not Available in a water data register will disclose there is an undisclosed data set about water.

DRAFT 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.

This recommendation effectively puts public funded entities, and the entities they in turn fund, under the same requirements as public sector agencies in recommendation 3.1 above with additional justification for why the data is not available. It may be that some public funded research groups will not have the capacity to comprehensively list data sets used or created. Similarly they may not have the data governance capacity to identify best practice. In these cases pubic funding agencies may need to ensure the funded party is given sufficient assistance to enable the data listing. As in 3.1 above a prioritisation as to the level of comprehensiveness is also required.

For research activities performing data collection, publishing prospective metadata, such as the methodology to be used, early in the research provides opportunity for others conducting related research to align their work and create comparable data sets.

It is CSIRO's position that there will, in some cases, be legitimate reasons for retaining or delaying dataset release including in order to meet CSIRO's other policy mandates and obligations in respect to privacy, ethics, contractual arrangements, and commercialisation and IP requirements.

Clarity on which publicly funded entities fall under this recommendation will be required.

DRAFT RECOMMENDATION 4.2

All Australian governments entering into contracts with the private sector, which involve the creation of datasets in the course of delivering public services, should assess the strategic significance and public interest value of the data prior to contracting. Where data is assessed to be valuable, governments should retain the right to access or purchase that data in machine readable form and apply any analysis that is within the public interest.

This recommendation refers to specific circumstances where private sector interests contract a public agency in an area of public interest. Outsourced services may be one such example of this. This is not likely to be relevant to CSIRO and other research agencies but it is our view that a better understanding of the types of circumstances that would fall under this proposed requirement is needed before any commitment could be made.

DRAFT 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.

The research community can assist in the ongoing development and improvement to de-identification/re-identification challenges, providing advice to agencies on current global best practice. Re-identification attacks through data linkage is a fast-moving area with new attack techniques and corresponding de-identification strategies continually emerging from the research sector. The new process recently put in place by PM&C to ensure that certain categories of datasets proposed for release are considered by third-party de-identification and re-identification experts (such as CSIRO Data61, ABS or relevant university experts) is a sound process ensuring that the circumstances around a specific release are considered.

DRAFT RECOMMENDATION 5.3

> The Australian Government should abolish its requirement to destroy linked datasets and statistical linkage keys at the completion of researchers' data integration projects.
>
> Data custodians should use a risk-based approach to determine how to enable ongoing use of linked datasets. The value added to original datasets by researchers should be retained and available to other dataset users.

This recommendation is partly supported. Linking and integrating authorities require a remit to use the results of research to improve custodian data holdings and practices. Propagating improvements to data holdings right back to the custodians is not always practicable. The ability to hold value-added data and to provide and register metadata for improving the original custodian data is a desirable outcome.

However in some cases, destruction of data may be necessary on ethical grounds, particular where destruction of data is built into experimental design of research involving human subjects and where consent to data release has not been obtained from subjects.

INFORMATION REQUEST

> The Commission seeks further views on the most practical ways to ensure improvements to linked datasets are available for subsequent dataset uses.

The value add of continued access to derived or linked datasets lies in the subsequent users' ability to understand how those datasets have been derived or linked. The key issue here is how information is provided to ensure the ongoing curation of these linked datasets. The provenance of the derived datasets needs to be tracked and preserved along with any algorithms used to transform the data from its original form. Some exceptions might be made where reverse engineering an algorithm might allow unauthorized removal of data restrictions on a dataset. It is also important to researchers and research organisations that attribution be maintained for work done to create the original linked datasets, and that this be maintained in the metadata describing the datasets in whatever repository or registry the data is made accessible from.

It is important to ensure that the provenance of source data is kept and maintained, especially though linkage processes. Data custodians should also receive prominent acknowledgement in all outputs of research resulting from use of their data.

DRAFT RECOMMENDATION 5.4

> To streamline approval processes for data access, the Australian Government should:

- issue clear guidance to data custodians on their rights and responsibilities, ensuring that requests for data access are dealt with in a timely and efficient manner;
- require that data custodians report annually on their handling of requests for data access;
- prioritise funding to academic institutions that implement mutual recognition of approvals issued by accredited human research ethics committees.

State and territory governments should mirror these approaches to enable use of data for jurisdictional comparisons and cross-jurisdiction research.

Clarity on rights and responsibilities should be backed by methodologies which allow straightforward risk assessment. Custodians that are confident of the risk of release will be more likely to share. Consideration should be given to what incentives and consequences are appropriate to support these requirements.

The broad area of data availability and use raises numerous significant ethical challenges requiring careful and detailed attention. The requirements of the Australian Code for the Responsible Conduct of Research (https://www.nhmrc.gov.au/guidelines-publications/r39) should be built into approval processes that relate to this recommendation.

DRAFT RECOMMENDATION 5.5

In light of the Australian Government's commitment to open data, additional qualified entities should be accredited to undertake data linkage.

State-based data linkage units should be able to apply for accreditation by the National Data Custodian (Draft Recommendation 9.5) to allow them to link Australian Government data, and the intention of 'open by default' should apply to these exchanges.

Some questions arise as to the relative roles and jurisdictions of a Linkage Unit (state based) and Integrating Authorities (national). For example, will state based authorities allow confidential Australian Government and state data to be linked and held by states?

DRAFT FINDING 6.1

The lack of public release and data sharing between government entities has contributed to fragmentation and duplication of data collection activities. This not only wastes public and private sector resources but also places a larger than necessary reporting burden on individuals and organisations.

Whilst this is true, the simple act of publishing data may not necessarily solve the challenge. Data also need to be discoverable (i.e. it can be found easily) and usable (the end user needs to be able to assess fitness for their purpose and to manipulate the data into a form that they can use).

DRAFT RECOMMENDATION 6.1

Government agencies should adopt and implement data management standards to support increased data availability and use as part of their implementation of the Australian Government's Public Data Policy Statement.

These standards should:

- be published on agency websites
- be adopted in consultation with data users and draw on existing standards where feasible
- recognise sector-specific differences in data collection and use
- support the sharing of data across Australian governments and agencies
- enable all digitally collected data and metadata to be available in commonly used machine readable formats (that are relevant to the function or field in which the data was collected or will likely be most commonly used), including where relevant and authorised, for machine to machine interaction.

Policy documents outlining the standards and how they will be implemented should be available in draft form for consultation by the end of 2017, with standards implemented by the end of 2020.

Agencies that do not adopt agreed sector-specific standards would be noted as not fully implementing the Australian Government's Public Data Policy and would be required to work under a nominated Accredited Release Authority (Draft Recommendation 9.6) to improve the quality of their data holdings.

A common view is that "data management" focuses on the day to day internal care of data in data systems. That part is important and adopting best practices can lead to efficiencies in agency operation and a consistent conceptual framework. However for data sharing and release it is the external view of the data that is most important.

The term "standard" commonly means a practice adopted by a standard setting body, codifying agreement within the relevant community. It may be preferable for the term "agreement" or "community agreement" to be used in contexts where there is resistance to the implication of 'top down' standardization. Before a de jure or consensus standard is made a community of practice establishes agreements (formally or informally) to deal with data in a consistent way. The standards creation process then clarifies, cleans up and codifies the practice that was already underway so that others can more easily join in.

For effective data transfer such agreements would cover protocols, structure and content of data. Appendix B of the draft report lists effective layers of interoperability and these agreements would need to cover all layers. However, not having an agreement on data or metadata may not be a reason to avoid publishing data and metadata. In fact publishing inconsistent or incompatible data and metadata in an active, functioning community can drive the process of clarification and consensus.

As indicated in the response to recommendation 3.1 an active agency governance process will seek to continually improve the quality, comprehensiveness and completeness of data and metadata. It will work with the community of data consumers and producers to reach the level of agreements necessary for effective flow of information.

The timeline given is inconsistent with the timeline for listing metadata in recommendation 3.1. Agreements in sufficiently large communities will be needed for comprehensive listing of data and metadata. However to recall Tim Berners-Lee's TED mantra 'raw data now' - i.e. you don't have to wait for standardisation and all QA/QC tasks to be complete before making datasets available. Knowing what is available can help forge communities and develop those agreements.

An agency needing help from an Accredited Release Authority (ARA) should not be considered in a negative light. Not all agencies have the capacity to deal with data and working with an ARA provides an economy of scale to data activities.

DRAFT 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.

Many standards used in the public and private sector are set by private sector organisations, e.g. the World Wide Web Consortium (W3C), the Open Geospatial Consortium (OGC), and the International Organization for Standardization (ISO). Governments can and do work alongside private sector interests in these organisations to achieve the best outcomes in both the private and public interest. Before these organisations set standards there is a degree of adoption of prospective candidates by members. Communities form around these informal agreements and the member work out the issues in practice.

These communities have a mix of private and public sector participants. The agreements are just as relevant for sharing data between private firms as sharing data in government. In many cases there will be coordination groups or peak bodies in place to represent communities and support development of agreements. In others the competitive nature may mitigate against coordination and standardization, even where this would be to the benefit of the community as a whole, as well as the broader community. Where there is a community need, but market failure, then the public sector should step in, recognising that it takes time and investment to gain agreement.

Clarity on where responsibility for such intervention sits within government and how this would be facilitated will be required. CSIRO has a track record of working to influence informal, national and international agreement for the benefit of both the private and public sector.

INFORMATION REQUEST

The Commission seeks more information on the benefits and costs of a legislative presumption in favour of providing data in an application programming interface (API) format, specifically:

- In which sectors would consumers benefit from being able to access data in an API format?
- What are the main costs and barriers to implementing APIs?

APIs enable reliable and reproducible machine to machine interactions. Appendix B goes into much detail about APIs. By not requiring a human in the loop the machines can retrieve the data as needed without delay. An API is useful where

- timeliness is important
- where multiple sources (such as different states) use the same API to access similar data or
- where many users are requesting the same data.

APIs can be provided by the controller of the data or by a third party where the third party hides dealing with the controller of the data. Appendix B lists four interaction patterns based on APIs. APIs can and do evolve over time. It is by experimenting with different flavours of things including APIs that the consensus emerges on the most effective approach. Any agreements on the form of an API are likely to be simple initially and grow over time.

Barriers include:

- Ability to reach an agreement on what the API should look like (to reduce diversity)
- Cost associated with coding an API (can be shared with multiple deployments).
- Cost associated with maintaining an API in effort, money and expertise.
- Cost of providing security on an API in two ways:
    1. Ensuring no security breach is allowed to essential systems. This can be mitigated by provide data access on a separate system to where the data in maintained.
    2. Preventing denial of service attacks. Though even with these the cost of accessing the data may still be lower than interacting with a person.

DRAFT FINDING 6.2

Data standards should aim to ensure that the content produced is usable by those who seek access to their own data. To achieve this, available data needs to be published in machine readable and commonly used formats that are relevant to the function or field in which the data was originally collected or will likely be most commonly used.

Formats are insufficient to adequately describe useability of data. As noted in appendix B, agreements are ideally achieved across all layers for interoperability for the purpose it will be most commonly used. As noted earlier, failure to reach this level of interoperability should not prevent the publication of data and metadata. On the contrary, publication without agreements, while not achieving interoperability, provides the raw material to determine what interoperability looks like if the community is active.

DRAFT RECOMMENDATION 7.1

Beyond achieving a ‘fit for release’ standard (Draft Recommendation 6.1), government agencies should only value add to data if there is an identified public interest purpose for the agency to undertake additional value adding, or:

- the agency can perform the value adding more efficiently than either any private sector entities or end users of the data; and
- users have a demonstrable willingness to pay for the value added product; and
- the agency has the capability and capacity in-house or under existing contract; and
- the information technology upgrade risk is assessed and found to be small.

It is in an agency’s interest to maximise the use and spill-over effects of their data. The key problem here is one of market failure. It is advantageous to encourage change in an industry or community sector by making critical data more easily accessible to others.

For example local councils are usually very data poor when it comes to managing the local environment or planning for emergency mitigation. This is because they lack the financial resources and local data analytics capability to access raw data from state and federal departments. Bringing together data from BOM, GA, ABS and other key agencies in a local government data portal with built in analytics could greatly improve services in regional Australia

Where there is poor use of an agency’s data, an agency might fund either lowering barriers to use or applications to showcase what could be done with the data to stimulate the market.

For research agencies an ‘identified public interest purpose’ may also include the use of data for research purposes, subject to necessary approvals including ethics approvals where personal data is involved.

DRAFT RECOMMENDATION 7.2

The pricing of public sector datasets to the research community for public interest purposes should be the subject of an independent review.

Ideally there should be no charge to the research sector for access to public data, particularly where research is aimed at generating meaning and additional value from public sector data.

DRAFT RECOMMENDATION 7.3

Minimally processed public sector datasets should be made freely available or priced at marginal cost of release.

> Where there is a demand and public interest rationale for value-added datasets, agencies should adopt a cost recovery pricing approach. Further, they should experiment with lower prices to gauge the price sensitivity of demand, with a view to sustaining lower prices if demand proves to be reasonably price sensitive.

CSIRO is currently working on how a data repository owner can assess the impact making their data publicly available will generate. A study of the impact of in-demand data could inform pricing policies. In implementing this it should not necessarily be the case that, where there is a public interest rationale for value-added datasets, there should be charging for that dataset. There are many situations where the most efficient and effective approach includes the release of open and free value-added datasets of high public interest. Conversely, where this is not a public interest rationale for value-added datasets, it may be appropriate to charge for that dataset subject to the pricing policies promulgated by the Australian Government. Pricing should reflect the purpose for which a value-added dataset is used for example national for public interest, or for commercial profit.

It should be recognised that there is likely to be a difference between the level of adoption of a free dataset and that of a dataset of a very low price. Online purchase of datasets (even when low price) creates both a barrier to the user and an overhead to the supplier and has more impact on the use of a dataset than the difference between a low price and a high price.

DRAFT RECOMMENDATION 7.4

> For datasets determined through the central data agency's public request process (Draft Recommendation 2.1) to be of high value and have a strong public interest case for their release, agencies should be funded for this purpose. Funding should be limited and supplemental in nature, payable only in the event that agencies make the datasets available through release or sharing.
>
> Aside from this additional funding, normal budgetary processes should apply for all agencies' activities related to their data holdings.

CSIRO notes and generally supports the intention of this recommendation. However, from a system wide perspective, it is important that this recommendation does not lead to agencies choosing whether or not to release data on the basis of funding. It is preferable that all data and metadata is published with appropriate access control and prioritisation of high value data sets.

DRAFT FINDING 8.1

> It is important governments and businesses maintain a social licence for their collection and use of data. This can be built through enhancement of consumer rights, genuine safeguards, transparency, and effective management of risk. Community trust and acceptance will be vital for the implementation of any reforms to Australia's data infrastructure.

Social licence to operate will be supported by a transparent approach to managing trust, privacy and security issues in relation to use of personal data. As mentioned in the introductory comments, consideration of the range of ethical issues that arise from proposed changes should be an integral component of the planning and change process.

DRAFT 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.

These rights are a positive step forward. Some considerations are:

- Digital rights management when an individual dies.
- Can an individual's digital footprint be inherited?
- Creating a mechanism where an individual can "donate" their data to research upon confirmation of their death by the appropriate births, death, and marriages registry.
- What rights family and descendants have to access an individual's data when finalizing their affairs or determining what constitutes reasonable family or historical record?
- The donation of complete unfettered comprehensive health records could add significant benefits to health and wellbeing studies when combined with such things as donations of an individual's body for research.
- Power of attorney over these rights.
- Assistance to people disabilities may be necessary to access these rights

Restrictions on allowable third parties for transfers could reduce fraud and identity theft. An exclusion list and/or a list of approved purposes and approved recipients (endorse by an industry association) could reduce leakage of data through scams. The ability to opt out of such measures would give flexibility in unusual circumstances.

Data transfer is not always effective for the consumer. Where the consumer holds a large amount of data, such as a video and photo archive, then transfer is not always practical. What could be required is authorised access with use restrictions, typically through an API. In other cases what is desired are events based on data holdings. For example an app may want to notify a consumer when the bank balance reaches a certain level or when a certain transaction occurs.

This requires a much more complex interface or the support for third party code to operate in a controlled environment over the account. Technologies for this exist but require parties to agree and an openness to change. The concept of transfer could be interpreted broadly to include third party use of the data on behalf of the consumer. The private sector can develop domain specific ways of making the data available. An arbitrator may be required where third parties cannot reach agreements with data custodians.

The right to appeal automated decisions implies the decision maker understands the provenance of the data. There is an implied right of the individual to be told the origin of the personal data and the requirement on the data holder to track the provenance. The right to appeal seems relevant to all decisions based on data.

INFORMATION REQUEST

The Commission seeks further views on datasets that are of national interest and that could feasibly be designated as such under the process proposed.

These data sets include:

- Foundational Spatial data as described on http://www.anzlic.gov.au/foundation-spatial-data-framework
- Integrated health data across MBS, PBS, hospital admissions, treatment, outcomes and e-health records to develop longitudinal analysis of illness, treatment and outcomes.
- Real time economic activity data from housing commencements, to grocery bills and credit card expenditure on a linked, de-identified and aggregated basis. This data is commercially sensitive and therefore commercial considerations need to be taken into account.
- Longitudinal education and employment outcomes to the first few years of employment. Linked, de-identified and aggregated.

DRAFT RECOMMENDATION 9.4

The Australian Government, in consultation with state and territory governments, should establish a process whereby public and private datasets are able to be nominated and designated as National Interest Datasets (NIDs).

Datasets (across the public and private sector) designated as NIDs would satisfy an underlying public interest test and their release would be likely to generate significant community-wide net benefits. Designation would occur via a disallowable instrument on the recommendation of the National Data Custodian.

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.

The process to designate datasets as being of national interest should be open to the states and territories in order to cover linked datasets, with negotiations undertaken to achieve this.

For community confidence, consideration should be given to use of a deliberative forum, such as a parliamentary committee, to take community input on and review nominations made, and to make proposals for future designations.

Private sector data gathering activities could add significant value to the Australian economy if they were combined with existing government datasets in many sectors. In areas as diverse and agribusiness and health, corporate Australia is generating huge amounts of data that, given the right agreements, could be ingested into some of the National Interest Datasets to inform government responses in specific sectors. The relationships created could be mutually beneficial in linking corporate and government data to allow business to make better strategic decisions and hopefully become more profitable where this is clearly in the national interest.

The nomination of these National Interest Datasets should also include recommendations for the ongoing curation and maintenance of data assets especially when there are changes in the Machinery of Government or when a private sector source ceases operations or is sold on to other entities. Examples of this might include the areas of health, water and energy security. Obligations to contribute or maintain national interest datasets need to be maintained if enduring value is to be derived. This is particularly relevant to small agencies that may hold important data sets. When they are absorbed into larger agencies the mandate to maintain the NID should remain. To gain value from NIDs they need a long life.

DRAFT RECOMMENDATION 9.5

The Australian Government should establish an Office of the National Data Custodian, as a new function within the Government to have overall responsibility for the implementation of data management policy.

Specifically, the National Data Custodian (NDC) would have responsibility for broad oversight and monitoring of Australia's data system, recommending the designation of National Interest Datasets, and accrediting Release Authorities and trusted users within the reformed data system.

In a data systems some core components and methods are best shared across agencies. The NDC could play an important role in informing decisions for establishing multi-stakeholder infrastructure.

The NDC has the potential to have a role in facilitating, advocating and coordinating activities. T and, ideally would include broad representation of stakeholders.

Box and Lemon (2015) provide an analysis and recommendations for governance and incentivising participation in distributed cross government national environmental information infrastructure.. http://www.neii.gov.au/system/files/filedepot/1/The%20Role%20of%20Social%20Architecture%20in%20Information%20Infrastructure.pdf

DRAFT RECOMMENDATION 9.6

Selected Australian and state/territory government agencies should be accredited as Release Authorities by the National Data Custodian. In considering applications for accreditation, the National Data Custodian should consult a wide range of parties and ensure Accredited Release Authorities (ARAs) have sectoral expertise. The current model used by the National Statistical Service for appointing data linkage authorities should be considered in developing a model upon which to base this process.

ARAs will be responsible for:

- deciding (in consultation with initial data custodians) whether a dataset is available for public release or limited sharing with trusted users
- collating, curating and ensuring the timely updating of National Interest Datasets.

ARAs will also perform an important advisory role in regard to technical matters, both to government and to the broader community of data custodians and data users.

Having ARAs aligned to sectors/domains is an attractive possibility permitting alignment across industry, government, research, academia and community with appropriate collaborative approaches to working flexibly across domain boundaries.

While many lead agencies are potential candidates as ARAs, for example Geoscience Australia, the ABS and the Bureau of Meteorology, many agencies service overlapping sectors with data sets relevant to many agencies. For example the Bureau of Meteorology is responsible for bore water data and Geoscience Australia is responsible for borehole data. An approach to attributing datasets to ARAs will be required.

ARAs will have a key role in advocating sector specific data governance and promoting the conversations on continuous improvement within the sector. ARAs, like all parts of the data ecosystem, should have ongoing dialogue with each other as there are significant commonalities in dealing with data in different sectors.

It is important that the sectoral knowledge of smaller agencies is not lost when an ARA brings their data into a NID. This is particularly important during Machinery of Government changes where an ARA may control the NID but the small agencies with the domain expertise may no longer exist.

DRAFT RECOMMENDATION 9.7

Trusted users should be accredited by the National Data Custodian for access to those National Interest Datasets (NIDs) that are not publicly released. Trusted users should be drawn from a wide range of potential entities, including: all Australian Government and state and territory government agencies; all Australian universities; and other entities (be they corporations, not-for-profit organisations or research bodies) that are covered by privacy legislation.

The default position should be that someone from one of these organisations would be approved for access unless the National Data Custodian transparently specifies a reason, on consideration, of why this should not occur.

For trusted users of NIDs, trusted user status should provide an ongoing access arrangement, with few restrictions on what could be done with the data. Trusted user status for NIDs should cease when the user leaves the approved organisation or be suspended if a breach occurs by any other trusted user in that same organisation and/or working on the same project.

A common way of determining trusted access to data across all domains should simplify the access process.

DRAFT RECOMMENDATION 9.8

Arrangements for access by trusted users to identifiable data held in the public sector and by publicly funded research bodies should be streamlined and expanded by the Australian Government. The National Data Custodian should be given responsibility to:

- develop, in consultation with data custodians, a list of pre-approved uses for a dataset, and make decisions on access to data for projects not consistent with the pre-approved uses list
- grant, on an approved project-specific basis, trusted user access to personnel from a range of potential entities, including: all Australian Government and state and territory government agencies; all Australian universities; and other entities (be they corporations, not-for-profit organisations or research bodies) that:
    - are covered by privacy legislation
    - have the necessary governance structures and processes in place to address the risks of inappropriate data use associated with particular datasets, including access to secure computing infrastructure.

Access would be granted for the life of the specific approved project. Trusted user status for use of identifiable data would cease when the user leaves the approved organisation; a project is completed; or if a breach occurs in that same organisation and/or project.

Most data can be used to identify characteristics of a person if enough is already known about the person. This risk increases when data from several government sources are cross linked, particularly in rural areas and small communities. A methodology or tool is need to assess the re-identification risk a data set poses.

Consideration should be given to a whole of government authentication and authorisation system perhaps similar to the Australian Access Federation used in academia. Using a federated access control system where not only a user's credentials are used but also a list of key attributes like organisation, project team membership etc. would greatly simplify the day to day operations of any system supporting this recommendation. An additional feature of this system is the ability to audit an organisations processes for identity management from time to time under a common agreed access framework.

The greater availability of personal data in the public space make some forms of authentication practices by business at risk. Many use name, address and date of birth. Some use data from statements which are now digital and could be transferred to a third party under the new rights.

The response to 8.1 is also relevant here in relation to how institutions recover from any breach and regain community trust.

Trusted access should also cover sensitive data (e.g. cultural sensitivities, species locations etc.) not just identifiable data.

DRAFT RECOMMENDATION 9.9

Public research funding should be prioritised on the basis of progress made by research institutions in making their researchers' data widely available to other trusted researchers on conclusion of research projects.

CSIRO notes this recommendation and observes that it would be difficult to implement at present as there is not currently an adequately standardised process for distribution of research data. It is also not always the case that research data should be widely distributed even when the research is fully publicly funded. However, the recently released Draft 2016 National Research Infrastructure Roadmap draws attention to the importance of "digital data and eResearch platforms" (see page 24) and, for coherency, it would be desirable that the consideration of this recommendation occurs in the context of government's decisions in relation to the Roadmap.

It should be recognised that research agencies will vary in their level of readiness and capability to respond to requirements in relation to data availability. Agencies with ARA status should be encouraged to provide support across the research sector to lift competency and capability, and progress measured relative to an agency's starting point on this journey.

DRAFT RECOMMENDATION 9.11

The Australian Government should introduce a Data Sharing and Release Act which includes the following:

- Provisions requiring government agencies to share and release data with other government agencies and requiring sharing between government agencies and other sectors.

  - These provisions would operate regardless of all restrictions on data sharing or release contained in other legislation, policies or guidelines.
  - The provisions may be waived in limited exceptional circumstances, and the Act should specify what these circumstances are.

- Strengthened provisions on access to data by individuals, including rights to access and edit data about them, a right to have data copied and transferred, and a right to request that collection cease.
- Provisions establishing the Framework for the governance of Comprehensive Rights of consumers, access to National Interest Datasets, approval of trusted users, and accreditation processes for Release Authorities.

CSIRO supports a cooperative approach to this issue in the first instance, rather than the introduction of legislation. In CSIRO's view, this being an area of significant complexity with multiple players and quite widely varying levels of sophistication across the system, a cooperative approach is more likely to lead to beneficial outcomes with reduced adverse impacts, than alternative approaches. However, CSIRO acknowledges that it might be the case that an analysis as to whether legislation would be required to *permit* government agencies to share data (i.e. to provide a legislated right to decide so to do) would lead to a motivation for legislation for that limited purpose.

CONTACT US

t 1300 363 400
+61 3 9545 2176
e csiroenquiries@csiro.au
w www.csiro.au

AT CSIRO, WE DO THE EXTRAORDINARY EVERY DAY

We innovate for tomorrow and help improve today – for our customers, all Australians and the world.

Our innovations contribute billions of dollars to the Australian economy every year. As the largest patent holder in the nation, our vast wealth of intellectual property has led to more than 150 spin-off companies.

With more than 5,000 experts and a burning desire to get things done, we are Australia's catalyst for innovation.

CSIRO. WE IMAGINE. WE COLLABORATE. WE INNOVATE.

FOR FURTHER INFORMATION

Dr Jack Steele

Tom McGinness

Data61
Gavin Walker
t +61 0 2 6216 7030