ADA AUSTRALIAN DATA ARCHIVE

Australian Data Archive
ada@anu.edu.au
www.ada.edu.au

Dr. Steven McEachern · *Director, Australian Data Archive*
Prof. Matthew Gray – *Director, ANU Centre for Social Research and Methods*

AUSTRALIAN DATA ARCHIVE and ANU CENTRE FOR SOCIAL RESEARCH AND METHODS

PRODUCTIVITY COMMISSION “DATA AVAILABILITY AND USE” PUBLIC INQUIRY

SUBMISSION ON INQUIRY DRAFT REPORT

12 DECEMBER 2016

ABOUT ADA AND THE ANU CENTRE FOR SOCIAL RESEARCH AND METHODS

The Australian Data Archive (ADA) provides a national service for the collection and preservation of digital research data and to make these data available for secondary analysis by academic researchers and other users. ADA was established at the ANU in 1981 with a brief to provide a national service for the collection and preservation of computer readable data relating to social, political and economic affairs and to make these data available for further analysis. ADA provides access to data from government, academic and private sector sources for researchers and policy makers.

A team of professional data archivists provides both stewardship and outreach services to the Australian community. The archive:

- acquires, documents, preserves and disseminates data online to a broad range of social science researchers in the university, government, and other sectors
- provides the only comprehensive social science data collection in Australia, with a catalogue of over 5000 data sets collected in over 1500 research projects dating back to 1947
- holds data from Australian surveys, opinion polls and censuses and includes data from other countries within the Asia Pacific region
- supports access to datasets from 10 different Federal government departments and agencies
- provides specialist services within specific subject areas, including social sciences, health, indigenous studies, electoral behaviour, criminology and some humanities disciplines, and within specific data types, including quantitative, qualitative, time series and panel data, and historical statistics
- locates and manages access to overseas social science data sets required by Australian based researchers
- adopts, develops and applies standards in line with international best practice
- belongs to international organisations (such as the International Federation of Data Organizations and the International Association of Social Science Information Service and Technology) and plays a major role in cross-national collaborative projects
- provides support for the management and dissemination of grant-funded data collections such as ARC and NHMRC projects

The ADA is located in the ANU Centre for Social Research and Methods (CSRM). The centre provides significant Australian capacity in undertaking multi-disciplinary research into major areas of applied social research, meeting a need for national leadership in research methods. The ANU CSRM produces credible evidence-based research to inform and evaluate policy, programs and initiatives.

COMMENTS ON DRAFT FINDINGS AND 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.

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.

Comment:

While these recommendations are worthwhile and commendable as statements of intent, we would suggest that both the Australian Government and publicly funded entities would struggle to meet the commitment to a comprehensive data register. It seems likely that most, if not all, universities and government agencies would, at present, be unable to comprehensively document the research data that they hold.

As such, an approach such as that outlined in Draft Recommendation 3.2 will require the implementation of relevant mechanisms within these organisations to enable the (preferably automated) capture of such information. Systems such as data management plans (e.g. DMPOnline - https://dmponline.dcc.ac.uk/) may be a means for capturing information on intended data creation, and catalogues such as data.gov.au and the ANDS Research Data Australia catalogue provide a means for hosting this information. There is however a significant amount of investment required to ensure that this information is captured and made available.

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.

Comment:

Appropriate data de-identification is an important part of an overall data access regime that provides a means for enabling access to government and other data sources. However there is the potential for these de-identification processes to result in data that is no longer fit for purpose. One example here is the use of top coding of age categories in population data – often to ages such as “65 and over” or “75 and over”. Another example is the top coding of income data, or making such data available in bands. While such techniques reduce identification risk, the method makes such data less useful for research studying those in retirement or later in life or those studying the upper end of the income distribution respectively.

We would recommend the use of de-identification techniques in the context of a broader trusted user model, such as the Five Safes model discussed in Section 5.6 of the report. De-identification, as a “Safe Data” technique, is one of a combination of approaches that can be used to enable improved access to data. Its use in conjunction with the other elements of the Five Safes model (safe people, safe projects, safe settings and safe outputs) should be a core part of the overall Trusted Access model. This point is reflected in Section 9.6 of the Draft Report on “Sharing Identifiable Data with Trusted Users”, which outlines one possible model for sharing such data in a trusted setting (such as the Secure Unified Research Environment), but should be made explicit in the Final Report.

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.

INFORMATION REQUEST

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

Comment:

ADA and CSRM support this recommendation in principle.

In terms of the Commission’s information request regarding practical methods for ensuring improvements to linked datasets, ADA would suggest initially applying the principles outlined in Draft Recommendations 3.1 and 3.2 to those datasets that have already been linked or will be linked as part of an approved research program[1]. This would include the implementation of a metadata registry for those datasets that have been already linked to then be made available for potential future use (subject to the conditions outlined in the proposed Trusted Access model).

[1] This is now occurring with datasets linked by the ABS in their role as a Data Integration Authority (e.g. http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/3418.0.55.001Main+Features12009-10%20and%202010-11), but should be extended to cover other integration authorities and data linkage projects.

More generally, we would recommend that newly created linked datasets be documented, curated and managed as data assets in the same way as other national interest datasets.

In terms of enabling improvements to linked datasets being made available for future use, this forms part of a greater set of challenge which are faced by data providers– how to enable feedback from users on the use of their data, and to incorporate that feedback into improvements in subsequent versions of the data. There is also a related issue of how to ensure that data is appropriately cited and that data creators are given adequate recognition for their intellectual contribution.

To date, there are no clear solutions that have emerged for enabling these feedback loops. One possible solution may lie in the use of open science frameworks (possibly in conjunction with the use of APIs for data access), which involve the publication of code and source data – or links to data – in conjunction with research publications[2]. Access to the open code may be a means of facilitating improvements to the source data by incorporating the documented changes that occur in the analyses conducted with the 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.

Comment:

ADA and CSRM strongly support each of these requirements. In addition, we would suggest that the development of "best practice" procedures for data access would provide an important complement to the other guidance suggested above.

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.

Comment:

[2] See for example the Open Science Framework: https://osf.io/

ADA and CSRM support this proposal. However we would recommend two extensions:

1. The extension of the capacity for accreditation to organisations outside the government sector, in particular to suitable organisations within the university sector but also potentially elsewhere.

2. The further extension of accreditation to enable non-government organisations to be accredited as "Approved Release Authorities"

We would also note that ADA would be willing and qualified to act as either (or both) an accredited integration authority and an approved Release Authority should such possibilities be available under legislation. We also note that other non-government organisations (such as the Population Health Research Network) might also be in a position to provide such services, addressing possible capacity constraints that may slow down access under the proposed access model identified in the Draft Report.

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.

Comment:

ADA is a contributor to standards development and implementation in both Australia and internationally. We believe that the adoption of data management standards is a critical requirement for the preservation, management and reuse of data.

In terms of sector-specific standards, we would particularly argue for the adoption of the DDI (Data Documentation Initiative[3]) standard for the management of unit record data (or microdata) within the social, behavioural and economic sciences and related fields. DDI is used by both government and academic organisations in over 80 countries around the world, to document both survey and administrative data, and enable machine readable access to data for research and administrative purposes. In Australia, the DDI standard is used by both ADA and the Australian Bureau of Statistics for the management of microdata for preservation and access.

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.*

Comment:

The resourcing of additional activities to enable value adding is a challenge in the management of data such as that provided under the Public Data Statement. However we would suggest that there is value in engaging in some limited value adding to ensure the capacity for data to be made adequately suitable both human and machine-readable purposes. For example, the application of value-added information to variables in datasets (such as occupational classifications or industry codes) may assist to enable users to automatically work across multiple datasets using a consistent set of indicators. Similarly the identification and harmonisation of geographic information within datasets would assist to ensure that machine-based mapping tools (such as the NationalMap technology now available through data.gov.au) can be applied on a consistent basis.

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.

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,

[3] http://ddialliance.org

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.

Comment:

ADA strongly recommends the adoption of a freely available access model wherever possible (and this is our own policy as of October 2015). We would note that this does not automatically imply an open data model, but it is a necessary condition for open data access such as that provided through data.gov.au.

We would also note from experience the need to provide funding particularly for the curation of data to enable it to be made available through both open and trusted access mechanisms. Many government (and academic) datasets include significant amounts of often implicit knowledge which needs to be incorporated into metadata and other documentation – for both access and preservation purposes.

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.

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.

Comment:

ADA and CSRM support the establishment of both the proposed agencies above in principle. We do however have some additional suggestions as to the scope and operation of these agencies

In relation to the National Data Custodian (NDC), we believe that a central coordinating agency would enable the establishment of consistent and coordinated practices across multiple agencies, which in our experience have varying levels of both knowledge and experience within this area. However we would like to see some means for ensuring that the NDC provides more timely and coordinated data access in practice, rather than an additional layer of bureaucracy and management for both data providers and users to negotiate.

In regard to the Accredited Release Authorities, we would like to note that such agencies will have an important role, but that (as noted in our response to Recommendation 5.5 above) that there should be consideration given to providing the capacity for organisations outside government to be approved as ARAs. The Australian Data Archive has in fact fulfilled a role very similar to that described in the Draft Report for numerous government department agencies, including 10 different agencies at the time of writing. Similarly in other sectors there may be equivalent non-government organisations who could perform similar roles.

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.

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.

Comment:

ADA is broadly supportive of the proposed framework identified by the Commission for access to National Interest Datasets and identifiable datasets. We would note that, in our experience, there are some additional administrative burdens that come from managing access to data for finite periods (rather than indefinitely), particularly where the project may require extensions of time or changes in personnel over the life of the project. These are however manageable.

We would however highlight that there may be capacity for reducing the need for project-based approval where the data has been de-identified (Level 2 in Figure 3 of the Draft Report). In terms of the Five Safes model discussed elsewhere, the use of de-identified data (Safe Data) and access by approved Trusted Users (Safe People), potentially in secure access environments (Safe Settings) may be sufficient risk management to enable relaxing the requirement for specific project approval (Safe Projects).

OTHER COMMENTS

There are two additional comments that we would make in this submission. First, in terms of the relationship between data providers and users in the government and academic sectors, the Draft Report makes regular note of the need for researcher access to government data for research purposes, particularly in those areas which might be considered in the public interest. The report also notes that there are other developments that have occurred in the academic sector (such as the SURE Secure Unified Research Environment) which may be utilised for access to government data.

There does however appear to be a significant division that is drawn between government and academia in terms of supporting access to government data. The proposed access models in the Draft Report all refer to government agencies in providing approval for access, support for curation and dissemination of data, and technical expertise. While there are good governance reasons for managing these processes within government, this does however provide a capacity constraint in terms of the technical and administrative skills required to support the proposed model.

We would note that this is one area that academia is particularly well-positioned to contribute to. Organisations such as ADA and the Population Health Research Network (PHRN, the providers of SURE) have a long history in enabling academic and government access to data for research and other purposes. Similarly universities across the country provide much of the research and statistical training required in the skills that are necessary to implement the model being proposed in the Draft Report.

We would recommend therefore that consideration be given to:

1. Articulating more clearly the role of the Australian academic sector in supporting and contributing to the proposed data access model

2. Extending the capacity for both the Accredited Release Authority mechanism and the data access environment

The second additional comment that we would make is that while ADA and CSRM are strongly in support of the default stipulated in the Productivity Commission draft report that de-identified data be made available for re-use and analysis, it also needs to be recognised that certain types of data collections may require additional levels of control. In particular, there is a strong need to ensure that data that is collected on or by Aboriginal and Torres Strait Islander communities is made available in a way that protects the rights and ownership of that data. That does not necessarily mean that such data is not made available to other researchers from outside the community. Rather, it means that the communities have sufficient control over access, and are sufficiently resourced to ensure that the data is used in a way that directly benefits the community. The ADA has a long-standing and successful model whereby data custodians retain some control over the release of datasets, whilst minimising the costs and burden to legitimate data users.

CONTACT FOR FURTHER INFORMATION

For further information on any of the comments in this submission, please contact either of the authors of this submission, Dr. Steven McEachern, Director of the Australian Data Archive and Prof. Matthew Gray, Director of the ANU Centre for Social Research and Methods at the Australian National University.

Dr. Steven McEachern
Ph. (02) 6125 2200
Email steven.mceachern@anu.edu.au

Prof. Matthew Gray
Ph. (02) 6125 8265
Email: matthew.gray@anu.edu.au