Submission to the Productivity Commission Inquiry into Data Availability and Use

The Australian Academy of Science National Committee for Data in Science (NCDiS) aims to foster the discipline of data science in Australia, link the Academy to relevant Australian scientists, practitioners and societies, and serve as a link between Australian and overseas scientists, primarily through establishing links with international unions in the field of data science.

The overarching purpose of this committee is to enable Australian scientists to more effectively exploit the emerging data-rich research environment in order to conduct data-intensive science. This purpose is achieved through a number of mechanisms, including providing strategic science policy advice on scientific data management, and through promoting the adoption of best practices by scientists working with data in Australia. This submission is directed towards these goals.

The NCDiS broadly supports the conclusions and draft recommendations of the Productivity Commission Draft Report on Data Availability and Use. The NCDiS supports the proposition that Australian publicly and privately held datasets are not used to their best capacity. Many datasets are difficult to access, whether for reasons of regulatory restriction or technical incapacity. Government agencies may be unwilling to share data with researchers because of concerns regarding how data will be interpreted, or because of the additional costs of making data accessible for reuse by others, or due to the concern that data may be misused, or used beyond its known limitations. The burdens imposed by these restrictions have a disproportionate effect on people seeking to make use of these datasets for purposes of research directed towards improving public policy. The NCDiS strongly agrees that there needs to be a shift away from only releasing data on request for particular projects and for an *ad hoc* charge, towards actively publishing data in a more coordinated and cost-effective way.

The recommendations in this submission must also bow to physical reality and capability limitations: the internal ICT infrastructure and expertise of some agencies simply cannot support modern forms of online delivery of information. Circumstance is often as much of a barrier to data access as regulation is; many agencies suffer from narrow pipelines, inaccessible data and limited resources and expertise.

The National Data Custodian and *Data Sharing and Release Act*

The NCDiS strongly supports the establishment of a National Data Custodian (NDC) and a *Data Sharing and Release Act*. While the Accredited Release Authorities have responsibility for curation and stewardship of data under the Framework described in the Draft Report, it would be expected that the NDC would also have an overall stewardship role to ensure that the data and the metadata that is released is of the highest quality. This agency would also be responsible for providing guidance on data release, data licensing, data re-use, data quality and metadata standards.

Ideally, the NDC will also provide a Dataset Registry that enables nationally significant datasets to be discovered and accessed via APIs. The commission and the government should consider the work of the Australian National Data Service (ANDS) in developing *Research Data Australia*[1], which provides a Web Portal to enable the discovery of nationally distributed research datasets.

Nationally Important Datasets

The NCDiS supports the concept of Nationally Important Datasets, assuming it is coupled to an agreed process for identifying and prioritising these datasets. Researchers, government employees, policy makers and planners should be able to submit proposals identifying datasets that they believe should

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

be considered “nationally important”. Each submission should include a justification outlining the dataset’s national significance, and its re-use value, if it was made openly available. The selection criteria should include attributes such as: the uniqueness of the dataset; the quality of the dataset and metadata; the ability and cost of regenerating the dataset; the number of potential users; the relevance of the datasets to real-world problems; the temporal and spatial scope of the dataset; the estimated cost of making the dataset accessible online.

A panel of experts should then assess these proposals periodically based on an agreed set of criteria, with assessments communicated to data holders and Accredited Release Authorities.

A list of requested datasets under consideration should be published on a public web page, and registered users encouraged to submit feedback to proposals. This would allow the extent of community support for proposed datasets to be assessed.

Until recently, many datasets from government research agencies such as CSIRO, the Bureau of Meteorology, Geoscience Australia and the Australian Institute of Marine Science were inaccessible, including multi-decade Landsat Earth observation datasets for Australia, climate records, geophysical data, bathymetry and elevation data. Access to selected subsets of data from these agencies, has been made available, via a 10PB data store at the National Computational Infrastructure at ANU, through funding from the Research Data Services Program. This has led directly to a number of new research projects and innovative studies. However there are many more legacy and emerging datasets being generated by these agencies that are not available. It is vitally important that datasets acquired by publically funded research organisations through government funded research be made generally accessible to the research community, in order to maximise the use of this data.

The release of this data will foster an environment in Australia that favours data-intensive research. Coupled with the considerable investments in recent years in research supercomputers, cloud computing and data storage, open access to significant datasets would enable researchers to explore new research directions, particularly in cross-disciplinary issues such as social impacts of climate change, improved community resilience to natural hazards, efficient use of and planning of government services, and sustainable use of our water, mineral and energy resources.

Datasets that are most useful for planning and policy making are those that are available at the finest spatial and temporal granularity, across the entire nation (e.g., all of the ABS Statistical Area 1 regions). Some example datasets that should be given high priority include:

- Data on activity and usage of government services and facilities. For example, how often are they being used, by whom, by how many, and how much are these services costing the government:
    - Government services (hospitals, medical services, Centrelink, educational institutions (schools, universities), community housing, public transport, welfare and immigration services);
    - Government facilities (cultural institutions and recreational facilities) and infrastructure (roads (both public and toll roads), bridges, tunnels, etc.)
- Datasets produced by the NCRIS facilities, such as the Terrestrial Ecosystem Research Network, the Integrated Marine Observing System, the Atlas of Living Australia, Population Health Research Network, Australian Phenomics Network and Plant Phenomics Facility, Astronomy Australia Limited, AuScope, Australian Urban Research Infrastructure Network (AURIN), Bioplatforms Australia, National Imaging Facility etc.

- Health, medical, hospital data, to enable policy makers and researchers to understand population health patterns and develop early diagnosis and preventative processes
- Health insurance data, both public and private
- Environmental data such as air quality and water quality data (waterways as well as drinking water), collected by both state and national environmental monitoring agencies
- Election data from the Australian Electoral Commission.
- Police data that records the location and frequency of crimes and type of crimes, accidents etc.
- Emergency services data: fire, ambulance, floods, emergency responses, etc.
- Courts statistical information for the Supreme, District and Magistrates Courts across the states, as well as the High court (not already available through the ABS).

In addition, both existing and emerging datasets being generated by government research and other agencies[2] such as CSIRO, the Bureau of Meteorology, Geoscience Australia, the Australian Institute of Marine Science, Great Barrier Reef Marine Park Authority (GBRMPA) should continue to be made accessible through funded programs such as the Research Data Services Program.

Release of datasets

The NCDiS considers that there should be a general expectation that publicly funded research datasets should be made more accessible. This includes grant-funded research at universities, research at publicly funded research agencies, and research conducted by government itself. There may be exceptions to this expectation, such as for sensitive data (medical or social science data, for example, or location information for indigenous artefacts or rare fossil or mineral specimens). Specific conditions restricting access to datasets will necessarily be dependent on the specific attributes of each dataset and the necessity to maintain confidentiality of private or personal data.

With respect to payment for datasets, the ideal is “open access” and “freely available”, followed by “cost recovery”. These models will allow the most use of the datasets. The current model is focussed on releasing data only on request for specific projects, at a cost that is not transparent; the NCDiS supports a move towards actively publishing datasets in a more coordinated and cost-effective manner. The current situation, where agencies such as the Australian Bureau of Statistics can charge different researchers (from the same or different institutions) for release of the same dataset, should be avoided as it represents a significant waste of research funding.

The main values that should be added to all datasets before release are:

- Standardisation of data formats and metadata, across government, research and private datasets. High quality, standardised data formats and metadata schemas will enable users to trust the source and the quality of the data.
- Aggregation of datasets to the finest spatial granularity, whilst still maintaining anonymity (i.e. Statistical Area 1 for geographical data).
- Application Programming Interfaces (APIs), so that the datasets can be queried, accessed and retrieved programmatically.
- The use of “unique identifiers” or “linkage keys” that enable the linking of datasets to other related datasets.

[2] http://www.australia.gov.au/about-government/departments-and-agencies/list-of-departments-and-agencies

- Identification of a long-term data steward, who can preserve the data through iterations of hardware and software evolution.

Where data is standardised, the standards should be fit for purpose, widely accepted and under a clear governance structure that allows the standards to evolve. Such standards should be specified and maintained by the NDC or a similar central body.

Registries of datasets

The NCDiS supports the publication of registries of data holdings by publicly funded agencies and publicly funded research groups. For example, infrastructure facilities supported under the National Collaborative Infrastructure Scheme, and collaborative facilities supported as Cooperative Research Centres, should have data management plans which specify where data collected by the facility will be stored, the format(s) that it will be stored in, the metadata schemas that apply and the protocols by which the data can be programmatically accessed. Additionally, there may be schemes (such as that implemented within Curtin University's geochemical laboratories or within the OzTrack animal tracking data repository) whereby access to facilities or services will be prioritised or costs reduced, for researchers who agree that their data will be made publicly available within a reasonable timeframe (e.g., after a 1 year embargo period).

The NCDiS notes that both Australian Research Council and National Health and Medical Research Council grant applications currently require data management plans. However, although these plans are required as part of the proposal, they are not actually assessed during the evaluation, or monitored over the duration of the grant. There is scope to more actively manage research project data management plans. The Research Data Alliance is actively working on this topic[3]. If data management plans were given unique Digital Object Identifiers (DOIs), then a permanent public record would be kept of which data collections were funded by the Research Councils. Ideally, such a process would also apply to research projects undertaken within publicly funded research agencies such as CSIRO.

Researchers and policy makers who use published datasets to derive new valuable data products should also be obliged to register/publish the derived data products as new datasets, and to publish the methods used to derive the data products.

If public research funding is to be prioritised on the basis of the extent to which research datasets are made accessible, the criteria for this prioritisation should be made clear. There are some models that might be considered:

- Formally integrating the evaluation of research data management plans into the grant evaluation process, along with research track record and quality of the research proposal
- Staggered funding, whereby a proportion of the grant value (~80%) is paid up front, with the remainder withheld until the project data is properly documented and accessible from a public repository. Ideally a proportion of the project budget (e.g., 10%) should be allocated for data management and publishing.
- Including dataset publications as reportable grant outputs.
- Including the researchers' track record in publishing and sharing their research data generated from previous grants, as part of the grant assessment process.

[3] https://www.rd-alliance.org/group/active-data-management-plans/case-statement/active-data-management-plans-ig.html

Trusted users

The proposal to identify "trusted users", who would be able to receive more sensitive data, will allow a certain level of administrative efficiency as well as greater use of released datasets. If these users are restricted to researchers at universities or other research institutions, or to state and federal government employees, for the duration of a specific project, then pre-processing of potentially private datasets will be greatly simplified.

It is important to note that as well as the trustworthiness of the user, there should be a consideration of the trustworthiness of the infrastructure with which the user will store and process the data. Vetting the trusted users should be the responsibility of the NDC, as described in the Draft Report.

Personal data

Personal, individual data carries the highest risk of misuse, security breaches and privacy breaches. To expedite the release of datasets, aggregated, anonymised datasets should be prioritised for early release; personal, individual data should be considered at a later date. All released datasets should be aggregated at a level that does not allow personal identification, using practices already in place at the Australian Bureau of Statistics and the Australian Institute of Health and Welfare.

The remit of the NCDiS does not extend to considerations about access of individuals to their own data. In this case, ethical principles such as those expressed in the NHMRC/Universities Australia *National Statement on Ethical Conduct in Human Research* should apply.

Other comments

The NCDiS would like to note that a significant amount of work has already been undertaken to develop best practices, services and infrastructure for data sharing and reuse. In particular, the NCDiS draws the Commission's attention to:

- Report on Best Practice for Research Data Management Policies published by the ICSU Committee on Data for Science and Technology (CODATA)
- *Legal Interoperability Of Research Data: Principles And Implementation Guidelines*, published by the Research Data Alliance and CODATA
- *The Value of Open Data Sharing*, a report by the Group on Earth Observations (GEO) and CODATA
- Income Streams for Data Repositories published by CODATA
- Existing national and international initiatives including:
    - *ICSU-World Data System Data Sharing Principles*
    - Group on Earth Observations (GEO)
    - *G8 Science Ministers' Statement* and *Open Data Charter*
    - *OECD Principles and Guidelines for Access to Research Data from Public Funding*
    - *Science International Accord on Open Data in a Big Data World* enunciated jointly by the International Council for Science, the InterAcademy Panel, the International Social Science Council, and the World Academy of Science.

Draft recommendation 5.5 focuses on accreditation of entities to perform data linkage, but accreditation of data repositories is also necessary. Previous work on certification of trusted repositories has also been undertaken by ICSU World Data System (WDS):

- *ICSU WDS: An Introduction to the Core Trustworthy Data Repositories Requirements*
- *ICSU WDS: Core Trustworthy Data Repositories Requirements*

The Draft Report does not specify particular types or formats of data under consideration, and appears to be mainly considering quantitative and numerical datasets. These are relatively simple to manage. It is not clear whether other datasets such as textual documents (letters, memos, emails, reports), Web sites, social media (facebook, Instagram, twitter), audio recordings, images, video recordings, 3D formats or software, are envisaged as falling within the scope of the NDC.

The NCDiS strongly recommends consultation with expert bodies at each phase. Within Australia, such bodies would include the NCDiS, ANDS, as well as State and National Libraries and Archives. In particular, there exists significant expertise embedded in the university/academic environment, which should be drawn on. Internationally, recommendations and best practice guidelines published in reports by groups such as ICSU CODATA, ICSU World Data System (WDS) and the Research Data Alliance (RDA) should be taken into consideration.

The NCDiS welcomes the opportunity to respond to the Draft Report and looks forward to being involved in future consultations.