Australian National Data Service (ANDS): comments on the Productivity Commission's Draft Report on Data Availability and use.

ANDS supports the overall thrust of Draft Report, which proposes quite fundamental changes to improve the availability (and sharing) of Public and Private Sector data. The research sector has significant dependencies and symbiotic relationships with Public and Private Sector data, and the recommendations of the draft report would reduce friction and increase efficiency considerably; indeed they may even open up totally new methods and areas of research.

The Commission proposes a different set of measures (non-legislative and non-structural) to increase the availability of research sector data, which we believe is appropriate. Our comments below are focused on:

1. **Data for Research** (from this or previous research projects; from the public or private sector)
2. **Data in Research** (under analysis, informal, dynamic, raw, confidential[1]); this class of data are not usually the likely candidates for data publication.
3. **Data from Research** (a subset of the above, quality assured to support findings and to serve as input to further research, and to translation into broader outcomes); ***this class of data are the likely candidates for data publications***[2].

1. Why treat the research sector the same as the public and private sectors?

Research sector data shares the same potential value and is subject to the same commercial, security or other sensitivities[3] as any other data. The research sector is not 'special' in this regard. They inherit a "by default openness" principle from the broad policy framework of NISA[4] and in particular, the Public Data Policy Statement[5] and the extensions proposed by the Productivity Commission. Like public and private sector data, data from the research sector provides significant societal value. It provides open evidence of the workings of research, increased return on investment through re-use, and the basis for an economy-wide translation of benefits including to industry, education, public policy, and health.

The Draft Report contains a comprehensive (Pp 61-187) review of the potential benefits of better access to and reuse of public and private sector data (Chapters 2-5), supported by case

[1] Here we refer to confidential data that has been obtained with permission for a particular research project
[2] Current ARC words on data publications "Since 2007, the ARC has encouraged researchers to deposit data arising from research projects in publicly accessible repositories."
[3] See the ANDS Guide: http://www.ands.org.au/working-with-data/sensitive-data/sharing-sensitive-data
[4] http://apo.org.au/resource/national-innovation-and-science-agenda-welcome-ideas-boom see p.15 "Non-sensitive data will be made openly available by default ... so that the private sector can use and reuse it to create new and innovative products and business models."
[5] "In short, "where possible ensure non-sensitive publicly funded research data is made open for use and reuse" https://www.dpmc.gov.au/sites/default/files/publications/aust_govt_public_data_policy_statement_1.pdf

studies and economic analysis; there can be no doubting of the potential here. But there also many references to the quality of these data, and metadata and a range of other attributes which are necessary for the full range of benefits to be realised.

The same principles apply to the research sector but it would be fair to say that the issue of data 'quality' within the sector is not as well understood or practiced as in the Public Sector, for example. To this end, a set of principles have been developed for the sector by an international community of scholars, librarians, archivists, publishers and research funders, known as FORCE11[6]; The principles are Findable, Accessible, Interoperable and Reusable or FAIR. Another issue is reliability, which is also discussed in Section 2.3 of this submission. This is an essential attribute of data if it is to be fit to support research findings and hypotheses and as well as providing the base of evidence to support translation into innovation, policy, or improved practice.

The Commission's attention is also drawn to the Review of Research Policy and Funding Arrangements[7] (the Watt Review), and particularly the 47 case studies, which make a powerful case for the value of data from the research sector to business and industry. Several other sets of case studies are available from the ANDS website[8]. Collectively, these reports and case studies show the overwhelming potential of data from the research sector to benefit the wider community, especially business and industry, probably more so than the other sectors.

This value is only available if the data from research is routinely available, and of high quality.

2. Why treat the research sector differently from the public sector and private?

On the above basis, the "publicly- funded assets should be publicly available by default" principle should apply squarely to data from the research sector, but, as the Productivity Commission has noted, there are legitimate reasons why not all research sector data can be openly shared:

- **Data in Research,** because researchers may be actively working on the data, generating new data based on that data, using and testing new hypotheses and models; the data may be the outcome of a very large number of experiments so the size does not justify publication. In particular, it is not feasible to make all data in research of sufficiently high quality to make it available.

---

[6] https://www.force11.org/group/fairgroup/fairprinciples

[7] https://www.education.gov.au/review-research-policy-and-funding-arrangements-0

[8] http://www.ands.org.au/working-with-data/publishing-and-reusing-data/data-reuse (green tab left side of page)
http://www.ands.org.au/partners-and-communities/projects/open-research-data-collection

- **Data from Research** because of commercial, security or other sensitivities, which can, and do, apply to all kinds of data, not just research sector data; these limitations on openness would be managed in the same way as in public and private sector data.

Research sector data differ from other kinds of public and private sector data in the way that an individual researcher's (or research group's) intellectual stamp is reflected in the choice of collection methods, observable properties, and subject areas; the data therefore holds an ongoing strategic value for the researcher's career. This needs to be balanced with public good considerations.

Data is also not a tightly defined concept and so what should be shared is not always clear. In the same way that not **all** ideas discussed within a research project end up being published as formal articles, so too not **all** data used in a research project ends up being published as a shareable product. But in both cases there is a reasonable expectation that the important findings and observations of the project are widely disseminated, so that new research can build on old. And given the right reward systems, researchers don't need mandates to publish.

An important element of increasing access to research sector data is therefore "incentive" within the research system. Formal data citations (contributing to impact measures for researchers and organisations), inclusion of data publications as reportable grant outputs, and inclusion of data publications as part of track record for grant applications and promotions are among the practical measures that would go a long way to creating positive feedback systems. We return to incentives in Section 2.1 of this submission in the context of Australian Code for the Responsible Conduct of Research.

Therefore the two specific elements of the research system need to be addressed as special cases for a data availability policy framework:

- **Research sector-specific policy principle**: a default policy setting acknowledging the value for the research system of data products from research projects and creating a responsibility to select and disseminate the data that supports research findings.

- **Incentive**: research funders and research organisations treating data publications as first class outputs for the purposes of scholarly referencing, track record, and reporting.

Inherent in the transition from "Data in Research" to published "Data from Research", there is some selection and quality assurance applied to appropriate data by the research group or research organisation, which has resource implications. Those resource implications are part of the increasing responsibility of burden of proof of research[9] and the changing societal expectations of the re-use of digital objects.

[9] This issue is clearly being recognised by journal publishers, as per section 2.3

2. Comments on section 8.5 Productivity Commission's proposed reforms to open up re-use of research data

2.1 Strengthening government policies

The draft report identifies the *Australian Code for the Responsible Conduct of Research* as a key policy document and discusses potential updates with principles to provide guidance resolving competing considerations.

The Code would be a good place to reflect the considerations of both "Research sector-specific policy principle" and "incentive" (dot points immediately above). Responsible researchers would maintain data that underpins research conclusions and disseminate any valuable data products from research projects; they would formally cite data products used from colleagues, thus contributing to healthy feedback systems.

A revised Code could send the same signal to the research sector (including health and medical research) as is being sent to the other sectors—accompanied by highlighting some of the advantages of data sharing, like attracting collaboration, raising one's profile in the research world or having data publications counted a part of one's track record[10].

The main draft recommendations of the Productivity Commission have the potential to free up the flow of data from the Public and Private sectors; assuming that the main recommendations of the Productivity Commission come into being, and the research sector should benefit from this approach as well.

2.2 Conditions of funding

The draft report contemplates recommending funding agencies to "*require by default* researchers to make their data available". This seems to be similar to the previous section on "strengthening government policies" and would seem to require a similar approach, viz. establish a responsibility to select appropriate data to disseminate; create incentives in the reporting and evaluation systems; and provide principles and guidelines for resolving competing considerations.

[10] Wellcome Trust, NSF, NIH, RCUK all allow 'data publications' to be counted as legitimate research outputs.

The NSF and NIH, for example, allow both publications and data publications to be counted explicitly in applications[11] [12], as do many other international funders. The Wellcome Trust takes into account data sharing practices in final reports when allocating new grants[13].

Australian funders could send a strong message to researchers and research organisations by making relatively minor changes to their reporting and application procedures.

2.3 Building on existing journal publication requirements

As the Draft Report says "... given the plethora of journals based overseas, a more comprehensive and coordinated approach may be difficult to implement from Australia's perspective". There is an international trend emerging of journals requiring 'click-through' access to the data sets which support the publications. In some cases (after some high-profile retractions) the provision of data is mandatory. Examples include: Nature[14], Science[15] and PLOS Medicine[16]. It seems likely that this trend will continue to spread to the publication sector more generally. The following link from Victoria University lists many other journals requiring data deposit: http://guides.library.vu.edu.au/content.php?pid=489543&sid=4015042.

2.4 Institutional issues

The draft report explores the role of institutions in contrast to the role of a more centralised service (extending ANDS). The future state of ANDS and related NCRIS facilities is at the moment under recommendation by the National Research Infrastructure Roadmap, who also call for a more "... integrated national data-intensive infrastructure system".

[11] https://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/gpg_2.jsp#IIC2fic
[12] http://grants.nih.gov/grants/policy/data_sharing/
[13] https://wellcome.ac.uk/sites/default/files/summary-of-phrdf-funder-data-sharing-policies.pdf
[14] http://www.nature.com/authors/policies/availability.html#data
[15] http://www.sciencemag.org/site/feature/contribinfo/prep/gen_info.xhtml#dataavail
[16] http://journals.plos.org/plosmedicine/s/data-availability

3. Comments on specific points raised in the Draft Report

On page p.136 we would recommend a re-wording of the following description of ANDS support for discovery and access:

> **FROM:**
> The central service of ANDS is its Research Data Australia discovery portal, which includes 73,453 records. ANDS does not have a curation role, and data is not directly accessible through the portal. Contact details of data providers are listed, along with information on accessibility. A similar tool is offered by CSIRO, with many datasets featured on both portals.
>
> **TO:**
> ANDS operates Research Data Australia, a national registry and discovery portal of research datasets, which includes entries for 130,000 collections. ANDS does not have a curation role, and data is not directly accessible through the portal. Contact details of data providers are listed, along with information on accessibility and links to download or access services where they exist. Research Data Australia is a national federated catalog, with dataset descriptions “harvested” from over a hundred Australian research organisations, including for example CSIRO which operates both catalog and repository/access services for CSIRO’s own data assets.

There is a reference to the lack of a registry of datasets that result from the funding of Australian funders (p.136) which lead to recommendation 3.2 that “Publicly funded entities, including the Australian Research council, should publish up to date registers of data holdings, including metadata, that they fund or hold”.

The principle of this recommendation is sound; its implied implementation might require some thought since it seems to imply scattered, minimal, and new registers of datasets maintained separately by a long list of government funded research programs. Potentially a more cost-effective and nationally cohesive implementation would be to leverage the existing national register of research datasets (NCRIS funded Research Data Australia) which is already populated by all Australian research organisations. Links from dataset description to funders and grants are also already in place and could be made comprehensive through policy-driven reporting requirements.