Productivity Commission: Issues Paper on Data Availability and Use

In this submission the Australian National Data Service (ANDS) makes the case for recognising the Research Sector as a major provider of data, along with (the other major providers), the Public and Private sectors. As the Terms of Reference are currently framed (TOR 1, 2), the Research Sector is characterised as a user (or recipient) of data from the Public and Private sectors, but not as a donor (or provider). Data that is provided should be of high quality and able to be used as evidence (see d. The reuse value of research data); achieving these qualities has costs but very substantial benefits. The arguments for recognising the value of research data are assembled under the headings of national policies and agendas, case studies, and economics—research data is estimated to be worth $2-6 billion annually in Australia, and accumulating. That said, ANDS cannot understate the importance of TOR 1 (increasing availability of public sector data[1]) to the research sector; research uses—and in some cases depends on—public sector data to develop new ideas, products and methods, which in turn benefits all other sectors.

1. Introduction: the rising importance of research data

In 2015 the Review of Research Policy and Funding Arrangements (a.k.a. the Watt review) included 48 case studies[2] involving 32 universities across medical, agricultural, aerospace, manufacturing, mining, oil and gas and automotive industries. Soon afterwards, the Department of Education and Training published NCRIS case studies[3] in which the accessibility of research data has made significant contributions to university-industry collaborations across multiple sectors (crop, steel, indigenous health, etc). Virtually all of these collaborations relied upon—or heavily involved—research data.

This submission will demonstrate that research data is also at the heart of the National Innovation and Science Agenda[4] (NISA), the Public Data Policy Statement[5] and increasingly, the major research funders, here and overseas. Research data is critical infrastructure to researchers—most research cannot be undertaken without it. This data is also the key to effective collaboration between research and industry—it fosters deep engagement, builds trust, and avoids the problem of traditional outputs (e.g journal articles) being the only information ‘product’ industry normally sees.

Put another way, the ANDS submission is about the proposition that research data (and research itself) will add value to both public and private sector data, and, at the same

[1] Examine the benefits and costs of options for increasing availability of public sector data to other public sector agencies (including between the different levels of government), the private sector, research sector, academics and the community.

[2] https://docs.education.gov.au/system/files/doc/other/20151202_case_studies_volume_nc_0.pdf

[3] https://www.education.gov.au/ncris-case-studies

[4] http://www.innovation.gov.au/page/agenda

[5] https://blog.data.gov.au/news-media/blog/australian-government-public-data-policy-statement

time foster industry engagement, enhanced public policy, education, and better research. TORs 1 and 2 are important in this context but do not satisfactorily capture the benefits of including research data in the overall picture. In this sense, an additional TOR could identify the research sector as the third major provider (and user) of data; were this to happen, there would be three major providers (public and private sector, research sector) without the need to change any of the parties in the users' (or recipients) categories.

Apart from the growing 'physical' evidence, as above, economic studies (later section) have demonstrated that research data has value well outside of business and industry; research data is increasingly being used by government (viz. evidence-based policy), education[6] and the wider community.

2. Different kinds of data (from a research perspective)

In the research and innovation sector, several kinds of data are commonly recognised[7]. ***Research input data*** includes data from other researchers, administrative and government data (e.g. health records, tax data, land parcel information) and private sector data. The research process sometimes generates 'temporary' data during investigation, which is sometimes referred to as ***Research working data***; these kinds of data are not usually retained—they may be erased, overwritten, replaced etc; they could be early runs or tests of models, ideas etc and could be considered as being similar to drafts in the literary world. ***Research output data*** are one of the main outputs of the research process and are the data that support the models, hypotheses or theories; in operational terms, these kinds of data could be considered as the numerical underpinnings of publications. Increasingly, research data is taking on a more generic meaning, to include techniques, methods, algorithms and even software (basically, the non-publication outputs of research).

Repeating an earlier point, research uses—and in some cases depends on—public sector data to develop new ideas, products and methods, which in turn benefits all other sectors. Public sector data is one of the major inputs to the research process.

3. Implications of the current Terms of Reference

The current terms (TOR 1, 2) characterise the research sector as users of public[8] (government) and private data. While this is certainly true, the following sections will demonstrate that data from the research sector has major and increasing value to all other sectors, so much so that research data has been identified in national policy and several national agendas—particularly in relation to business and industry; economic studies have estimated the value of research data to be between $1.9 to $6 billion per year (and accumulating) at current levels of expenditure and activity[9]. A proportion of this value is

[6] http://www.ands.org.au/__data/assets/pdf_file/0008/385019/teaching-with-research-data-report.pdf

[7] These are operational (working) definitions only

[8] The definition of public sector data used in the Productivity Commission paper is 'data collected by Government' which might otherwise be called administrative data, but which would not include research data generated by Government agencies, like the Bureau of Meteorology or Geoscience Australia

[9] http://www.ands.org.au/working-with-data/articulating-the-value-of-open-data/open-research-data-report

research data generated by certain Government agencies (e.g. Geoscience Australia, ABARES, Bureau of Meteorology) or directly funded by Government (CSIRO, RDCs, ACIAR). The balance involves research funded by Government funding agencies like the ARC, NHMRC, Department of Education and usually involving the university sector. In this sense the current TORs seem to be missing that component of Government data which comprises research data being generated within or directly funded by Government. Later in this submission we suggest adding the research sector as a major provider of data (alongside the public and private sectors) and it is important for the Productivity Commission note that this includes Government research data in significant terms. Repeating an earlier point: TOR 1 implies that public sector data in scope is administrative, rather than research data (see footnote 8).

4. Research as a major source of data

a. National policies and agendas

In December 2015 the Australian Government released two related initiatives, the National Science and Innovation Agenda (NISA) and the Public Data Policy Statement[10] (PDPS). PDPS says this about research data: "where possible, ensure non-sensitive publicly funded research data is made open for use and reuse;[11]"

Prior to this, the Australian Government's position on data (specifically meaning both government data and research data) was fragmented across agencies. The underlying premise of the new policies is that data which has been paid for using public money is now to be considered an asset with potential benefits for researchers, business and beyond. The PDPS also recognises those benefits cannot be fully realised without proper data management, standards, licences, repositories and services to ensure the data can be discovered, shared and reused effectively.

NISA has many references to data and the opportunities around its clever reuse. The PDPS recognises the potential for innovation which can only be realised by increasing access to public data, including both the data behind the administrative functions of government as well as the data that comes from publicly funded research.

This quote from the PDPS outlines the importance of the other (non-publication) outputs: "Australia's capacity to remain competitive in the digital economy is contingent upon its ability to harness the value of data."

b. Australian research funders' policies

Australian research policies also highlight the importance of research data as an output of research. The Australian Code for the Responsible Conduct of Research says: "The potential value of the material for further research should also be considered, particularly where the

[10] https://www.dpmc.gov.au/sites/default/files/publications/aust_govt_public_data_policy_statement_1.pdf

[11] https://www.dpmc.gov.au/sites/default/files/publications/aust_govt_public_data_policy_statement_1.pdf

research would be difficult or impossible to repeat."[12] The ARC policy states "The Final Report must outline how data arising from the Project has been made publicly accessible where appropriate" (Discovery Projects Funding Rules for funding commencing in 2014). These policy settings are entirely consistent with NISA which emphasizes the increasing importance of data outputs in research, "A unifying thread across all the domains is the importance of data. Australia's key research challenges will be increasingly data intensive and data driven."[13]

The Australian Research Council has recently announced changes to the funding rules for schemes under the Discovery Program for 2014 and 2015 and similar changes are expected for relevant Linkage Schemes under the National Competitive Grant Program (www.arc.gov.au/ncgp/). These changes will have implications for how research data is managed. The new rules strongly encourage researchers to make their data available: "Researchers and institutions have an obligation to care for and maintain research data in accordance with the Australian Code for the Responsible Conduct of Research (2007). The ARC considers data management planning an important part of the responsible conduct of research and strongly encourages the depositing of data arising from a Project in an appropriate publicly accessible subject and/or institutional repository". [14]

The NHMRC's 2015 data sharing statement[15] includes this quote: "NHMRC encourages data sharing and providing access to data and other research outputs (metadata, analysis code, study protocols, study materials and other collected data) arising from NHMRC supported research. This aligns with researchers' responsibilities under the Australian Code for the Responsible Conduct of Research (2007), which provides advice on the storage, management and privacy of research data (section 2.5-2.7) and states: "Research data should be made available for use by the other researchers unless this is prevented by ethical, privacy or confidentiality matters."

c. Case studies demonstrating the importance of research data

In late 2015 the *Review of Research Policy and Funding Arrangements* (a.k.a. the Watt review) included 48 case studies[16] involving 32 universities across medical, agricultural, aerospace, manufacturing, mining, oil and gas, and automotive industries. Soon afterwards, the Department of Education and Training published NCRIS case studies in which the accessibility of research data has made significant contributions to university-industry collaborations across multiple sectors[17]. Virtually all of these collaborations relied upon—or heavily involved—research data.

[12] Australian Code for the Responsible Conduct of Research, Section 2: Management of Research Data and Primary Materials, Introduction. http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/r39.pdf

[13] 2012 National Research Investment Plan p 61 http://www.innovation.gov.au/research/Documents/NationalResearchInvestmentPlan.pdf

[14] http://www.arc.gov.au/pdf/DP15/Funding%20Rules%20for%20the%20Discovery%20Program.pdf (p.18)

[15] https://www.nhmrc.gov.au/grants-funding/policy/nhmrc-statement-data-sharing

[16] https://docs.education.gov.au/system/files/doc/other/20151202_case_studies_volume_nc_0.pdf

[17] https://www.education.gov.au/ncris-case-studies

In 2015 the The Australian National Data Service gathered representatives from over 30 Australian research institutions to showcase their most recent Open Research Data Collections. These nationally and internationally significant research data collections are described as a series of electronic posters here.

d. The reuse value of research data

Data—whether as input to research or an output of research—can be more or less valuable, depending on a number of attributes involved with the veracity of the data themselves, how they are stored and can be found, as well as other associated attributes. One way of describing the kind of properties that make one data set more'valuable' than another was suggested by an international scholarly data community called Force11[18]. The following attributes are based on their 'FAIR' principles but include additional attributes to deal with the rising issue of reproducibility of science:

- **Findable, accessible, interoperable and reusable (FAIR) research data has more value**: to be valuable, data must be easy to find by either a human or machine, must be clearly described by fit-for-purpose matadata, have clear providence and standards, and so on as described here.
- **Reliable** data has more value: Increasingly, and internationally, the issue of reliability of research data is being linked to research reproducibility, potential usefulness as and an evidential basis for further research, for public policy and increasing potential for utility for research translation.

In the Australian context, the Chief Scientist, heading an expert panel, has been working on National Research Infrastructure Roadmap, which will have far reaching strategic implications for Australia. The following quote is from the recent National Research Infrastructure Capability Issues Paper[19] and demonstrates that these principles are well understood and established within Government.

> *Globally, data is rapidly becoming an important research output in its own right and so a research data output needs to be as reliable and trusted as a research paper. In order to achieve this, Australia needs the infrastructure, processes and policies that will ensure that all research data outputs are reliable, able to be reproduced and able to provide a sounder basis for further research and improved translation outcomes. This infrastructure would be comparable to and be able to leverage international opportunities such as the European Open Science Cloud and other similar global initiatives.*

The European Open Science Cloud initiative is described here and includes a discussion of the importance of reliable research data here.

[18] https://www.force11.org/about

[19] https://docs.education.gov.au/documents/national-research-infrastructure-capability-issues-paper

As the preceding section has shown, to be really valuable, data needs to be findable, accessible, interoperable and reusable (FAIR), and, increasingly, **reliable**. These principles go well beyond what is popularly called 'open' research data. Being open is only one aspect of what is required for data to be really valuable. These principles apply to all categories of data.

e. The economics of research data

In 2014 ANDS commissioned the Victoria Institute of Strategic Economic Studies (VISES) to estimate the value of the data created during the research process, along with an estimate of the benefits of curating and openly sharing public research data. Using conservative methods, the report estimated that the value of data in Australia's public research to be at least $1.9 billion and possibly up to $6 billion a year at current levels of expenditure and activity; this is against a background expenditure of approximately $10 billion per annum[20]. Importantly, the approach used by VISES has been validated and reproduced in many international studies, as shown here (see reference list).

Overall the report shows that a relatively small investment in a combination of data policy and infrastructure provides a significant increase in value to Australian innovation, research, and the broader economy.

A later report[21] attempted to enumerate the value of data infrastructure and services. Based on a well established data centre (European Bioinformatics Institute). It showed that the direct benefits were about 6 times its direct costs — this is measured in terms of researchers' time, based on current users. But the benefits of this facility were much greater than might be judged from the user base alone; benefits of 20 to 50 times the direct costs were calculated, representing the efficiency of having well developed research data infrastructure and services, as well as a measure of the value of research which probably would not have otherwise been done, were it not for the facility.

5. Conclusions

In this submission the Australian National Data Service (ANDS) makes the case for recognising the Research Sector as a major provider of data, along with (the other major providers), the Public and Private sectors. As the Terms of Reference are currently framed (TOR 1, 2), the Research Sector is characterised as a user (or recipient) of data from the Public and Private sectors, but not as a donor (or provider). The data that is provided should be of high quality and able to be used as evidence (see d. The reuse value of research data); achieving this quality has costs but very substantial benefits. The arguments for recognising the value of research data are assembled under the headings of national policies and agendas, case studies, and economics—research data is estimated to be worth $2-6

[20] From budget papers: http://www.industry.gov.au/innovation/reportsandstudies/Pages/SRIBudget.aspx

[21] http://www.beagrie.com/static/resource/EBI-impact-report.pdf

billion annually in Australia, and accumulating. That said, ANDS cannot understate the importance of TOR 1 (increasing availability of public sector data[22]) to the research sector; research uses—and in some cases depends on—public sector data to develop new ideas, products and methods, which in turn benefits all other sectors.

This submission will also show that for data to be really valuable, it needs to be findable, accessible, interoperable and reusable (FAIR), and, increasingly, reliable. These principles go well beyond what is popularly called 'open' research data. Being open is only one aspect of what is required for data to be really valuable.

The contact for this submission is Dr Adrian Burton: adrian.burton@ands.org.au

[22] Examine the benefits and costs of options for increasing availability of public sector data to other public sector agencies (including between the different levels of government), the private sector, research sector, academics and the community.