University of Sydney submission to the Productivity Commission's inquiry into Data Availability and Use, July 2016

Introduction

Data availability, use and management are integral to all aspects of the University's research strategy. Recent and rapid advances in information technology have facilitated an exponential increase in generation and storage of digital data. Accessing and using these data have created enormous opportunities and challenges for researchers today, and will continue to do so for the foreseeable future.

1. The University of Sydney supports enhanced access to public datasets to drive innovation, investigation and research

The data from public datasets form an intrinsic part of the input required for high quality research outcomes, and provide the platform for researchers to work with government to enhance productivity and economic outcomes, support consumers, and improve government policy, services and programs. In many of the University's key research areas, researchers source their data from public datasets. In some studies, these datasets may comprise the sole source of data, while in others they are used to complement and enhance data collected or generated elsewhere.

1.1 What makes data from a public dataset valuable?

Public datasets created by federal and state governments and their agencies are expensive to generate and maintain, and we believe there are strong economic and moral imperatives to maximise the public-good return (research outcomes that feedback to the community) on the investment of public funds.

Public datasets contain large collections of data that are unique; no other commercial or academic organisations have the access, resources, opportunity and permission to compile similar datasets, e.g. census data, health records. As such they represent an enormously valuable asset for research institutions. However, the value of publicly funded datasets is maximised where (de-identified) data are freely available to researchers, where data is updated regularly and where investment in good data management and curation practices ensure that data is FAIR (Findable, Accessible, Interoperable, Reusable[1]). Conversely, their value is greatly diminished if access is difficult or denied, or if the data are not curated well.

1.2 How should public dataset access be improved?

We recognise the need to maintain privacy, security and confidentiality of public-derived data. However, for non-sensitive data existing limitations to access seem unnecessary, for examples Australian soil data, fine-scale ABARES farm survey data, and many transport databases. In many cases, access is permitted only to aggregated data for determining large-scale trends, and not to more detailed individual records, which would enable more powerful analysis. For example. if data from individual farms was able to be accessed, econometrical analyses could be used to plot fine-scale water use across the agricultural landscape that would increase our understanding of the constraints on agricultural production and the efficient use of scarce environmental resources.

---

[1] Wilkinson, M. D. *et al.* 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018. doi: 10.1038/sdata.2016.18

Removing unnecessary barriers to non-sensitive data must be a listed priority for this Inquiry. Other countries have already recognised the need for improved access to public datasets. For example, the UK government 'is committed to making open data an effective engine of economic growth, social wellbeing, political accountability and public service improvement'[2]. Similarly, in 2013, the US government implemented its Federal Open Data Policy stating that 'making information resources accessible, discoverable, and usable by the public can help fuel entrepreneurship, innovation, and scientific discovery'[3], and created the 'data.gov' portal that allows open access to almost 185,000 public datasets[4].

Additionally, we contend that financial costs associated with accessing public datasets should be minimised or waived for non-commercial research institutions, particularly when research is funded from public research funding bodies (NHMRC, ARC). However, we recognise that 'value-added' complex datasets that require advanced data skills/translational expertise could be provided on a cost recovery basis (as is the case with health and welfare datasets managed and collated by the Australian Institute of Health and Welfare[5]).

1.3 How are research outcomes improved?

Increased access to public datasets will result in immeasurable benefit to the Australian community, government and businesses. Enhanced research activity that drives innovation and investigation will result in advances in physical, life and social sciences, and influence and inform policy and improve governance across public institutions.

Many advances will be in the emerging 'big data' space, of which public datasets will play an integral part. An example of this is the work of the University's Centre of Translational Data Science[6], which is utilising extensive datasets combined with machine learning technologies and data science to address multidisciplinary issues in health, education, and technology. A current project that partners researchers from the University's Business School, Centre of Translational Data Science, University of Technology Sydney and NSW Dept of Industry uses data science to devise analytical tools for public datasets to guide resource allocation within the NSW Vocational Education and Training system, ultimately developing a dynamic model for mapping labour demand and targeting public investment in VET. [7]

2. Linking public datasets will improve and increase the effective and efficient use of data and is integral to future research outcomes

Linking datasets from multiple sources can create powerful data networks of far greater research value than their individual components. Creating dataset links is a relatively fast and cost-effective method of maximising the utility of discrete datasets, but is often impeded by difficulties in accessing data from different jurisdictions and agencies (e.g. difficulty accessing Australian soil data collected by each State and combining with national climate data from Bureau of Meteorology). A whole-of-government approach is needed to create a system that facilitates common data structures and the release of data between the states, as the larger the structured data the more effective the analysis.

---

[2] https://www.gov.uk/government/publications/open-data-white-paper-unleashing-the-potential

[3] https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf

[4] http://catalog.data.gov/dataset

[5] http://www.aihw.gov.au/custom-data-request-service/

[6] http://sydney.edu.au/data-science/

[7] http://sydney.edu.au/data-science/projects/health-education.shtml

2.1 Dataset linking services

It is our expectation that rapid advances in data science and technology will allow us to forge increasingly complex and powerful relationships between datasets. This in turn will lead to increasing demand for access to public datasets, and to data services supplied by skilled data managers/curators from federal and state agencies.

We propose that public dataset linking services, such as those offered by AIHW (Australian Institute of Health and Welfare[8]), and DLWA (Data Linkage Western Australia[9]), be extended to service all public agencies and departments that generate and manage datasets that are used in research. For example, in a recent study[10] a University medical research team was able to use three linked public datasets (NSW Perinatal Data Collection, NSW Admitted Patient Data Collection, and Registry of Births, Deaths and Marriages Death Registrations) to investigate the causes of acute gastroenteritis in early childhood.

2.2 Linking Health Records

Many clinical and academic researchers working in health care are frustrated by the lack of access to data and poor linking of health records. Access for researchers is permitted to some anonymised datasets, but not to Electronic Health Records (EHR) for individual patients, even in a de-identified form. While recognising the importance of patient confidentiality and consent, lack of access to EHRs can prevent discovery of better treatment and health outcomes for many diseases as well as prevent the use of computerized clinical decision support for helping individual patients in need of acute care e.g. data mining national cancer patient datasets to determine optimal personalised cancer treatment[11].

Deficiencies in health datasets are compounded by the lack of common identifiers that can link records across all health care datasets (such as using a universal health ID number for all health information, including hospital records). This problem is further compounded by a fragmented healthcare delivery system that not only includes public and private models, but even state-driven public systems are not using federated databases and records are routinely duplicated when public patients are treated in more than one facility. Since the introduction of the national MyHealth Record System (ADHA[12]), uptake amongst different health care providers (hospitals) has been patchy, some healthcare providers (dentists) are excluded from the system and uptake by individual consumers has been poor (opt-in model). For example, oral (dental) and medical clinicians are unable to share crucial healthcare information about mutual patients in their care, resulting in sub-optimal patient outcomes[13,14]. Giving patients

---

[8] http://www.aihw.gov.au/data-linking/how-to-link-data/

[9] http://www.datalinkage-wa.org.au/

[10] Bentley, J.P. *et al*. 2016. Gestational age, mode of birth and breastmilk feeding all influence acute early childhood gastroenteritis: a record-linkage cohort study. *BMC Paediatrics* doi 10.1186/s12887-016-0591-0

[11] http://spectrum.ieee.org/biomedical/diagnostics/big-data-beats-cancer

[12] https://www.digitalhealth.gov.au/

[13] Kalenderian, E., Halamka, J.D. and Spallek, H. 2016. An EHR with Teeth. *Appl. Clinical Inf.*, *7*(2), pp.425-429 doi 10.4338/ACI-2015-09-LE-0124

[14] Hansen, G.M. *et al*. 2016 Relation of Periodontitis to risk of cardiovascular and all-cause mortality (from a Danish nationwide cohort study). *Am. J. Cardio. In press.* Doi 10.1016/j.amjcard.2016.05.036

granular control over who has, and who has not, got access to their health data, combined with a central consent data repository would not only help researchers, but create a safer environment for care delivery with benefits to the individual patient in need.

We propose that the Inquiry specifically addresses impediments to medical/patient dataset linking.

3. The University recognises the value of both public and private sector datasets and builds partnerships/collaborates with data providers

One of the University's key research strengths is building and maintaining successful multidisciplinary collaborations and partnerships with other research institutions, government agencies and private companies. These relationships are crucial to outcomes in many research projects. Project partners commonly contribute dataset access, and these resources are often combined with the University researcher's own data and those sourced from other partner researchers and public datasets (e.g. in the University's Business School[15], and Charles Perkins Centre[16]).

We believe that a priority of the Inquiry should be to investigate ways to encourage partnerships and collaborations between academic researchers and commercial/industry partners that facilitate better linking of public and private datasets. For example, medical researchers and dieticians investigating causes of increase in the incidence of type II diabetes in Australian children could access food supply chain data (if made available from private companies) to analyse the population's changes in diet. We suggest that government has a role in 'brokering' such public-private partnerships to promote collaborations that would benefit the community.

4. The University Is committed to an Open Access policy for data generated by its researchers

The University recognises the high value of its research outcomes and has an Open Access policy that encourages researchers to share their completed datasets (or metadata where restricted or mediated access to data is appropriate). Many of the University's researchers contribute datasets or metadata to the University's repository and the national data registry Research Data Australia[17], thereby building, enhancing and sharing valuable re-usable data resources with all researchers. Many researchers also contribute to and use datasets in discipline-specific spheres, such as genome sequences to Genbank (National Centre for Biotechnology Information, US National Library of Medicine[18]). Increasingly, when publishing research in journals, authors are required to upload datasets to repositories with open access where possible, and the major research publishers now maintain indices where individual researchers are rated by the quality and re-usability of their open access data.

Researchers in open access data-rich environments have unprecedented opportunities to collate and manipulate datasets to produce powerful models and tools that are of enormous value to the community and the economy (e.g. EarthByte global geophysical

[15] http://sydney.edu.au/business/business_analytics/research

[16] http://sydney.edu.au/perkins/about/our-partnerships.shtml

[17] https://researchdata.ands.org.au/contributors/the-university-of-sydney

[18] http://www.ncbi.nlm.nih.gov/

and geological data visualization[19], TERN eMAST terrestrial ecosystem modelling data in NCI's national data store[20]). Public-facing research consortia that have a major commitment to open data can grow through the recruitment of others with related research interests through a process of "forced serendipity" (e.g. the ARC-funded Open Source Malaria consortium based at the University[21], which posts experimental results in real-time thus enabling researchers from around the world to instantly access data). Initiatives such as these produce value-added resources from open, well-curated data that have broad applications across many disciplines and benefitting the wider community.

We recommend that the Inquiry considers the emerging role of open access datasets in creating a data-rich environment for multiple beneficial uses.

5. The University maintains the privacy, security and confidentiality of the datasets its researchers use

While committed to sharing data through Open Access datasets where possible, the University recognises restrictions to access are necessary in many situations, including for personal, confidential or commercially sensitive data. However, in many cases data may still be made available to researchers if it is aggregated and/or anonymised, and it is here that perhaps some of the greatest improvements in data management policy can be made.

After consulting with our researchers, three main concerns emerged around the issue of access to sensitive data. We urge the Inquiry to address these concerns:

1) Restrictions to accessing individual health records by medical/clinical researchers due to 'automatic' patient consent and privacy restrictions. We need a single mechanism that allows consenting patients to grant access to their records for research purposes, with the option for data to be de-identified/anonymised if preferred.
2) Research institutions carry potentially unlimited liability for any breach of confidentiality associated with the use of sensitive data.
3) Re-identification of previously anonymised data sets can never be 100% prevented as technology advances rapidly providing previously unforeseeable linkages. Thus a legal framework is required that makes any attempt at re-identification a criminal offence that will discourage re-identification and protect the public and researchers.

[19] Müller RD, Qin X, Sandwell DT, Dutkiewicz A, Williams SE, Flament N, et al. (2016) The GPlates Portal: Cloud-Based Interactive 3D Visualization of Global Geophysical and Geological Data in a Web Browser. PLoS ONE 11(3): e0150883. doi:10.1371/journal.pone.0150883

[20] http://dap.nci.org.au

[21] http://www.ands.org.au/news-and-events/share-newsletter/share-23/open-source-approach-puts-malaria-on-notice