UNIVERSITY OF TASMANIA RESPONSE TO THE PRODUCTIVITY COMMISSION INQUIRY INTO DATA AVAILABILITY AND USE

The University of Tasmania welcomes the opportunity to provide a submission to the Productivity Commission's Inquiry into Data Availability and Use (the Inquiry). Universities stand in a unique position in terms of new policy that may increase access to data. They are both creators and consumers of data. Data is fundamental to research. This position enables the Higher Education sector to understand fully both the implications and benefits of improving data availability and use.

Universities create, collect and record data for both research and administrative purposes. The Inquiry's recommendations will therefore need to take account of the differing purposes for which data is collected.

Research data

In terms of research data, University of Tasmania researchers have created, and/or have enabled access to large data sets in many different disciplines including diverse areas such as:

- The Tasmanian Data Linkage Unit which is based at the Menzies Institute for Medical Research http://www.menzies.utas.edu.au/research/research-centres/data-linkage-unit and is a node of the Population Health Research Network http://phrn.org.au/ funded by the National Collaborative Research Infrastructure Strategy (NCRIS);

- The Institute for Marine and Antarctic Studies (IMAS) which maintains a significant catalogue of marine research data derived from studies conducted by University researchers. This catalogue is publically available online through the Metadata Catalogue and the IMAS Data Portal.

  IMAS also operates facilities and host datasets on behalf of others in the national interest, often with international focus and scope (eg. Reef Life Survey, TemperateReefBase, Redmap and the soon to be released Seamap Australia);

- Sense-T http://www.sense-t.org.au/ - a partnership between the University, CSIRO and the Tasmanian Government which is a user of big data, sensing technologies and data analytics to create opportunities, improve decision making and realise productivity growth in a variety of industries including agriculture, aquaculture, viticulture and tourism; and

- In the humanities, the Centre for Colonialism and its Aftermath (CAIA) where the records of 75, 000 convicts are being used to study the effects of intergenerational health issues.

In these and other large data sets, it is the ability to link individual data records that provides answers to research and policy questions. Such data linkages must occur with regard to data management codes and legislation, and require both appropriate infrastructure and trained personnel.

In regards to research data management, the University subscribes to the data management principles outlined in the Australian Code for Responsible Conduct of Research and agreements with funding agencies such as the Australian Research Council (ARC), National Health and Medical Research Council (NHMRC), state and federal governments and others. Many of the University's large research data collectors and users have adopted additional data management policies in keeping with norms and requirements of international associations and discipline groups.

In terms of research data infrastructure, the University also has found the NCRIS-funded Australian National Data Service (ANDS) to be an essential mechanism for making data more "discoverable" through Research Data Australia and providing resources to facilitate data sharing. Of particular relevance to this Inquiry will be the importance of ensuring continued funding for this service. This will be even more important as the requirement by journals and funding bodies for researchers to share data will undoubtedly be a major driver for behaviour change.

In terms of a trained, data literate workforce, we strongly support the position outlined in the Universities Australia response that *'the Productivity Commission consider strategies to harness existing talent and increase the knowledge, skills and abilities of the Australian workforce to nurture and sustain a digital economy and develop an evidence based and data oriented culture."* In this regard we welcome the programs designed to enhance STEM literacy through the National Innovation and Science Agenda and the mooted increased flexibility in researcher training. However, incorporating additional elements into researcher training (such as increased data literacy) will be challenging and may be difficult to achieve under existing funding structures.

The Inquiry asks for specific comment regarding efficiencies in managing the costs of data availability and use. In regard to the ongoing collection and use of research data we would suggest the following principles:

- Support for continuing national programs like NCRIS which ensure a consistent, universal approach to all research infrastructure;
- Support for a simple funding model vs a user pays model. The exponential growth of data threatens to reach a ceiling of current funding levels and the introduction of 'user pays' systems are likely to be necessary but will require critical analysis [level, dimensionality, subscription model] to review the need to pump prime this model given the inherent overhead in administering funding;
- Adherence to international standards and best practices for metadata management and technical data delivery services to ensure a high level of interoperability between disparate sources of data and reduce the need for costly 'one-off' data processing and delivery tasks; and
- Re-use of existing technology. There is a proliferation of software infrastructure for managing data, including within specific disciplines. Ensuring the re-use of existing technologies, particularly open source where a contribution can be made to maintaining and improving functionality, will significantly reduce the costs associated with optimising data use.

Administrative data

The changes proposed in the Inquiry will have significant implications for the administrative systems in operation at Universities. The challenges arise from the integration and linking of various data attributes, given that students and staff are administered across numerous systems. The use of hosted, cloud-based, or software as service systems is also increasing. In turn, integration costs are greater and the process more complex. Whilst storage and communication costs per byte are decreasing, management and retrieval costs are growing.

Further issues arise from the requirements of privacy legislation, which has not kept pace with the rate of technological change. Universities have been left to fund increasing and exponentially growing costs of managing rights to privacy, information and security. This must not be the case with new policy to open up access to data.

The changes proposed in the Inquiry are likely therefore, to increase information costs in Higher Education.

Consequently, adequate funding will need to be provided to help public and private organisations meet the costs of open data access. Legislative changes should be accompanied by a detailed and rigorous examination of the costs of compliance and associated funding allocations made to the sector.

In this respect we lend strong support to the Universities Australia response to the Inquiry with its statement that the Productivity Commission consider "*how to create a legislative and regulatory environment that supports greater access to and use of data, that specifically considers the aspects of intellectual property with a view to providing clarity with respect to the ownership and use of data (and transformations thereof) that offers more clarity without increasing the barriers to the use of data for both research and commercial purposes*."

Specific questions raised in the Inquiry

As noted in the introduction to this response, the University is a major partner in Sense-T which has substantial experience in dealing with private industry and their data. The following responses to specific questions from the Inquiry arise from the experience of Sense-T researchers and their industry.

What are the main factors currently stopping government agencies from making their data available?

Key factors for the government groups that we have observed in our relationships are:

- The lack of infrastructure available:
    - The infrastructure must be reliable;
    - Groups creating data are not experts in the infrastructure to distribute their data;
    - The groups creating the data do not have the resources to support infrastructure or help desk operations to support the users of the data if use is scaled;
    - Effective infrastructure must support human browsing of the data and also be machine readable but the experience we have is that often that only one aspect has been addressed. This may be because a "visual" interface used by humans is easier to demonstrate to gain support in an organization but in contrast the sharing and re-use of coded information requires the data to be machine readable; and
    - The infrastructure must support the characteristics of the data and the delivery and distribution of that data. For example, aspects like real-time sensing data that is being automatically collected on a constant basis is not currently supported on the Australian Government Open Data websites that are based on the CKAN system. The absence or alternative standards for meta-data on scientific data is also a challenge when interpreting and re-using data collected by other organisations.
- The lack of discovery of the data:
    - A common experience we have is that the available public data is not discoverable, even when publicly available. In many cases data may be available but there are no public records linking or detailing where to find that data, instead the details have to be obtained from a person.
- The cost of the data preparation and lack of method to determine the fit between the users potentially interested in the data and the group controlling that data.

Should the collection, sharing and release of public sector data be standardised?

The metadata describing the data should be standardized. An example in scientific data is the standardization of the metadata format for units of measure. The technical openness of data through standards for access to the data should also be standardized.

Are there any legislative or other impediments that may be unnecessarily restricting the availability and use of private sector data?

In our experience private industry is concerned that the Right to Information (RTI) Acts will allow unintended access to the sensitive data they provide to universities or government agencies for the purposes of research.

Should these impediments be reduced or removed?

In our experience a single Federal Right to Information (RTI) Act would be better than the separate State Acts that currently exist, as this would remove the variations that occur between the States.

What are the reasonable concerns that businesses have about increasing the availability of their data?

The most common concerns from our experience are:

- Businesses are familiar with keeping and controlling the data within their own organization. They are concerned that providing data will expose trade secrets and operational processes to their competitors. They still have some concerns that they will be unable to correctly determine how much needs to be confidential to avoid competitors gaining advantage from their data even if they can exclude confidential details from a shared version. They are also uncertain what techniques might be available to their competitors to discover useful details if their data are combined with data from other sources to find larger patterns;
- They are concerned that data breaches are not detected by data distribution providers or could be left unreported to them when they happen. This may be an expression of the general awareness in the community that security breaches are happening more often than they are reported;
- There is a general concern that making data available means they may lose “ownership” of data. Generally explicit agreement is required to transfer contractual control of data but the concepts of “ownership” for data are not well understood in our experience and concerns will remain. For example, we have heard legal opinion that while the format of data may be protected the “data” itself as facts cannot be protected by copyright law or intellectual property law in Australia. This thesis requires further testing within a suitable legal framework; and
- Inherently the global nature of data distribution leads to the uncertainty for business in meeting their responsibilities under applicable legal frameworks.

What principles, protocols or legislative requirements could manage the concerns of private sector data owners about increasing the availability of their data?

The following privacy and confidentiality principles are taken from the Sense-T Privacy White Paper. A copy can be obtained by contacting Dr Paul Neumeyer

Data Management Principles

- P1 Privacy by design
- P2 Data is an asset to share
- P3 Data usability through open standards
- P4 Data integrity and attribution by design
- P5 Data protection through security and monitoring
- P6 Federated data is preferred over duplication

Confidentiality Principles

- C1: Disclose only if authorization given
- C2: Authorized level of detail
- C3: In accordance with legal and regulatory frameworks
- C4: Inform stakeholders on Confidentiality and Privacy

What would standards that are 'fit for purpose' look like?

A standard for "fit for purpose" would require sufficient mandatory meta data so that the data can be tested against the requirements of the data user. The fitness for a purpose will always have specific requirements for that purpose and the users of the data for that purpose are best able to confirm the fitness if they have sufficient meta data.

A proscriptive standard dictating the characteristics of the data itself is unlikely to be widely successful. A standard that sets certain minimum or maximum characteristics is likely to put unnecessary burden on data collectors to meet the purpose of all potential uses, limit the amount of data available as existing data may not meet the standard and because new types of data and new uses will arise, a proscriptive standard will become outdated.

To what extent can voluntary data sharing arrangements — between businesses / between businesses and consumers / involving third party intermediaries — improve outcomes for the availability and use of private data?

We have experience with voluntary data sharing arrangements with our industry partners and we have been involved in setting up voluntary agreements that have improved the availability and use of private data for research. There are some aspects that require mandatory data sharing of private data such as for regulatory requirements and where data is shared with others because of a legal requirement.

We expect the availability and use of private data will naturally increase where the value to business to share data is sufficiently high and the risks are low. The extent that business as a generalized group will make data available is likely to increase as they see other businesses make data available and can quantify the risks they currently perceive. Businesses will not want to make data available that puts them at a competitive disadvantage which is where voluntary data sharing arrangements are a better fit to encourage the availability of private data than mandatory data sharing requirements for business.

Who should have the ownership rights to data that is generated by individuals but collected by businesses?

In our experience a factor useful to differentiate the control of data for individuals is when the individual is generating material covered by copyright.

For example, when an individual generates photos, written opinions or other creative works they have the role of data controller over the copyright data they have generated. If they provide that data to a business they are still the data controller by default, but could transfer contractual control via an agreement with the business.

When a business (observer) is observing an individual (for example, how the individual interacts with the business) if the individual provides photos, written opinion or other creative works what those contain, or the details of a particular behaviour then the individual is the data subject and the business is the data controller of the collected data.

An individual (observer) can observe another individual, for example, how the individual interacts with the observer, the observer can take a photo which includes the individual, or the details of a particular behaviour then, like in the case of a business being the observer, the individual is the data subject and the observer is the data controller of the collected data.

A concept to give a limit on observation is that individuals and organization have private domains, which is a domain around an individual or organisation which includes all the things that are a part of it and that includes seclusion, limited accessibility or the ability to control information flow[1]. As observation techniques and technology becomes increasing sophisticated the concept of a private domain maintains a human perspective on what is a reasonable expectation on observation. For example, an individual will expect their home is a private domain and will not be under observation there. Something that was closely related to an individual as a private domain can change depending on how it is put it into a public context. For example, an individual's mobile phone, even though it is a private domain, if discarded in a public rubbish bin does not remain private.

Concluding comment

In summary, we welcome the opportunity to provide input to the Productivity Commission's Inquiry into data availability and use. Data is becoming a key economic asset and the ability to capture, store, make available and link data within a secure environment will be a driver of innovation and productivity. Universities are at the cutting edge of "big data" being both creators and users of data. We therefore take this opportunity to reinforce key messages to the Commission, namely:

- Continued federal support for data storage facilities such as ANDS and the RDDS will be essential to ensure data availability and use;
- There is a need for Australia's legislative and regulatory environment to support greater access to and use of data;
- The need for strategies to build a data literate workforce, with particular attention paid to the role and funding of universities in this process; and
- The potential for increased costs to universities that are likely to arise from requirements to make administrative data more readily available.

Questions concerning the University's response should be directed in the first instance to Professor Brigid Heywood, DVC Research,

[1] Yael Onn, et al., *Privacy in the Digital Environment* , Haifa Center of Law & Technology, (2005)