CSIRO Submission 16/560

Data Availability and Use

Productivity Commission

July 2016

Enquiries should be addressed to:

Dave Williams
Executive Director CSIRO
PO Box 52, North Ryde, NSW 1670

Main Submission Author:

David Chapman
CSIRO Data61

Table of Contents

Table of Contents 2
Executive Summary 3
Introduction 7
CSIRO response to the Productivity Commission Issues Paper on Data Availability and Use 8
1. Availability of public sector data (ToR 1) 8
1.1 Benefits and issues for increasing public sector data availably 8
1.1.1 Issues and impacts of increased data availability 9
1.1.2 Capability and Technology - Use 15
1.1.3 Promote long-term vision to standardise data through open data policies 15
1.1.4 Data Sharing 16
1.1.5 Data Management 18
1.1.6 Public registries as data services 19
1.1.7 Data community 20
1.1.8 Information supply chains 20
1.2 Emerging technologies to enhance responsible data availability and use 22
1.3 Examples of high value public sector data sets 24
1.4 Dataset and Education 28
1.5 Datasets provided by CSIRO 29
2. Availability of private sector data (ToR 2) 31
2.1 Issues in availability of private sector data and benefits of addressing these 31
2.2 Examples of valuable private sector data sets 31
2.3 Legislation and other impediments 32
3. Individuals' access to public and private sector data (ToR 3) 33
3.1 Datasets provided by CSIRO 33
4. Standardising, sharing and release of public and private sector data (ToR4) 34
4.1 Standardisation of data 34
4.2 Data sharing 35
4.3 Duplication in data collection efforts 37
4.4 Storage and publication of data 37
4.5 Data ecosystem 38
4.6 Confidentialisation in facilitating data sharing while protecting privacy 39
5. Enhance and maintain individuals' and businesses' confidence and trust in the way data are used (ToR5) 40
References 42
Appendix 1 – EUDM Work Findings 43
Appendix 2 – Data about Australian Businesses 45

Executive Summary

Increasing access to data, while managing potential privacy risks, will enhance consumer outcomes, better inform decision making, and facilitate greater efficiency and innovation in the financial system and the broader economy.

Whilst the size and nature of benefit differs depending on the analysis approach taken and the cases analysed, enormous value can be created in the Australian economy through responsible (open) access to public service data. Unlocking the potential of this data to improve policy and service delivery will differentially affect industry, the public service, academia and citizens. Publishing public sector data will enable innovation in the following ways[1]:

- Create markets for new/better services
- Enable benchmarking and innovation
- Identify customer/citizen needs through more effective analysis
- Improve and automate decision-making
- Provide transparency and accountability
- Reduce cost through reduction in duplication
- Potentially increase productivity

The key issues to be addressed include:

- increasing the quantity and quality of public service data sets and services which could lead to high-value outcomes;
- data standardisation, collation and data infrastructure;
- sharing data within the Commonwealth and across jurisdictions;
- data privacy preservation;
- transparent communication with the public to build trust;
- addressing the data literacy skills gap to collate and use data and;
- clear interpretation of related legislation

The availability of government and industry data is essential for many areas of CSIRO's research and innovation, involving both intensive and extensive use of data collected by many different agencies. CSIRO is both a significant generator and user of data, across a variety of disciplines.

Increased availability of data for research is necessary, however to achieve productivity gains integrated with requirements there should be a clear assessment of 'fit-for-purpose' curation of data for the end user needs, as well as ensuring that accessibility is simple and straightforward for the end user.

Release of public sector datasets relating to Australian business and other non-government organisations should be an important priority as these are of high potential value. Creating greater access to information and data will also be an invaluable resource for research and innovation, creating new services and insights on business activities in Australia. It will also reduce duplication of data sourcing and management thereby leading to a significant reduction in costs.

A policy framework relating to publication of publicly funded research data will need to address a number of legal issues, including managing the conditions of access to data and potential liability issues flowing from third party use of data.

There is a need for appropriate open data policies promulgated throughout all agencies of government (including CSIRO) and from that would flow organisational decisions in relation to funding data as

[1] https://www.dpmc.gov.au/sites/default/files/publications/public_sector_data_mgt_project.pdf

infrastructure. The Atlas of Living Australia (ALA) provides solid examples of the benefits and impact of making data open and available. The value of data improves when it can be integrated with other data within an environment that supports data discovery, visualisation and analysis. The success and achievements of the ALA have help to drive a culture shift in terms of data sharing.

Access to open government data in Australia is economically important, as confirmed by multiple theoretical and empirical studies, with varying estimates of its net positive benefit[2]. We are therefore supportive of an Open Data Policy where the basic principle is that data generated with public funding for widespread community use purposes should be made open and freely available. However, careful consideration of potential ethical, privacy issues, requirements in relation to intellectual property or under statute in relation to the data, and any other limitations must be taken into account, and this may result in time delays for release, and for certain types of data it may not be possible to achieve immediate release of data.

Open Data is only one component of the culture shift that needs to be achieved as we embrace the Digital Transformation. We need to promote open source software, open infrastructure, open services (including Application Programming Interfaces (API)) and a culture that embraces collaboration rather than the creation of data/infrastructure silos.

While we support the issues paper call for standardising, sharing and release of data (where for the most part there are existing protocols already in place), we would like to see additional emphasis on the creation of data. This should include new methods of data capture, “cradle to grave” data management policies, and rapid digitisation programs to bring physical information into the digital world.

A key recommendation is to develop a long-term funding model for managing and distributing public and private data and better coordination across institutions and infrastructure (recommendation 3). This long-term funding for existing data custodians will allow a federated model for data custodianship where the experts in the data have responsibility for keeping the data of high quality and being up to date as well as being a trusted third party. This allows data to be discoverable and explorable via a common access point (data.gov.au as outlined in the Australian Government Public Data Policy Statement, 7 December 2015[3]). This requires funding for both data.gov.au (which should continue to be overseen by PM&C due to its importance for government, industry and citizens) and funding for the data custodians to ensure they maintain high quality and up-to-date data. The team in government responsible for data.gov.au should also be responsible for promoting data standards, protocols for data use and facilitating data sharing.

We note that government has a process underway in determining its National Research Capability Infrastructure Roadmap[4] which should be considering the long term sustainable provision of data services for research purposes within Australia. We recommend that the development of public data infrastructure, such as coordination with data.gov.au, is included as a priority.

Our recommendations for Government to increase access to public and private sector data to boost innovation and competition are:

Recommendation 1: Continue to recognise and promote the significant contribution publically acquired data makes to the economy.

[2] Bureau of communications Research, *Open government data and why it matters*, February 2015 https://www.communications.gov.au/sites/g/files/net301/f/BCR%20Open%20government%20data%20report%20-%20final%20-%205%20Feb2016.pdf

[3] Australian government Public Data Policy Statement – 7 December 2015 https://www.dpmc.gov.au/sites/default/files/publications/aust_govt_public_data_policy_statement_1.pdf

[4] http://www.chiefscientist.gov.au/2016/07/national-research-infrastructure-roadmap-national-research-infrastructure-capability-issues-paper-released/

Recommendation 2: Continue its focus on public data availability as described in the Australian Government Public Data Policy Statement released by the Prime Minister on 7 December 2015.

Recommendation 3: Develop long-term funding models for managing and distributing public data. This would include providing support for government agencies in making available high quality, well-maintained, standards-conformant data services and Application Programming Interfaces (APIs) (in addition to periodic release of individual data files). This would also include ongoing funding for data.gov.au as public data infrastructure.

Recommendation 4: Develop a culture that sees data management as an investment rather than an overhead to overcome the issue that significant portions of the cost lie with those that acquire and share data rather than those that use it.

Recommendation 5: Clearly document the quality, collection methods and intent of public data sets, such that they can be used within an appropriate context and with sufficient understanding of potential biases and flaws.

Recommendation 6: Retain public data sets of high value to the public, industry and research sectors which provide 'baseline information', ensuring it is available to future users to save significant setup costs for projects and provide incentives to digitise and mobilise legacy data that significantly contribute to establishing baselines.

Recommendation 7: Assist in closing the skills and knowledge gap between data release and data use through greater promotion of government data services, provision of sample applications and greater level of developer documentation and support.

Recommendation 8: Reduce the friction in accessing information in registries, to enable new commercial value-added services, so as to multiply the broader economic benefits supported by government registries.

Recommendation 9: Recognise that new distributed privacy-preserving approaches to encrypted data analytics can allow for data to be joined and used for analysis without the data contents being viewable by any party (including the party doing the analytics) and consider relevant changes to legislation (including privacy laws) to distinguish data availability and use in a privacy-preserving mode from conventional data availability and use.

Recommendation 10: Align public and private data activities to achieve efficiencies and increase availability of private data, by making the guidelines and standards for public data available and enabling provision of de-identified datasets.

Recommendation 11: Play a catalytic role for the public sector to provide training and access to platforms, protocols and processes that facilitate the private sector to use and share public data and platforms (with appropriate protections in place) with the option of value adding and marketing that data to aggregators who have a demand for it.

Recommendation 12: Enable communities of practice as cohesive authorising environments for standards development, connected with international communities – addressing the social and technical aspects of data and information management.

Recommendation 13: Standardise data to enable combining of disparate datasets to develop multi-variable models, which will require development of a model for funding the processing of data to a standardised format at the level of institutions acquiring data.

Recommendation 14: CSIRO supports the principle of all data generated with public funding for widespread community use purposes being made open and freely available, noting that careful consideration of potential ethical, privacy issues, requirements in relation to intellectual property or under statute in relation to the data, and any other limitations must be taken into account, and this may result in time delays for release and for certain types of data it may not be possible to achieve immediate release of data.

Introduction

CSIRO welcomes the opportunity to provide input to the Productivity Commission's Inquiry into Data Availability and Use. Australia's future success is heavily dependent on our national ability to engage and lead in the use of data in all segments of the economy and to inform government policy and program delivery. The availability and use of data is essential for CSIRO's delivery of research and innovation services.

CSIRO has world class capabilities and delivers rigorous and comprehensive research in data and digital innovation, from data analytics and data network platforms, to collecting, managing, disseminating and using extensive amounts of data collected by CSIRO and other agencies. CSIRO is a major provider of data, data products, analysis and infrastructure, servicing industry, government, the science community and the public.

CSIRO is a significant generator and user of data in a variety of domains (agriculture, mining, energy, environment, health and biosecurity, astronomy). We are experienced with creating, managing and delivering exceptionally high volumes of data. In the astronomy domain, this is particularly evidenced in the successful development of the Australian Square Kilometre Array Pathfinder (ASKAP) radio telescope and for our aspirations to deliver the Square Kilometre Area (SKA) where the big data processing and analytics capability we have developed with our partners in the Pawsey Supercomputer is essential.

CSIRO's response to the inquiry draws on our broad range of scientific expertise. Our comments have been prepared by a team of scientists, engineers and specialists from across CSIRO with experience and international recognition in many facets of data research, creation, use and management. We would like to acknowledge input to this submission from CSIRO's research capability in agriculture, land and water, oceans, atmosphere, energy, astronomy, health and biosecurity, information management and technology, education, digital innovation and in national facilities, with the Atlas of Living Australia (ALA) a key example of a successful open data platform delivered by CSIRO on behalf of the nation.

Our submission addresses each of the Terms of Reference, providing background information, commenting on the state of the science, significant issues and knowledge gaps, and the priorities to address the issues paper. This submission is focussed on sections where CSIRO has undertaken research that is published or in the public domain.

CSIRO welcomes any opportunity to discuss any areas in more depth with the Committee.

CSIRO response to the Productivity Commission Issues Paper on Data Availability and Use

1. *Availability of public sector data (ToR 1)*

Given the enormous potential value in making public sector data available, we recommend that the Australian Government:

- **Recommendation 1:** Continue to recognise and promote the significant contribution publically acquired data makes to the economy.
- **Recommendation 2:** Continue its focus on public data availability as described in the Australian Government Public Data Policy Statement released by the Prime Minister on 7 December 2015.
- **Recommendation 3:** Develop long-term funding models for managing and distributing public data This would include providing support for government agencies in making available high quality, well-maintained, standards-conformant data services and Application Programming Interfaces (APIs) (in addition to periodic release of individual data files). This would also include ongoing funding for data.gov.au as public data infrastructure.
- **Recommendation 4:** Develop a culture that sees data management as an investment rather than an overhead to overcome the issue that significant portions of the cost lie with those that acquire and share data rather than those that use it.
- **Recommendation 5:** Clearly document the quality, collection methods and intent of public data sets, such that they can be used within an appropriate context and with sufficient understanding of potential biases and flaws.
- **Recommendation 6:** Retain public data sets of high value to the public, industry and research sectors which provide 'baseline information', ensuring it is available to future users to save significant setup costs for projects and provide incentives to digitise and mobilise legacy data that significantly contribute to establishing baselines.
- **Recommendation 7:** Assist in closing the skills and knowledge gap between data release and data use through greater promotion of government data services, provision of sample applications and greater level of developer documentation and support.
- **Recommendation 8:** Reduce the friction in accessing information in registries, to enable new commercial value-added services, so as to multiply the broader economic benefits supported by government registries.
- **Recommendation 9:** Recognise that new distributed privacy-preserving approaches to encrypted data analytics can allow for data to be joined and used for analysis without the data contents being viewable by any party (including the party doing the analytics) and consider relevant changes to legislation (including privacy laws) to distinguish data availability and use in a privacy-preserving mode from conventional data availability and use.

1.1 Benefits and issues for increasing public sector data availably

Whilst the size and nature of benefit differs depending on the analysis approach taken and the cases analysed, enormous value can be created in the Australian economy through responsible open access to public service data. Unlocking the potential of this data to improve policy and service delivery will differentially affect industry, the public service, academia and citizens. Publishing public sector data will enable innovation in the following ways[5]:

- Create markets for new/better services
- Enable benchmarking and innovation

[5] https://www.dpmc.gov.au/sites/default/files/publications/public_sector_data_mgt_project.pdf

- Identify customer/citizen needs through more effective analysis
- Improve and automate decision-making
- Provide transparency and accountability
- Reduce cost through reduction in duplication
- Potentially increase productivity

The key issues to be addressed include:

- increasing the quantity and quality of public service data sets and services which could lead to high-value outcomes;
- data standardisation, collation and data infrastructure;
- sharing data within the Commonwealth and across jurisdictions;
- data privacy preservation;
- transparent communication with the public to build trust;
- addressing the data literacy skills gap to collate and use data and;
- clear interpretation of related legislation

1.1.1 Issues and impacts of increased data availability

The availability of government data is essential for many areas of CSIRO's research and innovation, involving both intensive and extensive use of data collected by many different agencies.

Health and Medical Data

Data availability for health and medical purposes – both primary and secondary use – is vital for patient care, health service efficiency and improvement and for medical research. The use of health data falls into two categories – primary and secondary. Primary use refers to the use of health information for the direct treatment of the patient. Secondary care refers to the use of the data for purposes other than direct patient care – such as reporting, quality improvement, medical research etc.

Australia has a history of valuable health data collections at places such as the Australian Institute of Health and Welfare, our state and federal health jurisdictions and through specific studies. Much of this data until now has been "Administrative" in nature, requiring researchers to infer clinical utility from the data. Current Electronic Medical Record and Electronic Health Record initiatives mean that more clinical data is being captured, which if made available, would greatly increase the ability of Australia's medical research community.

Registries are an important part of Australia's health data landscape. At a state level, health jurisdictions are required to maintain various registries for public health, such as state based cancer registries. In addition, various clinical groups have developed disease specific registries, such as the trauma registry or prostate registry. In the case of mandated registries these typically contain a minimum data set and it is a legal requirement to submit this information. In the case of clinical registries, these are typically more detailed but are not mandated and will not capture all cases in Australia.

The linking of data from different data collections to these registries can add significant value. In the case of the cancer registries, the linking of treatment and outcome data provides a more useful set of data for clinical research.

Public Sector data about Australian businesses and other organisations

Release of public sector datasets relating to Australian businesses and other non-government organisations should be an important priority as these are of high potential value. These datasets are essential for accountability and transparency on corporate activities, as well as providing greater information symmetry for the efficient operation of markets. Creating greater access to this information will also be an invaluable resource for research and innovation, creating new services and insights on business activities in Australia.

The datasets relating to Australian businesses and other non-government organisations covers four types of activities. These include:

- Regulatory: where a government agency or authority collects and manages information about an entity for regulatory purposes.
- Procurement: where a government agency or authority collects information about an entity for the purposes of purchasing goods and services.
- Grants: where a government agency or authority collects information about an entity for the awarding and managing grants.
- Tax and Tax Offsets: where a government agency or authority collects information about an entity for the purposes of collecting and administrating taxes and tax offsets.

CSIRO is implementing a number of projects that could make use of the business data that has been identified above through use of open data complemented with other data sources to conduct analysis to generate improved data-driven insights on business and research activities. Some examples of relevant projects include:

- *Ribit* - (www.ribit.net) is an online platform to connect students to jobs, and universities to industry. Its initial focus is on matching students to meaningful work engagements including part-time roles, specific projects and internships.[6] The intention is to extend this to support the matching of researchers and businesses for short and longer-term collaboration, creating faster, simpler, easier connections across the entire innovation ecosystem. The platform directly addresses the low level of business and research collaboration in Australia, as well as the ambition of the university sector to introduce mandatory Work Integration Learning (WIL) requirements for all undergraduate students. The capability of Ribit to provide optimal matching between students, researchers and businesses will be supported by the use of open data regarding businesses and research activities to inform algorithms for appropriate recommendations.

- *Discovery of innovative companies* - A project to support the Ribit platform is the discovery of innovative companies using open data from both the public and private sectors. The objective is to identify relevant companies for matching to students and researchers using Ribit. It could also be used to help benchmark the digital maturity and innovative capacity of companies, thus better informing individual companies as well as improving the targeting of government research and innovation support programs.

Other researchers and companies have identified other opportunities and potential benefits from using open data relevant to Australian businesses and related organisations. The Lateral Economics Report, *Open for Business: How Open Data Can Help Achieve the G20 Growth Target*, has identified how governments could release data on employee job satisfaction and related employment information. Releasing such data could help people identify the best-matched workplaces to their skills and aspirations, thus potentially improving job matching, job satisfaction and productivity. The report notes that this example could generate large, economy-wide productivity gains for Australia (approximately $3.4 billion per annum over the first five years or 0.22% of GDP).[7]

Data about companies held by governments has also been used to create new services focused on promoting greater transparency about the business sector. A leading example is Open Corporates, a United Kingdom based service that uses company registry data from each country to create a global open reference website

[6] About Us, Ribit website: www.ribit.net

[7] Open for Business: How Open Data Can Help Achieve the G20 Growth Target
A Lateral Economics report commissioned by Omidyar Network, June 2014
www.omidyar.com/sites/default/files/file.../ON%20Report_061114_FNL.pdf

about companies. Open Corporates then integrates this data and seeks to create greater understanding about ownership and large corporate groups.[8] This value-added data is now being used by a range of other organisations such as LinkedIn, World Bank and various commercial credit and business advisory companies.

Easy access to spatial data

The NationalMap (http://nationalmap.gov.au) initiative of the Department of Prime Minister and Cabinet provides easy access to more than 4000 data sets and data services operated by more than 30 government agencies at all levels of government. It is free to use and can be used from any web browser, smartphone or tablet. The NationalMap was developed by CSIRO Data61 and uses the TerriaJS software also developed by Data61. It is now being used for providing mapping systems in other countries (e.g. by the US Geological Survey and other groups in the US, France and Latin America).

NationalMap uses a federated model and accesses data directly from the data custodian. This ensures that the data is always up-to-date and that it remains under the control of the data custodian. The NationalMap makes use of the data catalogue at data.gov.au and uses available government agency APIs for government data when available (including direct access to ABS APIs for demographic data).

The NationalMap uses open data, open protocols and open APIs. NationalMap software and the underlying TerriaJS software are both open source. This ensures that there is a growing framework of open data services that can be used by NationalMap and other mapping systems. For example the Australian Renewable Energy Mapping Infrastructure (AREMI) http://nationalmap.gov.au/renewables, has been developed by CSIRO Data61 for ARENA and is an important resource for the renewable energy industry including financiers, developers and policy makers. This uses many of the same data services as NationalMap and uses the same underlying open source software. There is a growing network of open data services and different themed mapping front-ends for these, including a map of resources in Northern Australia developed for Austrade (http://nationalmap.gov.au/northernaustralia) and the National Environmental Information Infrastructure (developed by the Bureau of Meteorology).

Some advantages of this approach to an open framework include:

- Low cost of the development of new sites: The cost of the development of a new themed government interactive mapping site is low through the separation of data service availability and map user interface availability. Themed maps such as those for agriculture, energy, infrastructure, etc. can be developed leveraging the data services already available in NationalMap.
- Innovation by the private sector: Many companies have adopted use of the data services and/or TerriaJS in their commercial products and several start-ups have developed products using these (e.g. Propeller Aerobotics, a Sydney company developing a cloud platform for transforming drone-captured imagery into 3D models and allowing tracking of construction, mining, etc. over time).
- Greater government data sharing: Government agencies can make better use of spatial data from other agencies through a single easy-to-use public data infrastructure.

Through the National Innovation and Science Agenda, CSIRO Data61 are working with PM&C to integrate NationalMap further with data.gov.au and to expand this to being an advanced public data infrastructure. The initial version of the new data.gov.au will be released in late 2016.

There are opportunities for integrating NationalMap with the Australian Geoscience Data Cube (AGDC) which provides access to earth observation satellite data including 40 years of Landsat imagery as well as imagery from other satellites.

[8] Open Corporates website, https://opencorporates.com/

The AGDC has come out of a partnership between CSIRO, Geoscience Australia and the National Computation Infrastructure (NCI). When combined with the long-historical and up-to-date earth observation satellite data from the US and Europe, sitting on public computational infrastructure such as the NCI in the form the AGDC, the access to valuable satellite data processing services and visualised through the NationalMap is greatly facilitated to many more users from research to industry. Developed as open-source code, it is already attracting interest and being implemented by the US Geological Survey (USGS) for use across the continental US, as well as by NASA who in collaboration with CSIRO is installing pilot "DataCubes" in Kenya and Colombia, to demonstrate how this Australian innovation is making satellite data much more freely accessible and useful to developing countries and international aid agencies interested in food-security, disaster response and climate & forestry applications.

The ease of use of NationalMap and the richness of the AGDC data and the related analytics (such as determining frequency of water coverage for any property in Australia) would provide a highly interactive and powerful platform for many Australian industries and citizens. This has applications in agriculture (such as for salinity detection over time), environmental management, resources, development, infrastructure planning and many other areas. The maturity of the NationalMap platform and the AGDC are both at a level that integration would be achievable and powerful and likely to lead to widespread adoption throughout industry and government to make better use of the rich satellite data that is available.

Data relevant to natural hazards and emergency management

Increased availability of data for research in the areas of natural hazards, emergency management, adaptation and infrastructure planning and risk assessment is necessary. However to achieve productivity gains integrated with these requirements there should be a clear assessment of "fit-for-purpose' curation of data for the end user needs, as well as ensuring that accessibility is simple and straightforward for the end user. The current practice generally in this space include:

- Data being available but in disparate locations
- Data being available but not fit-for-purpose. For example for flood modelling in urban environments one requires very high resolution terrain data (of the order of 1-2 metres, see Prakash et al., 2015) whereas for fire modelling (see Miller et al., 2015) the requirements are less stringent from a terrain perspective (of the order of 30-90 metres).
- The people responsible for preparing and curating datasets are generally big data specialists using IT infrastructure who do not have the necessary appreciation of end user needs. Close collaboration is therefore needed between data creators/curators/users to achieve desired productivity outcomes.
- Local councils and other planning authorities spend significant amounts of money to just curate and analyse the data for their needs. This investment could be utilised for infrastructure instead if tools were provided to these agencies to enable well curated and reliable data using uniform standards imposed throughout Australia.
- Widely varying data quality for different states. For example we have found through our projects that Victoria and Queensland have the best quality data in the country. Why is this not universally deployed in a nationwide approach?

This underpins the need for national collaborative bodies to work through the various community practices toward achieving census and broad agreement on standards to apply across jurisdictions, with open data sharing as the goal. An example of such an attempt supported by the Commonwealth government is the Essential Environmental Measures program. The first community under development is around vegetation. Two working groups are looking into 1) the problem of interoperability for sharing vegetation site data; and 2) observations and measures about vegetation that are essential inputs to a system understanding of dynamics and indicators of change; and therefore essential for ongoing monitoring against baselines. Rather than the issue being "Why is this not universally deployed in a nationwide approach" we should consider how to enable open collaboration to ensure communities of practice can discuss and work toward agreed best practice, and so forth.

Environmental Data

CSIRO is both a significant generator and user of environmental data, across all environmental disciplines. The data used for environmental research depends upon the nature of the research and typically integrates across social, cultural, political, economic and environmental sources of information. While it is not possible to nominate individual data sets that are more important than others, a priority list of environmental data sets was developed by the Australian Government – Environmental Information Advisory Group (AG-EIAG 2012), including input from CSIRO.

Environmental data sets are retrieved from wherever reliable sources can be found and access negotiated. Thus a mix of public and private data sets might be used for a particular project. The Bioregional Assessments Programme (to which CSIRO is the largest capability supplier) sources data from many public and private organisations. One of the major challenges for environmental researchers is the discovery of data for their work.

Remotely sensed data (typically satellite based imagery) is an increasingly important data source for environmental research. These data are typically sourced from public agencies via long term relationships.

Climate data sets are also important. Biophysical research into ecosystems uses historical climate time series and projections to understand and simulate the impacts of climate change. Individual researchers, in the absence of a consolidated view, will value-add to improve the relevance of interpolated climate data, for example, to more accurately represent interactions with land form by taking into account variation in exposure to radiation on hill slopes and associated ambient temperatures relative to a flat surface. Communities of practice are needed to increase efficiencies in spatial environmental data creation and use of best available science in their generation.

A key challenge for environmental researchers is the lack of definitive data sets representing the features of Australia's environment. That is, there does not exist national authoritative data sets of features such as streams, transport networks, etc. In particular there are no agreed and well governed identifiers for these features. The Foundation Spatial Data Framework (FSDF) initiative is seeking to address this issue (http://www.anzlic.gov.au/foundation-spatial-data-framework).

It is often the case that the outputs of environmental research must meet the requirements of open data. Data sourced for research that is restrictive in its use and so prevents open publication of derived products may no longer be viable for use.

Environmental research is often focussed on understanding the implications of particular decisions or policy, and how these interact with natural biophysical system processes. In this context, the most current information around the ownership and use of land, and all if its natural resources and capacity for recovery, is particularly important. In a future that is increasingly challenged by global change, more specific and dynamic information about how land is utilised or managed is important for developing understandings of practices likely to sustain future social and economic development.

Initiatives like the Australian Ocean Data Network (AODN) provide access to all available Australian marine and climate science data and IMOS data and metadata. Within the Research Provide Networks, it brings together the Heads of Research and Fisheries agencies from all jurisdictions to set consistent metadata standards, to improve data management and to share within the marine domain.

The AODN is increasingly relied upon by the marine science community for large offshore projects and aims to improve access to and integrated use of spatial data and information, to support decision making.

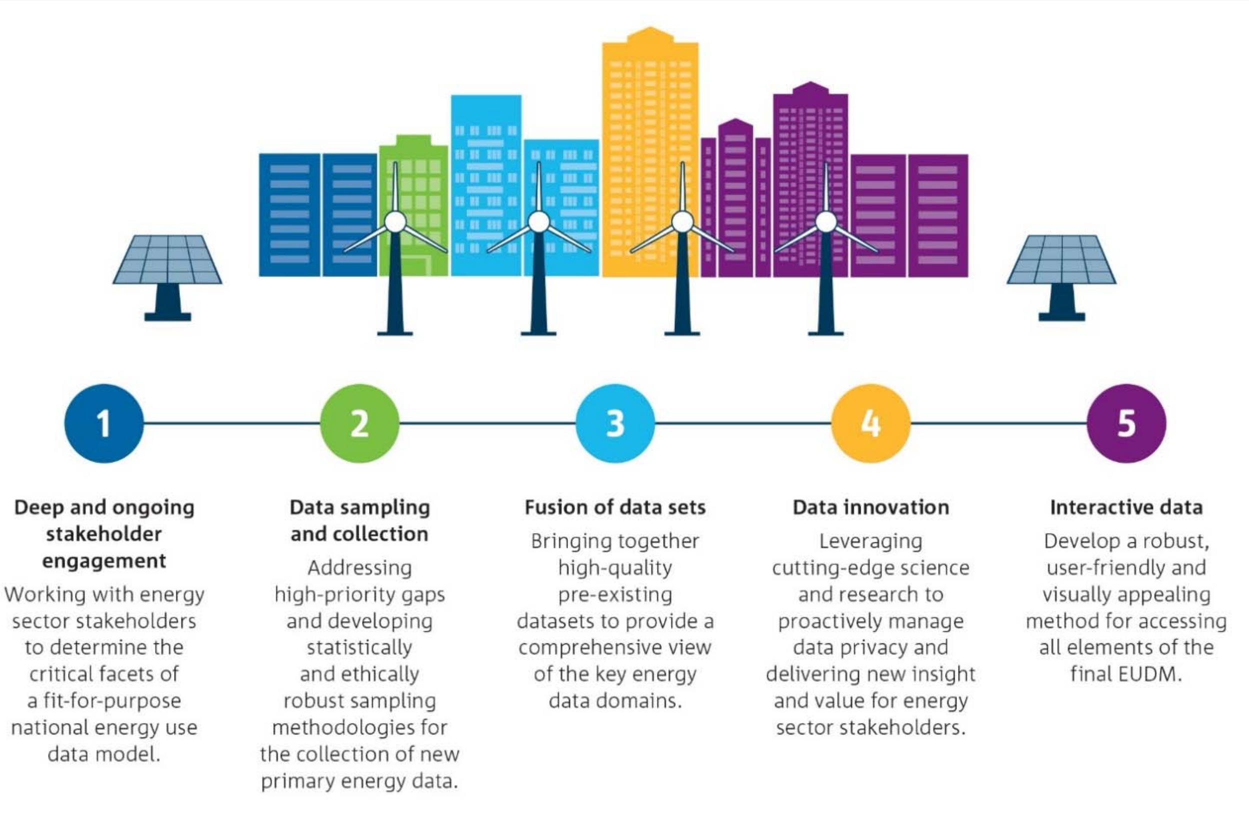
The National Marine Science Plan (NMSP) 2015-2025 notes that increased accessibility to data, regardless of who paid for the collection, improved delivery of information to both the private and public sectors, a national approach to data sharing and accessibility and developing high-performing data services to synthesise, analyse and access data are all integral to realising the opportunities offered by the Blue Economy.

Case Study: Energy Data

Australia's energy sector is seeing new technologies emerge and end-user behaviour change as markets deregulate and diversify. The predictability and planning assumptions about the nature and future of Australia's energy systems are beginning to break down as the sector innovates, adapts and disrupts. Without rich energy-related data to capture the effect of these changes on individuals, regions and the nation, it is almost impossible to forecast infrastructure requirements or optimise new business models. Equally difficult is the task of developing effective policies or fully understanding the customers that the sector aims to serve.

CSIRO and the Commonwealth Department of Industry, Innovation and Science are responding to the need for such data by developing the Energy-Use Data Model (EUDM). We are developing a data platform that takes end-user (researcher, government, consultant and data-holder) needs into account, and focuses on making high-quality energy-use datasets available to the energy industry. The data held in EUDM will provide the information required to develop evidence-based solutions to the critical issues facing the energy sector.

Building a meaningful platform for a rich energy data solution is a multi-faceted task. In EUDM, the goal is achieved by delivering work across five parallel streams (see figure 1). Together, these streams connect stakeholder needs, primary research, software development and data into a complete programme that is focussed on delivering a concrete product that is both innovative and fit-for-purpose


1
Deep and ongoing stakeholder engagement
Working with energy sector stakeholders to determine the critical facets of a fit-for-purpose national energy use data model.
2
Data sampling and collection
Addressing high-priority gaps and developing statistically and ethically robust sampling methodologies for the collection of new primary energy data.
3
Fusion of data sets
Bringing together high-quality pre-existing datasets to provide a comprehensive view of the key energy data domains.
4
Data innovation
Leveraging cutting-edge science and research to proactively manage data privacy and delivering new insight and value for energy sector stakeholders.
5
Interactive data
Develop a robust, user-friendly and visually appealing method for accessing all elements of the final EUDM.

The recent EUMD findings provides insight relevant to both Government and private datasets and is pinned to contemporary stakeholder engagement which are reflected in **Appendix 1**.

1.1.2 Capability and Technology - Use

In general, public sector data (such as environmental data) is routinely and regularly made available to the general public and is done so using appropriate open licensing. This is the case across federal, state and research sectors with most organisations having a stated direction or declaration regarding open data. A key gap exists, however, in the ability of end users to consume many of the complex data services which are produced by data providers. The focus on data interoperability standards (e.g. Open Geospatial Consortium, OGC, standards) has resulted in relatively complex services which require advanced knowledge to use effectively. These services are often developed and delivered on the assumption that end users can actually make use of what is built – an assumption that "build it and they will come". However, the uptake of these datasets and services into the decision making chain is often lacking, meaning that the benefits of open data are not being realised.

Decision makers want "tools" rather than raw data services, but often lack the specialist capabilities or staff to build these tools. In addition, the software industry has not been successfully incorporated into the open data journey, meaning that it is difficult to acquire the right skills to utilise complex data services. As a result, there is often a skills and knowledge gap between data release and data use which has limited the dispersed benefits of releasing the data in the first place.

Health related data is generally made available under strict ethics approval and through rules associated with the collection of that data. For instance, state health data is generally made available under the Public Health Act of each jurisdiction. The data made available is of variable quality – with some data collections consisting of high quality data with strong meta-data including data dictionaries. Other data collections are of low quality. With the investment into e-Health around Australia we have the opportunity to mandate vendors on the quality and format of their data collections, which would greatly increase the quality of the data available for sharing without greatly increase the cost.

1.1.3 Promote long-term vision to standardise data through open data policies

There is a need for open data policies promulgated throughout all agencies of government (including CSIRO) and from that would flow organisational commitments to fund data as infrastructure. At a whole community level, invest in ongoing quasi-independent facilities for example driven from NCRIS to ensure baseline data and ongoing environmental monitoring and data aggregation services (of whatever form – centralised or decentralised business models as appropriate) are continued.

As discussed above, mandating the use of standards for data collection at the point of care in healthcare software will greatly increase the quality of data available for both primary and secondary data use.

The climate, ocean, environmental and agricultural domains receive significant benefit from the consumption of open data, one leading example being satellite remote sensing data from global providers, of which none is directly collected by Australia. Significant cost and effort must be put, however, into developing locally relevant products from these international sources. From an Australian context, the delivery of the global data is not enough, local products will provide a greater benefit to end users.

Much of the national infrastructure related to public sector data collection, processing and delivery is supported through short-term funding from activities such as NCRIS or from internal capital budgets rather than any longer term operational activity. This uncertainty surrounding short-term budgets means that the capital expenditure is often not leveraged to its full extent or risks being lost once the initial investment has been made. Due to this, there is a perception that much of this data infrastructure is ephemeral and end users are subsequently wary of investing too much time making use of it. Additionally, many of the short-term funded repositories do not have plans in place to ensure the longevity of data if their funding is withdrawn. Key is to explicitly build in the development of long-term governance and custodianship models from project inception – this addresses (up-front) the risk of data graveyards, loss of dataset expertise and data loss.

In order to develop a culture that sees data management as an investment rather than an overhead it is proposed to develop a model to overcome the issue that significant portions of the cost lie with those that acquire and share data rather than those who use it by developing long-term funding models for managing and distributing public data to reduce the risk to repositories and the data they contain and recognising the contribution publically acquired data make to the economy. One aspect of the model is the measurement of data reuse for recognition and potentially reward.

1.1.4 Data Sharing

Morton and Tinney (2012), in a review of Australian government environmental information activities, note that:

Australian governments, industry and communities need comprehensive, trusted and timely environmental information to help ensure government policies and programs are properly evidence based. Decisions with an environmental component – either impacting the environment or dealing with an impact of the environment on society or the economy – are made daily by individuals, businesses and all levels of government.

This same review identified five themes of obstacles to efficient and effective use of environmental information. These are:

1. Cultural – characterised by a lack of communication between providers and users of environmental data;
2. Structural – characterised by a lack of coordination both with and between agencies maintaining environmental data sets;
3. Funding arrangements – a lack of coordination resulting in duplicated investment and missed opportunities;
4. Technical barriers – a lack of standards in some areas and a lack of use of standards where they do exist; and
5. Legal barriers – primarily licensing restrictions.

Morton and Tinney proposed a set of recommendations to address each of these themes to ensure Australia has an environmental information system that is responsive to policy and that uses and reuses the information generated by its agencies efficiently and effectively.

The Morton and Tinney report also addressed challenges created by the reality that many hundreds of organisations (public and private, federal, state and local) collect and manage information about the environment for various purposes. Apart from a few specific areas (e.g. water, biodiversity) there is no single organisation with a role to collate/aggregate complex environmental information on behalf of the community, although the Atlas of Living Australia has a mandate to fulfil this gap for biodiversity and environmental data. As a result, the need to share environmental information in order to undertake various activities (e.g. develop environmental policy, environmental management and research) is very necessary.

Open Data is only one component of the culture shift that needs to be achieved as we embrace the Digital Transformation. We need to promote open code, open infrastructure, open services and a culture that embraces collaboration rather than the creation of data/infrastructure silos.
Examples of benefits here, taken from the Atlas of Living Australia, include:

- The core ALA infrastructure now supports over 70 different hubs and portals delivering to a wide variety of outcome areas (biodiversity, biosecurity, citizen science). This is a key example of the benefit of open infrastructure – which is as important (if not more important) a story than open data.

- ALA infrastructure is now being used by 7 different countries around the world to support their own national biodiversity portals – with several more actively implementing the ALA tools.
- ALA web services (all completely open) are being used by various groups to support their own initiatives (e.g. NECTAR labs such as the Biodiversity and Climate Change Virtual Laboratory; QuestaGame – a game app to support biodiversity learning).

Within the environmental data sector, the most significant challenge with sharing data is that the costs, both the initial setup costs and ongoing costs, are primarily borne by the custodians of the data, whereas the benefits accrue to the end user. The traditional market based approach to solving this problem would be for data custodians to charge a fee for data access. However, this approach, has proven challenging and, in some cases counterproductive with respect to data sharing. As such, there appears little incentive for custodians of data to share that data in any way other than that which minimises the cost to them. Hence the need for government intervention to support data sharing in the environment domain.

Unfortunately, this approach of minimising cost greatly impacts the ability of potential users to discover, access and use the data for their purposes, leading to inefficiencies including replication of data generation processes. For many, in particular the research community, these tasks of data discovery and access are essential and represent a significant portion of the cost of research. These issues are so large and pervasive that they require further exploration, below.

Discovering, accessing and using data.

For a data set to be successfully shared, it must be discoverable, accessible and usable by those it is being shared with.

Discovering data

Within the environmental data community, a number of initiatives have been undertaken to create catalogues of data sets. These are often volunteer based or have limited funding and there is rarely any interoperability between these initiatives. As a result, those looking for data have little confidence that they will find what they are looking for, or that what they find is the most current or best practice, authoritative version.

Ideally, those looking for data should be able to:

- Be briefed or trained in the value or principles of managing information
- Use the discovery tool of their choice
- Be confident that their search will be comprehensive
- Determine fitness for purpose as part of their search

Accessing data

The challenge of accessing data has many facets. Not only does a potential user need to understand the processes of accessing data, which can often be ambiguous, they also need to have tools and, in some cases, the authority to do so.

The process to access data involves:

- The technical process of connecting tools to data.
- The process of negotiating often complex and restrictive licence agreements. Arrangements that are too restrictive can lead to potentially useful data being unable to be used for research and other (e.g. commercial) purposes.

Using data

Using the data in many cases represents the largest, yet often least understood, cost to a potential user. This cost is the time taken for the potential user to understand the data and then manipulate it into a form that is usable for their purpose. It includes:

- Understanding the vocabularies, semantics and concepts of the data set
- Understanding the full context in which the data were generated/captured
- Undertaking manipulation activities.

Many of the issues raised above are also issues with the sharing of data in healthcare.

The findings from recent EUDM work provides insight relevant to both Government and private datasets and is pinned to contemporary stakeholder engagement. A series of workshops across Australia with participants from across research, government and industry sectors focused on building an understanding of energy-data users, the role of energy data in their work, and the issues that are holding back progress in energy-data research. The themes and key findings from these workshops are reflected in **Appendix 1**.

1.1.5 Data Management

Despite recognition that data is a valuable resource, data management is still often perceived as an overhead rather than an investment. Yet many projects undertaken by the public sector would be unaffordable if archival datasets were not available and would have to be reacquired. Significant investment in data management expertise is required to enable the availability and linking of data, yet this is an area which tends to be under-resourced.

The Marine National Facility, through the CSIRO Oceans and Atmosphere Information and Data Centre, has ensured that data acquired through the Marine National Facility is consistently quality-controlled. Metadata is developed according to ISO19115 and made available in a machine-readable format through internet data portals with a CC-BY license.

Data has a lifecycle and it would be advantageous for the Commission to review a process to plan data management from inception. This could draw on considerable work on Data Management Planning that has already been undertaken in agencies including CSIRO and the National Archives of Australia such as the CSIRO Data Access Portal and CSIRO Astronomy Science Data Archive.

This would also involve the establishment of provenance for public data re-use. Provenance would also assist the assessment of the inputs of public data into innovation, as would the assignment of Digital Object Identifiers (DOI) which would enhance the ability to cite the data in publications. The establishment of feeds to tools such as the Data Citation Index would then allow it to be measured.

The Commission can draw on the work that ANDS has done in the research sector to determine what can be leveraged for managing public sector data. Particularly it can review the work that ANDS has done internationally, specifically with Research Data Alliance. This has taken the approach of addressing specific issues across geographic and jurisdictional boundaries. It encompasses standards, description, interoperability, security, licensing, attribution and more.

Effective management and use of public data requires increased focus on the associated human capital. This includes appropriate workforce expansion, career path establishment and course development/evolution in specialist data science, analytics and information services. It also includes training to broaden familiarity, literacy and numeracy for effective data management, across the communities that generate and use data.

Much of public data is poorly described and organised, and management of it is poorly resourced. If availability is to be improved to increase re-use and potential innovation, this will either need to be resourced or require some innovative approaches to automated capture, description and discovery. Part of this effort will also be to identify roles: owners, custodians, curators, infrastructure providers. There is already work in a number of agencies, such as Geoscience Australia, that can be leveraged here. Additionally there is work required to establish national vocabularies for data description. One example would be a definitive vocabulary for the description and changes to the names of government departments and agencies.

1.1.6 Public registries as data services

Just as the economy benefits from the release of public data, the economy benefits from the open availability of government registry services. Registry services can be seen as authoritative data mappings (e.g. from an ABN to a business name or entity name). Where the government has responsibility for assigning this mapping, it makes most sense for the government to have control over the authoritative data service that provides online lookups using these data services. The operation of authoritative, open registries can be seen as a core function of government.

The cost of running government registries (particularly in modern cloud deployments) is immaterial compared to the economic activity they support. If considering charging for use of such data lookup services, the government should carefully consider whether doing so will disincentive use and so work against the value the registry can provide to the economy and citizens. Instead of increasing costs to users, governments should consider lowering the friction of accessing information in registries, to enable new commercial value-added services, so as to multiply the broader economic benefits supported by government registries.

Equally important is the issue of regulatory agility. Agility is all about change. Agility is not just about change for today – it is a stance that keeps us ready for future changes. For government registries, the most critical emerging changes are to respond to new technology opportunities and the evolving policy context. Technological advances are enabling opportunities for rapid near-term innovation for government registries. Much of this opportunity derives not from new high-risk technologies, but from ideas about new applications of existing technologies. For registries, two examples of emerging opportunities are to link registries across domains and across jurisdictions, and to automate interfaces and access controls to reduce the friction of using registries and their information.

- Registries are often inter-linked. For example, companies are identified in a business name registry, and might then be referred to from many other industry licensing registries. However, currently these registries are usually maintained separately, and so may not be updated consistently. This degrades data quality, hurts regulatory effectiveness, and increases risk in those regulated industries. All of this degrades the broader economic benefit of the registries. With technology we can identify registry relationships, improve current levels of data quality across linked registries, and automate the consistent update of multiple registries. It should be noted that the challenges here are not expected to be primarily technological – instead the challenge to integrate government information is a policy challenge, across multiple departmental or jurisdictional silos. If the government were ever to consider outsourcing individual registries to different commercial bodies, this would only exacerbate these policy challenges.
- Access to registries today is through an interface which is either paper-based (or over-the-counter), through a web-based GUI, or through a programmatic API. There may be a variety of privacy or confidentiality policies that apply to these registries, depending on the registry and its industrial context. These policies determine the authentication and authorisation conditions to update or view the information in the registry. These access control conditions should apply equally regardless of the form of the interface used.

Another driver for regulatory agility is broader policy change. For example, technology adoption within society has and will continue to change community expectations around privacy and confidentiality. Government policy will need to change in response. If the government were ever to consider outsourcing registries, this would cement government policy, contractually. Many government policies on the operation of registries can currently be changed administratively, without legislation. If these registries were ever outsourced, government policy changes for those registries are likely to require additional cost and time arising from contract re-negotiation. This would reduce the ability of governments to effectively regulate the operation of registries.

1.1.7 Data community

We recognise the important work of the Data Champions group in Federal Government which includes key senior people with data responsibilities in key government agencies. We think that it would also be worthwhile to establish a Whole of Government data community of practice (CoP), for example by leveraging the Science Agencies Data Stewardship Working Group. Members include:

- Geoscience Australia (host Agency)
- Bureau of Meteorology
- Prime Minister & Cabinet
- Australian Bureau of Statistics
- CSIRO
- Murray Darling Basin Authority
- Defence Science Technology Organisation
- Australian National data Service
- Department of Environment
- Department of Health

1.1.8 Information supply chains

Box *et al* (2015), when studying information supply chains within the geospatial data community, identified many challenges to information supply that are common with those found in the environmental data and possibly other data domains. They stated that to find a solution, it will be necessary to *address a range of interwoven technical and social challenges caused by the fragmented and heterogeneous production, management, supply and governance of data across multiple levels of government.*

The most critical challenge to be addressed within the geospatial data domain is the need to integrate a patchwork of data sources with different structures and semantics, developed under different business contexts, into a coherent suite of maintainable products. This challenge is largely a function of the federated government structures in which geospatial data production and delivery takes place across all levels of government.

These issues play out at the level of the individual researcher, where a lack of leadership and training in the benefits of sharing data and information becomes an embedded cultural constraint on open release of data. New incentives to encourage an open data culture at all levels of governance and use are required.

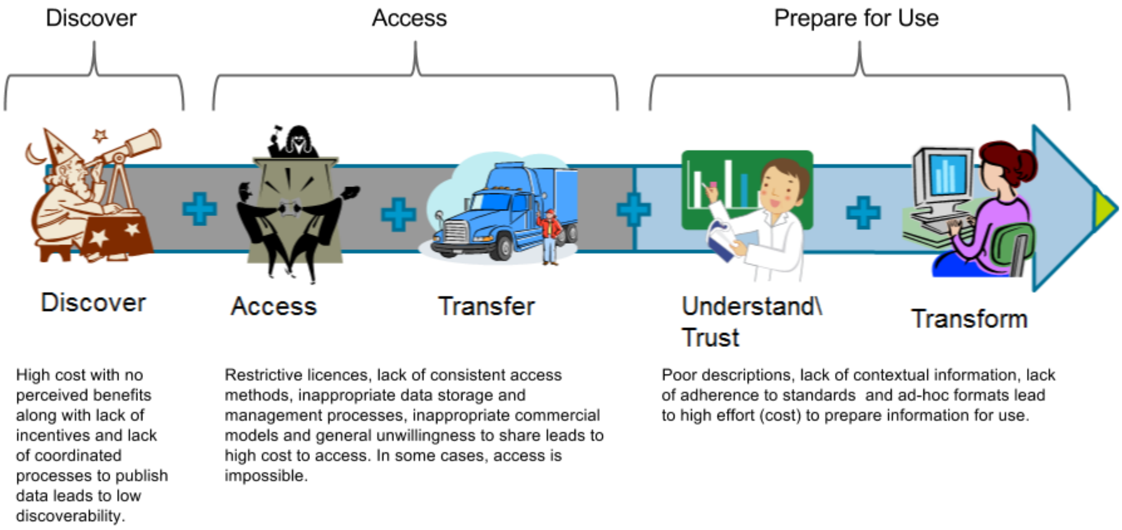
Data is a key input into research. As such, the challenges of inefficient information supply chains are felt acutely within the innovation sector, particularly when conducting trans/inter/multi-disciplinary research.

The key to transforming information supply chains is to recognise that they are socio-technical in nature. As such, the barriers and risks to be addressed are a mix of social and technical concerns and relate to such things as:

- **Social context** – attitudes to data sharing, ethics and privacy
- **Existing agreements -** policy regulation, legislation, licensing and standards
- **Institutions -** institutional arrangements and organisational and group behaviours

- **Economics –** the economics of information, return on investment for infrastructure and funding models
- **Technology -** implementation patterns, technology standards, service interfaces
- **Information content -** semantics, description and structure

These barriers are encountered at different points in the supply chain, but to solve the challenges of data availability and access, each must be addressed. The figure below describes the key steps and provides examples of typical barriers as they apply to an information supply chain. Note that not all barriers exist in all supply chains.



High cost with no perceived benefits along with lack of incentives and lack of coordinated processes to publish data leads to low discoverability.

Restrictive licences, lack of consistent access methods, inappropriate data storage and management processes, inappropriate commercial models and general unwillingness to share leads to high cost to access. In some cases, access is impossible.

Poor descriptions, lack of contextual information, lack of adherence to standards and ad-hoc formats lead to high effort (cost) to prepare information for use.

A central issue is that significant portions of the *cost* of providing discoverability, accessibility and usability of a data asset to a level where the asset can be re-used or repurposed, sits with those that share it (the owners/producers). Whereas, the potential *benefits* lie with users. As a result, owners/producers seek to minimise these costs with a subsequent impact on discoverability, accessibility and usability. Users are then faced with the challenge of either spending extra resources to meet their data needs, or potentially compromise the outputs of their project by using limited data. Either way, the end result is that less value is generated than might otherwise have been.

This pattern is more problematic when an information supply chain spans more than one organisation or business unit as the reconciliation of costs with benefits is more difficult.

Currently, the challenges of information supply chains are being addressed independently within specific information domains or initiatives. These initiatives range in scale, scope and maturity. Importantly, there is largely no coordination across these initiatives and solutions typically cannot be easily applied in other contexts.

Public data must have a very real value proposition for the data provider. There are multiple opportunities here: providing pathways for publish-once, access-many models that reduce the burden on data holders who must manage frequent (ad-hoc) requests; maximise opportunities for automated linkage to other public datasets on publication (instantly increasing the value of the original set to the data holder); openly acknowledge and champion the role of the data holder; develop systems for third-party custodianship, management and data expertise, thus moving the burden from data holders who do not see themselves first as data providers; provide systems that give confidence, particularly around fraught areas such as privacy and ethics.

It is important to recognise that these issues are both social and technical in nature, so solutions need to be cohesive across both. For example, mandating data standards without a process of consultation and inclusion across communities may not achieve the outcome. Building standards through enabling communities of practice and processes of consensus decision-making may be more effective in the longer term and provide a more lasting authorising environment.

1.2 Emerging technologies to enhance responsible data availability and use

When data is being used, it is being analysed to produce some insight that will drive a decision or action. The quality of these decisions and actions is dependent on the quality and scope of the data that is used, the effectiveness of the analysis that is being applied, and the ability of the organisation to use the resultant insights effectively.

The quality and scope of data that is available for any particular analysis is bound by the sensitivity and value of the data in comparison to the value of the analysis. Within a single organisation, most or all of the organisation's data may be available for use. However, when external data is included in an analysis, that data must be sourced, merged with existing data, and then analysed. In general, low value or low sensitivity external data is readily available, while more valuable or more sensitive data is more difficult to access, and is often protected. Technology solutions have been developed over many years to enable more data to be available for use, such as anonymisation and data obfuscation techniques, and this activity continues today, with many exciting technologies for improved data availability on the horizon.

To illustrate the evolving technology landscape for improving data availability for use, it is useful to breakdown the flow of data within the data driven decision-making process into four steps:

- **Extract** – obtain access to data from a source system, e.g. a sensor, a data warehouse, an online data API etc.
- **Join –** Relate the data from one source to data from another source, e.g. combine based on postcode, user id, industry etc.
- **Protect –** Ensure that information that should not be exposed is kept hidden, e.g. obfuscation, aggregation, anonymisation.
- **Use –** Provide the ability for the data to be analysed, e.g. release on data.gov.au, API access, containerised access, secure facility.

In each step there are a variety of options as to how the data is treated, the combination of which defines how secure, how private, how convenient, and how useful the data is for decision-making.

Extract

The following facilities or services would greatly improve the ability of both public and private organisations to provide data that can be used effectively by other organisations:

- Guidelines and API services for extraction of data from source systems according to agreed standards, e.g. for administrative boundaries through G-NAF, and other official categorisations for non-spatial data.
- Guidelines and API services for publication of data from source systems using standard systems, e.g. geospatial services through WMAP etc.
- Services to expose logical relationships between datasets to providers of data, and suggest consistent or compatible aggregation levels, units of measurement, etc. that allow them to be merged at a later stage with other datasets.

- Services to measure and rate quality of data, based on intrinsic features of the data (e.g. completeness, consistency etc.) and extrinsic features, (e.g. use and issues reported by other organisations).
- Services to track and expose the provenance of data for use by subsequent analysis systems.

Join

Joining aggregate data requires the use of common aggregation standards. These standards should be defined and services that provide transformations between commonly used forms and these standard forms should be available to both public and private organisations.

Currently joining unit record level data involves disclosing the data to a linkage agency. New technology for doing privacy preserving data linkage has being developed over the last 6 years (e.g. by Dr Peter Christen at ANU) that can greatly speed up this process, at the cost of some accuracy. These capabilities enable the joining of data with disclosure of PII.

Protect

Existing methods for protecting data include simple methods such as aggregation and suppression, such as only releasing data at postcode level and only with a sufficient number of counts. These methods can be augmented with perturbation methods, such as that used by the ABS Table Builder product, which protects tables released from unit level census data from deidentification attacks. These methods should be made available in a standardised form to both public and private organisations, and this is part of the work the Data61 is undertaking within the NISA framework.

There are several recent and new methods of data protection that CSIRO is actively researching and developing into services. These will be appropriate for a variety of sensitive data.

Synthetic data release

There is a recent advance in privacy technology known as Differential Privacy, introduced by Dr Cynthia Dwork at Microsoft. Differential Privacy is a quantifiable measure of the privacy of certain data analytics techniques that involve random perturbation of either the data being analysed or the analysis itself. CSIRO Data61 is working on a variety of differentially private mechanisms to allow the release of synthetic unit record datasets that contain statistically similar data to the original data, but can guarantee that the released data cannot be re-identified. Data61 is undertaking investigation of these methods within the NISA framework to allow the release of a significant federal government dataset with less restrictions than is currently possible. These techniques involve adding noise to the data, so have some impact on the utility of the data for analytics.

Confidential Computing

Another area of intense academic and commercial interest currently is the field of privacy preserving machine learning or data mining. There are many focus areas – for instance homomorphic encryption, secure-multi-party computation, oblivious computation, and zero-knowledge proofs. There are a variety of research groups in major universities and corporations around the world developing new technologies in these areas.

CSIRO has expertise in homomorphic encryption, which enables calculations to be done on data while the data is encrypted; and secure multi party computation, which allows data to be computed on by multiple parties, where no party can learn the data without colluding with a majority of other parties. Both of these approaches are considered very promising as a long term solution to the data protection problem, however “fully homomorphic encryption”, which is a recently discovered capability, is not yet practical for large scale data analysis problems.

As part of its "confidential computing" platform, CSIRO Data61 is developing a combination of "partial homomorphic encryption" (which is more limited but more efficient than fully Homomorphic encryption), distributed computing and machine learning. This platform enables the provision of services that allows organisations (both public and private) to do joint analysis of data without exposing their data to any other party. These methods are being applied to federal government data within the NISA framework as a proof of concept.

Use

Data use occurs through a large number of mechanisms, including direct consumption of data by individuals through visualisation in dashboards or through web and mobile applications. These approaches are well supported through the use of data publishing APIs.

Another use of data is by researchers, data scientists, and analysts who are investigating particular problems and wish to find and analyse all the available data relevant to that problem. Currently there are a variety of mechanisms to allow access to that data:

- Open data access – the data scientist downloads the data directly
- Containerisation – the data scientist can access the data on a secure platform through a containerised application (e.g. the SURE platform used by the Sax Institute)
- Secure Facility – the data scientist goes to a secure site to access the data (e.g. the ABS ADSL)

Unfortunately, the last two of these greatly restrict the number of people that can access the data, and the convenience of that access. The "Confidential Computing" platform promises another approach, which allows access to a proscribed set of analytics functions that are performed over encrypted data that is not disclosed to the data scientist or analyst. This enables a new, low friction, method of doing exploratory linkage and analysis of datasets. This approach may allow the discovery of new connections and insights without the overhead of the training, authorisation and provision of current approaches, while still maintaining the confidentiality of the data. More expensive access to the data directly can still be obtained through current methods, particular if justified through exploratory analysis over encrypted data. This capability is equally relevant to intra-government data collaboration, government-private data collaboration, and private-private data collaboration.

Standardisation

There are a wide variety of options available for each stage in the flow of data from production to use. The somewhat idealised view described above is actually more complicated when considered in more detail. As such, standards and guidelines for the application of different components in the flow in different scenarios need to be developed at a national or international level. These standards should address which mechanisms for protection and use are appropriate for which applications on data of a given sensitivity.

1.3 Examples of high value public sector data sets

The Open Data 500 Australia project recently surveyed many Australian sectors, agencies and research organisations, to understand what data is most valuable to users[9]. The top 5 most requested data sets from research and industry are: Census and Labour Force Longitudinal data sets, Geo-coded National Address File, Personal Income and Business Tax, Health Benefits Data (MBS, PBS) and Social Security Payments.

Public data sets of high value to the public and research sectors are those which provide 'baseline information'. This is data for which collection costs are high, for which the data is unrepeatable (i.e., time series data) or reduces the requirement to gather data on subsequent occasions, and is not ephemeral in

[9] http://www.opendata500.com/au/about/

nature. Baseline data provides a reference point to a current circumstance. Retaining this data and ensuring it is available to future users can save significant setup costs for projects. Examples of datasets of high value are those which provide information on the state of the environment or other variables at a given time, i.e., coastal retreat information, tides, bathymetry, fish catch data, hydrological data, fisheries licencing information. Data which provides links between datasets (i.e., market data) is also valuable. These datasets can be key net contributors to the economy.

There may be an activity for identifying the '100 year data' or national collections – that data which is core valuable data and should be preserved. There is potential alignment with the National Facilities and Collections in CSIRO, as well as potential for collaboration with Australian National Data Service (ANDS) and the identification of 'trusted repositories' for data collections.

The Atlas of Living Australia combines mainly public sector data sets (over 1000 data sets), with notable contributors being Commonwealth Government Departments, all State and Territory Governments, Australian biological collections (museums and herbaria), NGOs, NRM and CMA groups and Citizen Science initiatives.

The Atlas of Living Australia has achieved a great deal in establishing the benefits of sharing public data sets. These include:

- The ALA currently holds over 60 million records, and to date over 7.5 billion records have been downloaded by a wide user community to support research, education, conservation management, biosecurity, environmental assessment, collection management and citizen science.
- Over 800 academic articles have referenced the ALA since 2007 (Source: Google Scholar)
- ALA represents the largest partnership for biodiversity in Australia
- Digivol – an online platform enabling volunteers to digitise museum collection specimens – has reached over 350,000 digitisation events
- There are over 1 million images and 4500 sound files on the Atlas
- In 2015 the Atlas of Living Australia had over 625,000 users, over 5 million page views, and an average 3500 users per day (Source: Google Analytics)

In the Energy domain the Australian Energy Market Operator (AEMO) datasets are impressive, including residential, commercial and industrial energy metering and tariff data (at fine spatio-temporal granularities). Release of even aggregated views of such data would enable statistically robust assessment of proposed energy tariff changes, enable the exploration of policy impacts on energy consumers (i.e. energy efficiency and building code standards) and would reveal market opportunities for new entrants looking to provide bespoke energy services (where, at present, there is insufficient data to build a compelling business case).

Sharing of Clean Energy Regulator (CER) datasets, which include large-scale industrial and commercial energy consumption datasets by facility type, and solar generation uptake (captured at the building level). Though aggregated data is published, alternative (and finer-grained) aggregations would provide new insight into the energy consumption behaviour of different commercial/industrial segments, provide clearer linkages between energy productivity and the commercial sector, and would highlight the demographic predictors of solar PV uptake and the dominant market players in the PV space.

MBS (Medical Benefit Schedule) and PBS (Pharmaceutical Benefit SCHEME) are two of Australia's largest health data sets. These capture information where federal government funding is used to fund some or all of a patient episode of care or pharmaceutical script. However these are examples of administrative data sets and don't provide direct clinical information, although this can sometimes be inferred. Linking of this information to other data sets will add considerable value to these data collections.

Regulatory data on Australian businesses

The scope of these datasets relates to information about individual businesses and related organisations (e.g. charities, non-government organisations, etc.). The primary datasets include the ASIC Company Register and Australian Business Register (ABR). There are however other important secondary datasets relating to Australian companies and related organisations (e.g. large taxpayers, government procurement, research funding, patents and trademarks and many more related to the regulatory functions of agencies such as employment, environment, etc.). These datasets are different to confidential survey information collected by the ABS for official statistical analysis.

The Australian Government has made good progress with releasing a limited part of the ABR for widespread industry use but could release further information that would be of high value to external users. There is an opportunity to open up more of the ABR information for wider public use as open data. The release of data such as an entity's standard industry code, description of main business activity, other business locations and business size would be extremely valuable for reuse in a range of business and research activities.

If there are concerns about the sensitive nature of some of this data, it can be modified to release it at a more aggregated level. For example, the release of public data on business size could be based on standard categories such as micro (less than 5 employees), small (5 to less than 20 employees), medium (20 to less than 200 employees) and large (200 or more employees).

A recent study by the Open Data Institute identifies that only 3% of countries provide free access to ownership information about companies[10] , although company registers in the UK and New Zealand provide free access to full company records.[11] Several groups have proposed that the detailed information on Australian companies in the ASIC Registry should be provided free of charge. This would make the ASIC Registry an open data repository, allowing far greater access to researchers, journalists and the general public. It would also enable other businesses to use this data in innovative ways to create new value-added services.

If Government sees merit in there being free and open access to all relevant company information on the ASIC Registry this will need to be considered in the current sale process for Registry being managed by the Department of Finance.

See Appendix 2 for more information about the importance of regulatory data about Australian businesses and other organisations.

Research and Development Grants and Offsets

Data providing details of R&D grants to or support involving specific research organisations, businesses and other organisations is more available as open data compared to other grant data. This data however could be made available in a more structured manner to make identification of organisations and individuals easier and more consistent. This includes the provision of ABNs as a unique identifier for businesses and other non-research organisations and Open Researcher Contributor ID (ORCHID) as a unique identifier for researchers.[12]

Data providing details of the R&D Tax Incentive to specific businesses is not currently available to the public. Given that these tax offsets amounts to just over $2 billion per year in Australian Government support for

[10] The Open Data Barometer Global Report – Second Edition: World Wide Web Foundation, January 2015, p. 6. http://opendatabarometer.org/assets/downloads/Open%20Data%20Barometer%20-%20Global%20Report%20-%202nd%20Edition%20-%20PRINT.pdf

[11] New Zealand Government Companies Office, www.business.govt.nz/companies Companies House website, United Kingdom Government ,https://beta.companieshouse.gov.uk/

[12] NHMRC and ARC Statement on Open Researcher and Contributor ID (ORCID) NHMRC website www.nhmrc.gov.au/grants-funding/policy/nhmrc-and-arc-statement-open-researcher-and-contributor-id-orcid

businesses (approximately 25% of the total support for R&D), there is an argument that such information should be made available as open data, in a manner similar to R&D grant data.[13]

The three main sources of R&D grants relevant to Australian businesses and other organisations include:

- Australian Research Council (ARC) administers the National Competitive Grants Program (NCGP) and provides on its website open data on grants approved for research organisations. This includes a spreadsheet with partner organisations associated with the grant program dating back to 2001.14 This spreadsheet could be improved with the inclusion of ABNs to assist correctly identify the partner organisations, as well as information about the level and nature of industry support for the Linkage grant projects.
- National Health and Medical Research Council (NHMRC) administers Australia's medical and health research grants and provides on its website open data on grants approved for research organisations. This includes a spreadsheet showing recipients of the grant program dating back to 2000.15 This spreadsheet could be improved with the inclusion of ABNs to assist correctly identify the partner organisations.
- Innovation Connections Grants are a series of grant programs to support the placement of researchers in businesses funded by the Department of Industry, Innovation and Science. This includes the Researcher in Business, Research Connections and the current Innovation Connections program.16 It would be valuable if the data on the companies that have been recipients under the program were provided as an open data set (including ABNs as a unique identifier).

Industry Development Grants

Grants constitute approximately 6% of Australian Government expenditure, with further expenditure undertaken by state, territory and local government jurisdictions. Grants for industry development are a subset of overall government grants, with the other purposes including delivery of services on behalf of government policy, and research and development.

The Commonwealth Grant Guidelines require that all grant payments that are awarded to recipients are published on agency websites (Department of Finance and Deregulation, 2013), but does not require this to occur in a consistent format or location. Currently, there is no comprehensive Australian Government wide repository of information about grant activity. As the National Commission of Audit has noted, grant data can be difficult to find, assimilate and analyse.[17]

The Department of Finance has commissioned the development of a whole-of-government grants advertisement, lodgement and reporting system. The project is currently in development and is expected to be fully operational in 2017, with some functions, such as grants advertising and notification, to be rolled out in 2016.[18]

While the grants.gov.au initiative is likely to greatly improve the availability of grant data, there are immediate steps Australian Government agencies can take to improve access to and the quality of their grant data. This includes publishing lists of their current and recent grant recipients (e.g. covering data for at least the last ten years) including an ABN identifier in a simple machine readable format (e.g. csv) on data.gov.au.

[13] National Commission of Audit, February 2014. 10.2 Research and development http://www.ncoa.gov.au/report/appendix-vol-2/10-2-research-and-development.html

[14] Grants Dataset, ARC website, http://www.arc.gov.au/grants-dataset

[15] Research funding statistics and data, NHMRC website, www.nhmrc.gov.au/grants-funding/research-funding-statistics-and-data

[16] Innovation Connections Programme, www.business.gov.au/assistance/innovation-connections

[17] National Commission of Audit, February 2014, 10.17 Grants programmes http://www.ncoa.gov.au/report/appendix-vol-2/10-17-grants-programmes.html

[18] Australian Government Grant News, June 2015 http://www.finance.gov.au/sites/default/files/grants-news-june-2015.pdf

Even when the grants.gov.au system is fully implemented, there is still value in having access to this data in both a simple csv format as well as through an API.

Some of the priority datasets for release as open data relating to Australian business include:

- Export Market Development Grants (EMDG) scheme, administered by Austrade, provides grants to for aspiring and current exporters to develop export markets. While Austrade provides case studies of selected recipients on its website, it does not publish a list of all the recipients in either its Annual Report or website.[19] This list should be provided in machine readable data with the inclusion of ABNs to assist with the identification of each business.
- The Export Finance and Insurance Corporation (EFIC) currently publishes its Annual Report the full list Australian businesses provided with finance and other assistance. This list should be provided in machine readable data with the inclusion of ABNs to assist with the identification of each business.[20]
- The Department of Industry, Innovation & Science administers a number of industry advisory and support programs including the Entrepreneurs', Small Business Advisory Services and Accelerating Commercialisation Programs.[21] It would be valuable if the data on the companies that have been recipients under the program were provided as an open data set (including ABNs as a unique identifier).

1.4 Dataset and Education

CSIRO Education and Outreach has been working with CSIRO research teams to deliver research datasets to students and teachers. The aims of providing this data are:

- In a safe and secure environment to allow students to make their own scientific discoveries from the data (and where significant enough co-author scientific papers with the researchers).
- Provide real-life experience using real data for students to learn skills including coding, statistical analysis, scientific, and computational thinking, logical thinking and reasoning to design their own research questions for analysing the data; visualisation and visual interpretation of data and potentially to inspire invention and innovation by using data to create products and services.
- Depending on the datasets include associated information such as ecological or social contexts; scientific collection methodology, research questions being investigated etc. to raise awareness and knowledge of the discipline or area of research.
- Provide links and access to researchers and contemporary scientific processes and methodologies.
- Gain experience and confidence in handling real data and making their own decisions about how to investigate that data.
- Build trust in datasets and scientific organisations.
- Develop cohorts of students with an awareness of data collection, storage, analysis and the skills to interpret data; and to be scientifically and digitally literate and potentially innovation literate.

Impediments to making research datasets available to students and teachers include:

- The time (and therefore cost) of the researchers to ‘clean’ up the datasets so they can be used by the teachers and students.

[19] Austrade Annual Report, 2014/15, http://www.austrade.gov.au/ArticleDocuments/1401/Austrade_Annual_Report_2014_15.pdf.aspx

[20] Export Finance and Insurance Corporation Annual Report 2014/15, http://www.efic.gov.au/about-efic/our-governance/reporting/

[21] Grants and Assistance www.business.gov.au/assistance

- For complex and large datasets providing subsets that are manageable but that do not compromise the integrity of the data (and therefore the trust of the user) e.g. stream flow data that records information every minute for years.
- The diversity of digital, mathematical and scientific literacy among students and teachers and therefore needing to cater for a range of skills and provide appropriate level instructions/ manuals to access and make use of the data. The options are to create scaffolded curriculum-linked lessons targeted at key year groups (e.g. years 5 &6; 7&8 and 9&10) to provide a range of entry points with easy entry, low skilled options and enough extension to challenge highly skilled students and teachers.

Most of these impediments could be overcome by the standardisation of data set formats and naming (and metadata) protocols thereby allowing standardisation of tools for analysis; grouping datasets together in a meaningful way (and to allow comparison between datasets); datasets ‘served’ on a common platform (e.g. similar to the Atlas of Living Australia for museum collections). Use and uptake by schools and students would be helped by links to the curriculum, making the datasets easily available and inclusion of data analysis in key national education strategies.

1.5 Datasets provided by CSIRO

The ALA combines mainly public sector data sets (over 1000 data sets), with notable contributors being Commonwealth Government Departments, all State and Territory Governments, Australian biological collections (museums and herbaria), NGOs, NRM and CMA groups and Citizen Science initiatives.

The Atlas of Living Australia has achieved a great deal in establishing the benefits of sharing public data sets. These include:

- The ALA currently holds over 60 million records, and to date over 7.5 billion records have been downloaded by a wide user community to support research, education, conservation management, biosecurity, environmental assessment, collection management and citizen science.
- Over 800 academic articles have referenced the ALA since 2007 (Source: Google Scholar)
- ALA represents the largest partnership for biodiversity in Australia
- Digivol – an online platform enabling volunteers to digitise museum collection specimens – has reached over 350,000 digitisation events
- There are over 1 million images and 4500 sound files on the Atlas
- In 2015 the Atlas of Living Australia had over 625,000 users, over 5 million page views, and an average 3500 users per day (Source: Google Analytics)

CSIRO provides astronomy data sets that are of high value to the research sector, academics and education. These are generated from observations taken with radio astronomy facilities including the Parkes radio telescope, Australia telescope Compact Array (ATCA) and Mopra radio Telescope. Commissioning Data products from the Australian SKAP Pathfinder (ASKAP) array are also available and a large volume of ASKAP data will be available from the start of Early Science later this year.

The primary use of these data is for astronomy research carried out by scientists located in Australia and overseas. There is also some use for educational programs, targeted at high school students. There is huge potential for school students to make a significant and meaningful contributions to the analysis and therefore understanding of astronomical data with the Australia Square Kilometre Array Pilot (ASKAP) project and then the Square Kilometre Array coming on line.

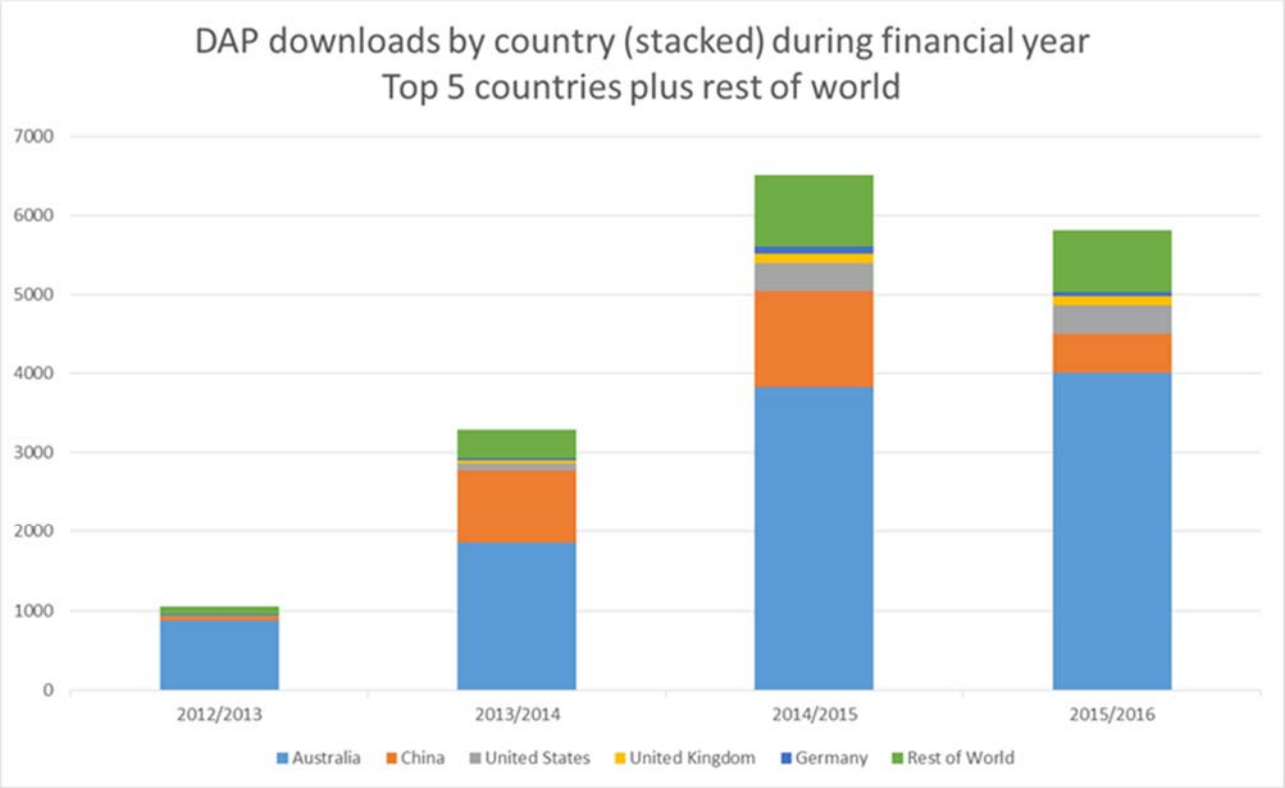
Radio astronomy is at a tipping point, moving from relatively small data sets that can be processed by desktop computers or small clusters, to extremely large data flows that require real-time processing and high performance computing. For this new generation of radio astronomy data, therefore, data availability is not sufficient: there also needs to be available computing infrastructure available to the research community for processing and storage.

There are few legal impediments to the Open Access release of these data. Parkes, Mopra and ATCA data are made openly available following a proprietary period of 18 months during which the data are restricted to science teams. ASKAP data are made openly available without any proprietary period but following a validation process to ensure data quality.

The Marine National Facility provide important environmental and marine datasets over a time span of more than 35 years, which are used to understand changes in our marine ecosystems and environment through time. The data is made available publically through the CSIRO DAP.

The CSIRO Data Access Portal (DAP) is part of the Australian Research Data Commons and enables research data assets to be published. It has been developed by CSIRO and research partners as a secure repository for CSIRO's research data assets. The DAP offers self-serve capability for researchers to deposit, describe and manage access to research data assets. It also offers options for searching, retrieving and downloading research data. Researchers can assign access permissions and select from a range of licences. DOIs are automatically generated for publically available collections.

The DAP also accommodates the publication of software and tools which facilitate the use of data; and it links to publications associated with data. All of this improves the context in which data is presented. It provides information about some of the issues that researchers need to consider in making data available and choosing terms on which it is made available – sometimes those terms being dictated by the nature of existing arrangements concerning the data. The issues that need to be considered include privacy and ethics, IP protection (often going to the timing of the release of data) and contractual obligations relating to projects in which the data is generated, quality and reliability of datasets.



2. *Availability of private sector data (ToR 2)*

Given the potential value in making private sector data available, we recommend that the Australian Government:

- **Recommendation 10:** Align public and private data activities to achieve efficiencies and increase availability of private data, by making the guidelines and standards for public data available and enabling provision of de-identified datasets. .
- **Recommendation 11:** Play a catalytic role for the public sector to provide training and access to platforms, protocols and processes that facilitate the private sector to use and share public data and platforms (with appropriate protections in place) with the option of value adding and marketing that data to aggregators who have a demand for it.

2.1 Issues in availability of private sector data and benefits of addressing these

For the most part, private sector environmental data is not available for open use. In some regions, such as those with large extractive industries, private sector data can far outweigh data collected by public sector organisations. This has resulted in large quantities of coastal and ocean data (including water quality monitoring, coastal habitat extent or species tracking) which is inaccessible for environmental reporting purposes outside of the specific industry study. Significant efficiencies could be achieved through aligning public and private data activities, particularly related to regional or national ecosystem or State of Environment style reporting.

In Agriculture there is an argument that there is a catalytic role for the public sector to provide platforms, protocols and processes that facilitate the private sector (including farmers) to share data (with appropriate protections in place) with the option of marketing that data to aggregators who have a demand for it. Otherwise lock-in of data ownership and use will occur between the generators of that data and private sector interests, preventing its potential wider use.

There is the potential to provide de-identified datasets from the private sector for student use. This will not only greatly increase the number of datasets that are available but also allow students to choose areas that interest them, work with data that may have direct relevance to them and their families e.g. energy use in homes including the use of and efficacy of alternate energy supplies. The context provided where the data is made available provides the opportunity for learnings related to privacy, the importance of having access to and knowledge of all datasets and being able to compare datasets to look for similarities and differences.

There is little information from private data sets in health that are made available. This often creates considerable “gaps” in the data as people move between the private and public system as part of their treatment. This reduces the ability of researchers to use data for both health system improvement and medical research.

2.2 Examples of valuable private sector data sets

Availability of energy retailer datasets, which include residential tariff uptake and gross solar generation output. Release of aggregated views of this data would provide a true baseline for understanding the impact of tariff changes in the energy sector, while gross solar generation is critical to understanding actual solar production in Australia (other data sources are typically net metered).

While mainly working with public sector data, the largest single data contributor to the Atlas of Living Australia is BirdLife Australia (12.5 million records). Whether an NGO is public or private sector is perhaps debatable, but this is an example of a non-publicly funded organisation that saw the value of contributing its data to the ALA.

2.3 Legislation and other impediments

As with public data, the primary concerns raised with respect to the sharing of private data are associated with privacy and the cost to share. The other critical barriers to releasing/sharing/publishing private datasets, including legislative restrictions, are a lack of ongoing custodianship practices, perceived privacy risks and potential backlash, an insufficient business case for releasing data, and competitive disadvantage.

That is a significant portions of the *cost* of providing discoverability, accessibility and usability of a data asset to a level where the asset can be re-used or repurposed, sits with those that share it (the owners/producers). Whereas, the potential *benefits* lie with users. As a result, owners/producers seek to minimise these costs with a subsequent impact on discoverability, accessibility and usability. Users are then faced with the challenge of either spending extra resources to meet their data needs, or potentially compromise the outputs of their project by using limited data

Unlike public data, one extra concern for private data holders are the twin challenges of avoiding providing a commercial advantage to competitors and avoiding cartel behaviour through knowledge sharing.

Solutions to the privacy (e.g. customer data and sensitive data that might mean customers are identifiable) and cost to providing data and/ or loss of market advantage (where data allows a company to have advantage in the market) barriers might be to include protocols for only including de-identified data available or having time limits on when the data can be made available (i.e. when there is no longer market advantage in the data). Another solution is the concept of the 'trusted third party' whereby an organisation (public or private) acts on behalf of private organisations and is trusted to protect the interests of those organisations.

The wider EUDM programme and initiative, including its approach to providing a platform for integrating energy use data from across the sector while explicitly managing privacy and delivering ongoing custodianship (see also case study in section 1.1.1 and **Appendix 1**).

Vertical integration of information supply chains is the process by which companies seek to control data collection and analytics in order to provide convenience for users through newly created services. However, as a result of this integration:

- the capacity to uniformly increase competitiveness across an economy, by reducing the costs of information supply chains, remains low; and
- opportunities to create new value from the combination of data and analytics in previously unforeseen ways are limited.

As such, this trend towards vertical integration of information supply chains represents a risk to public research and the economy at large.

There are currently differences between state and federal Privacy legislation that causes issues when requesting similar data from multiple jurisdictions.

3. Individuals' access to public and private sector data (ToR 3)

3.1 Datasets provided by CSIRO

Individuals currently access data through CSIRO data archive services. These include the Australia Telescope Online Archive and the CSIRO Data Access Portal (DAP). The DAP also connects to the Research Data Australian[22] - a national repository of research data as well as providing the ability to syndicate CSIRO metadata to other subject portals. In addition data are made available through the implementation of relevant domain standards, such as Virtual Observatory standards in the case of the astronomy data. These are used in astronomy to facilitate the international sharing of data between different research facilities.

Considerable resources have been used to make radio astronomy data as easy to access as possible and this is an area of ongoing efforts. The experience of the astronomy community is that making such datasets openly available and relatively easy to access and process encourages excellent science.

The key radio astronomy data sets are:

- Data for observations of pulsars taken with the Parkes radio telescope;
- Other unprocessed data from the Parkes radio telescope, ATCA and Mopra radio telescope;
- Processed data products from Mopra and ASKAP.

In Australia, Australia Astronomy Limited provides an important role in assessing data requirements for astronomy across Australian astronomy, including optical, infrared, radio and theoretical astronomy. National coordination helps with the adoption of protocols and a more standardised approach to interfaces. These are of benefit to users. Centralised support services could also help in these regards.

On an international level – discussions between Australian developers with groups from other astronomy institutions and with the International Virtual Observatory Alliance are also important for the development of applications targeted at end users. The standards set by the International Virtual Observatory Alliance intend to make astronomy data across all wavelengths more easily available. The astronomy community in Australia is at the forefront of applying these standards both in optical and radio astronomy.

While CSIRO increasingly seeks to make datasets available freely, not all CSIRO data is made available to the public. There are some cases where privacy, ethics, contractual arrangements (e.g. projects conducted with others), commercialisation and IP requirements and sensitivities and so on make providing data immediately and freely available not possible.

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

4. *Standardising, sharing and release of public and private sector data (ToR4)*

We recommend that the Australian Government:

- **Recommendation 12:** Enable communities of practice as cohesive authorising environments for standards development, connected with international communities – addressing the social and technical aspects of data and information management.
- **Recommendation 13:** Standardise data to enable combining of disparate datasets to develop multi-variable models, which will require development of a model for funding the processing of data to a standardised format at the level of institutions acquiring data.
- **Recommendation 14:** CSIRO supports the principle of all data generated with public funding for widespread community use purposes being made open and freely available, noting that careful consideration of potential ethical, privacy issues, requirements in relation to intellectual property or under statute in relation to the data, and any other limitations must be taken into account, and this may result in time delays for release and for certain types of data it may not be possible to achieve immediate release of data.

4.1 Standardisation of data

In the environmental domain, some areas such as ocean and climate data are relatively well standardised due to the long history of data sharing at the global level. In the coastal domain, however, standardisation of data collection is much less prevalent due to a very large number of organisations collecting data for different purposes.

At the applications level, however, where the data is used, for example to evaluate ecosystem responses to climate change, fire regime changes, impacts on reef, etc., which requires value adding and downscaling for domain relevance, and combining with other environmental datasets, this breaks down and standards are not effectively supported due to lack of integrated disciplinary focus on requirements

Data should be in a relevant standardised format, recognising that some domains will have domain-specific requirements for data and metadata formats. The marine research community currently uses domain-specific (but open-source so interoperable) data formats, and agreed metadata standards such as ISO19115 allied with domain-specific vocabularies (i.e., Marine Community Profile) to ensure that data can be shared as widely as possible. Open-access rather than proprietary formats should be used. Metadata should be described using machine-readable standards – not all institutions use interoperable or machine-readable standards. The real value of data is in its analysis, and unless data is well-described and in an accessible format it cannot be analysed effectively.

Benefits include the ability to combine disparate datasets to develop multi-variable models, particularly in the environmental and natural hazard response domains. However a model for funding the processing of data to a standardised format must be developed at the level of institutions acquiring data. The value here is in simplifying the process of bringing together datasets from across domains and sectors to explain the predictors and drivers of all sorts of behaviours. Whole communities of practice involved in the supply chain of data use need to be involved to ensure an institutions standardisation lines up with requirements for use of that data.

The environmental and biodiversity data community has well-established global standards, with several options for data sharing already in place and working. These could be further developed by expanding the existing collaboration between the NCRIS funded environmental facilities (e.g. ALA, IMOS, TERN). Data standards are already in place (and widely used) for much of this information.

Australia’s e-Health infrastructure is standardising on the data formats and terminologies which are to be used when capturing and exchanging health information. It is important that state health jurisdictions in particular make these standards part of the procurement process so that health data can be more successfully used for primary and secondary purposes.

Some government departments still maintain part of their data separately from the NCRIS environmental initiatives. While holding and managing the data separately may be sensible, the sharing of the data would be of mutual benefit.

While we support the issues paper call for standardising, sharing and release of data (where for the most part there are existing protocols already in place), we would like to see additional emphasis on the creation of data. This should include new methods of data capture, “cradle to grave” data management policies, and rapid digitisation programs to bring physical information into the digital world. e-Infrastructure such as the ALA is only as good as its content – so initiatives that increase the flow and quality of data are always welcomed.

4.2 Data sharing

It is axiomatic that the value of data is increased as it is shared. CSIRO is currently working on an assessment of the economic value of publishing research data. Public data, while not necessarily collected for that purpose, is a valuable input into research. It would be desirable to examine the relationship between public data and research and assess the economic value to the innovation system, thus extending the Houghton[23] and Gruen[24] reports. This relationship includes community good as well as commercial opportunity. It is expected that the economic value assessment will be completed in the coming months and CSIRO would welcome a discussion with the Productivity Commission once the assessment report has been finalised.

In order to increase the value of public data, it needs to be brought together in meaningful ways. The presentation and/or availability of data on given topics or of related types will enhance its comprehensiveness and thus potential value. Numerous initiatives have been undertaken at different levels of government, the research sector and across different aspects of the environment to address the issue of data sharing. For example, three National Collaborative Research Infrastructure Strategy (NCRIS) facilities currently aggregate, maintain and publish information around:

- Atlas of Living Australia (ALA) – biodiversity
- Terrestrial Ecosystems Research Network (TERN) – soils, vegetation, and others
- Integrated Marine Observing System (IMOS) – marine

Similarly many jurisdictions have implemented (or are implementing) some form of environmental information portal (e.g. NSW Environmental Data Portal) and may have many such portals, each focussing on a different aspect of the environment (water, biodiversity, land use, etc.). Enabling the connection of these various portals would enable more comprehensive vies of and access to the data on a given topic, for example the nation’s water data or soils data, or LIDAR data.

However, whilst there is significant activity, there remains very little coordination and significant ongoing costs. Where two portals nominally provide access to the same or similar data, they are unlikely to operate in the same way, the data provided will not be in standard forms nor have consistent semantics (meaning) and is therefore difficult to combine. Furthermore, there is little or no governance in place and a range of ad-hoc bilateral agreements between organisations resulting in duplication of effort and confusion of

[23] http://www.ands.org.au/__data/assets/pdf_file/0004/394285/houghton-cost-benefit-study.pdf

[24] http://www.ands.org.au/__data/assets/pdf_file/0019/393022/open-research-data-report.pdf

stakeholders. Belbin and Williams (2016)[25], while outlining the achievements of the ALA Spatial Portal, point to the problem of no central authority responsible for ongoing aggregation of the best available and definitive spatial and contextual data 'library' supporting environmental modelling, lacking the authority to effect change. For example, at the Australian Government level, the Environmental Information Advisory Group (EIAG) provides a forum for commonwealth agencies to discuss issues and initiatives associated with environmental information. However, this has no authority and no ability to direct resources or individuals to take particular actions. Activities undertaken by it are largely based on the goodwill of participants.

This would require the Commission looking at accessibility of data across jurisdictions, i.e. state collected data feeding into collections for the national good. In addition to this, identifying industry data that is of national benefit and assisting in the availability and accessibility would also be beneficial. Data on resource consumption is an obvious example.

There may be advantages in developing an infrastructure to address the coordination and management of spatial environmental data (development and delivery), in the same way that the Atlas of Living Australia acts as an aggregator via agreements with a range of data owner/custodians. Such an infrastructure could also support standards development and delivery models for research as well as dashboards or portals to support the general public or community-based interest in viewing environmental information.

Box and Lemon (2015) proposed a 'social architecture' designed to tackle some of these issues. However, implementing such an architecture has proved troublesome as this requires individual agencies, often driven by their own needs, to relinquish authority and, in some cases, control, of certain data assets to others.

This report identified, through a review of a number of information infrastructure initiatives, that to achieve impact a number of issues related to governance, participation and agreements need to be addressed.

Governance

- The need for broad inclusive governance mechanisms including participation of state/territory governments (which hold significant environmental data), research, not-for-profits and industry;
- The need to improve alignment of activities within thematic policies as well as crosscutting policies such as those related to eGovernment and Open Data;
- The need to link technical outcomes to high level government policy drivers and outcomes;
- The critical role of whole-of-government leadership and priority setting to provide top down drivers and coherence for collective activity; and
- The risk of lack of 'buy-in' if some stakeholders are not engaged early.

Participation

- The importance of key individuals, such as trusted leaders, who have built community trust based on strong track records within domains. Strong senior champions are needed to drive the initiative within participating organisations;
- Ensure recognition and support is given to participating individuals and organisations;
- The need for coordination, communication, and awareness at senior levels; and
- The critical role of networks and the need to tap into and leverage self-organising and formal communities that typically form around technical and domain issues and expertise.

Agreements

[25] Belbin L, Williams KJ (2016), Toward a national bio-environmental data facility: Experiences from the Atlas of Living Australia. *International Journal of Geographic Information Sciences* **30**(1), 108-125.

- Agreed, standard data licensing frameworks are key for effective data sharing. The recommendation to use the Australian Governments Open Access and Licensing Framework (AusGOAL) 1 to support and guide NEII open access was strongly expressed;
- The major barriers to data access are policy, organisational, legal, and cultural. Open Data initiatives have helped overcome some organisational and legal barriers and should be leveraged wherever possible;
- Promoting the importance and adoption of common standards, particularly for system and data providers, is an important area of work;
- Designated central authority plays a key role in coordinating the specification and implementation of standards for collection, integration and dissemination of information; and
- Obtaining permission from and ensuring attribution of data rights holders is key to ensuring continued participation.

The critical role of whole-of-government leadership and priority setting to provide top down drivers and coherence for collective activity; and risk of lack of 'buy-in' if some stakeholders are not engaged early**.**

4.3 Duplication in data collection efforts

There is considerable duplication in the collection of data from industry on farm performance, practices and attitudes. The ABS, ABARES, Rural Research and Development Corporations (RDCs) and other organisations (like agribusiness companies, research organisations) collect similar information at varying degrees of spatial and temporal granularity. There is an obvious opportunity for harmonisation, efficiencies, reduction in respondent burden and consistency in data collection.

There is also considerable duplication of effort in the creation of various access points such as portals into data of various types. Having multiple access points can be advantageous – it increases exposure and pulls relevant data together for a specific purpose. Reinventing the technology and re-doing the description is less desirable. There should be an ability to discover infrastructure in place, understand it and re-use or modify it for purpose where possible. Data should only ever be described once and those descriptions shared and transformed where appropriate. The model for resource sharing employed by the National Library may be worth examining.

4.4 Storage and publication of data

The storage, documentation and accessibility of research data particularly that which is publically funded, is not mandated by research funders or providers. This means the dissemination and re-use of such data is patchy and often limited.

Precedents for mandating publication of certain research data that is publically funded exist in the USA. Learnings from such implementations of publication policy would be valuable in developing an Australian legal and policy framework for mandating publication of suitable categories of publicly funded research data.

Examples of suitable categories of publicly funded research data for mandating publication of are data in the nature of 'public' data, such as examples given in this submission above, e.g. spatial data, environmental data etc. In addition to such areas of research, CSIRO also carries out research in areas that are sensitive, including for national security and commercial reasons.

The data produced in CSIRO research thus may not always be appropriate for public release, either at all, or only after a period of time. This means that the framing of a of a mandatory publication regime for suitable categories of publicly funded research data must be carefully crafted. For example, data is often collected in research as a backbone to Government economic or social planning or for security. In these cases it is clearly important that open access to those data do not allow any commercial pre-empting of government decisions or impact on security. Thus the timeliness of releasing data may be crucial and in some areas may not be possible.

It is important for the legal and policy settings concerning publicly funded data to recognise and accommodate the need, in relevant areas, for intellectual property protection to drive practical benefits from publicly funded research, for example through encouraging investment in technology development and innovation. Accordingly, any obligations to publish publicly funded research data should be carefully framed so as to permit research agencies to appropriately sequence the pursuit of relevant intellectual property protection with data publication.

CSIRO's DAP guidelines ask researchers to consider issues that go to the ability and the desirability of making data available and the point to be made is that there needs to be a nuanced approach to mandating publication that could cut across other goals of fully or partially funded public research.

Data does not exist in isolation. The infrastructure required to generate, process and store data is significant, especially for large volumes or complex/critical assessments. Increasingly, rather than simply download or transfer data, it is necessary to be able to bring significant computing resources to the data location. This paradigm also brings significant opportunities to cross-analyse multiple datasets. In addition, data facilities can support a full ecosystem of associated services, software, tools, models, visualisations, publications etc. as well as associated expertise. Shared facilities have also proved effective in building inter-institutional relationships at all levels.

As an example the CSIRO DAP is an open repository that could both support broader data availability. Its infrastructure supports availability for over 7 peta bites of curated data. Data sets are accessible 24/7 and are currently being accessed by researchers around the globe. CSIRO would be happy to provide more details on the DAP its architecture and possibilities for greater use.

Peak or supercomputing capabilities are required to undertake advance simulation, massive data analysis (e.g. genomics) and experimental modelling. Australia's two current national facilities, NCI and Pawsey, are not currently set up for success with no agreement for capital refresh. In the absence of a national peak computing strategy Australia's ability to utilise data for economic benefit will be constrained. CSIRO supports a reimagining of this capability linking the two facilities as one to drive a consistent national strategy supporting the data and digital innovation agenda. This capability would not only directly support the research sector but also provide a peak computing capability for Australian industry. Thus reducing industry investment risk in this area by further developing the capability from the existing successful research peak computing community.

The Agriculture sector does not have a peak body for dealing with data issues. This could have a role in promoting data standards, protocols for data use, facilitating data sharing.

4.5 Data ecosystem

The biggest difference between enterprise and the external data world is the enterprise data ecosystem. This ecosystem develops as new enterprise IT systems create more transactional data. Over the last ten years there has been a move to connect systems for a range of efficiency reasons. The effect of this connectivity has been the creation of an internal data ecosystem. Within the ecosystem, rules for accessing the data, confidence in its accuracy and availability and prospective use are known. As a result Chief Information Officers, and Chief Data Officers can apply analytics and deliver insight linking data and developing insight with high levels of confidence.

Unlike the enterprise there is no national ecosystem for data. Government announcements about making its own data open are important cultural signifiers. However, this call has been made in the absence of an ecosystem even between Federal Government agencies themselves. The delivery of value more generally from data in Australia will require measures that support the development of a data ecosystem. Measures will include the introduction of frameworks for integration. This is not about delivering a single approach (we need to resist the temptation to attempt to reduce data management to a lowest common denominator, as implied in ToR 4). To be effective there is a need to support a networked ecosystem. An approach of a single system, storage platform, descriptive framework or licence would be to ignore the complexity of the types and usage of data. The challenge is to introduce the ability to make connections and to deliver consistent levels of assurance supporting insight.

4.6 Confidentialisation in facilitating data sharing while protecting privacy

There are a large number of proposed approaches to confidentialisation to facilitate data sharing for research while protecting privacy. All of these have been used in successful, large scale implementations in Australia and internationally, [O'Keefe and Rubin]. Relevant arrangements include:

- User agreements for offsite use (licensing), in which users are required to register with a custodian agency, and sign a user agreement, before receiving data to be analysed offsite.
- Remote analysis systems, in which the analyst submits statistical queries through an interface, analyses are carried out on the original data in a secure environment and the user then receives the (confidentialised) results of the analyses.
- Virtual Data Centres, which are similar to remote analysis systems, except that the user has full access to the data [68], and are similar to on-site data centres, except that access is over a secure link on the internet from the researcher's institution.
- Secure, on-site data centres, in which researchers access confidential data in secure, on-site research data centres.

Each setting makes available data at a specified level of detail, where detail can be reduced in personal data by methods including:

- Removal of identifying information
- Confidentialisation of the data by one of a range of methods, including aggregation, suppression or the addition of random "noise"
- Replacement of sensitive variables or data with synthetic (made-up) data.

Additional emerging methods for data sharing are described in section 1.2.

5. *Enhance and maintain individuals' and businesses' confidence and trust in the way data are used (ToR5)*

Data quality and appropriate use are key issues. Government and research providers often generate high quality, complex data products which sit alongside less accurate datasets from alternative providers which are delivered through more usable platforms or in more user friendly ways. In these instances, simplicity and modern web design can be mistaken for quality, resulting in the more authoritative data being underutilised or ignored altogether. There is a significant need for stakeholder engagement and education at the same time as simplifying and demystifying the delivery of complex data to address this issue.

Privacy, in particular is a key issue – it reduces the appetite for publication (often due to perceptions around what is appropriate, rather than a rigorous assessment of disclosure risk and utility) and it exposes those who do publish to risk (i.e. due to the Privacy Act).

A lot of datasets (public and private) are ethically fraught: minimal or tightly-bounded consents; personal and/or sensitive data; potential disclosure through cross-set linkage; etc. More work needs to be put in to defining privacy preserving frameworks and techniques, managing data security, and building a robust approach to ethics into the formational stages of data collection activities. Moreover, for non-expert data holders, they need thoughtful educational programs and support structures to enable proper publication procedures.

Access to data can and should varying according to the sensitivity of the data. At present rapid access is problematic and not timely. Establishing a framework for accessing sensitive data (e.g. medical data, endangered species locations, fishing grounds etc.) in a way that ensures authenticated use and timely access will assist research and, importantly, emergency response.

A recent report commissioned by the (UK) Economic and Social Research Council (ESRC) and the Office for National Statistics (ONS) described the findings from a public dialogue on administrative data and data linking [Cameron et al 2014]. Participants had little prior familiarity with or understanding of the purpose and processes of social research. Through the dialogue, they were shown information about research projects that have used linked administrative data, asked questions of experts, and heard the opinions of other participants.

An effective governance body is required for establishing the policies that will maximise the value of Australia's data assets. This will require representative expertise across government, industry and research stakeholders and social, economic and environmental lenses. A model for this across the research sector is ANDS.

There is ongoing discussions with the National Safety Agency based in Victoria and a few UAV/drone manufacturers around integrating data obtained through these devices closely with the computational models that CSIRO has developed around bushfires (and floods) to reduce levels of prediction uncertainty.

The report contains many insights into building, maintaining and enhancing the public's confidence and trust in the way data is used. For example, some of the relevant findings include:

- Participants' opinions of the ADRN plans changed, sometimes moving between negative and positive perceptions several times. The broad overall pattern was that as participants gained understanding of social research and the ADRN plans, they tended to become less concerned and more supportive.
- The findings suggest that the public would be broadly happy with administrative data linking for research projects provided
    i. those projects have social value, broadly defined
    ii. data is de-identified,

iii. data is kept secure, and

iv. businesses are not able to access the data for profit.

- The key area where there could be public concern about the UK Administrative Research Data Network (ADRN) plans is around de-identification. Confidence in the process by which this happens is crucial to creating support for linking administrative data. But the dialogue shows that this process is very difficult and time-consuming to explain.
- Transparency was seen as both important and sufficient, including ensuring public representation in the decision-making process.

Australian examples of good practice in communicating with the public about data use include:

- SA-NT Datalink animation, see https://www.santdatalink.org.au/animation
- Population Health Research Network (PHRN) case studies, see http://www.phrn.org.au/for-the-community/what-we-have-learnt/
- Population Health Research Network (PHRN) Consumer and Community Participation Policy, see http://www.phrn.org.au/for-the-community/consumer-and-community-participation-policy/

References

Australian Government Environmental Information Advisory Group (2012), Statement of Australian Requirements for Environmental Information, Bureau of Meteorology, Canberra. pp. 86 http://www.bom.gov.au/environment/Statement_AGREI_web.pdf

Morton, S. and Tinney, A. (2012) Independent review of Australian Government Environmental Information Activity final report, Canberra. https://www.environment.gov.au/system/files/resources/06e5e5b5-4584-4bd9-b2fd-05a790d0b2c4/files/eia-review-final-report.pdf

Belbin L, Williams KJ (2016) Toward a national bio-environmental data facility: Experiences from the Atlas of Living Australia. *International Journal of Geographic Information Sciences* 30(1), 108-125.

Box P. and Lemon D. (2015) The Role of Social Architecture in Information Infrastructure: A report for the National Environmental Information Infrastructure (NEII). CSIRO, Canberra, Australia. http://www.neii.gov.au/system/files/filedepot/1/The%20Role%20of%20Social%20Architecture%20in%20Information%20Infrastructure.pdf

Box P, Simons B, Cox S, Maguire S, (2015) A Data Specification Framework for the Foundation Spatial Data Framework. CSIRO, Australia. CSIRO. http://www.anzlic.gov.au/sites/default/files/files/FSDF-Data_Specification_Framework.pdf

Prakash M, Hilton J, Ramachandran L (2015) Integrated hydrodynamic and hydraulic modelling for evaluating future flood mitigation in Urban Environments, International Symposium on Environmental Software Systems, 282-292

Miller C, Hilton J, Sullivan A and Prakash M (2015) Spark- A bushfire spread prediction tool, International Symposium on Environmental Software Systems, 262-271

Cameron D, Pope S, Clemence M, Dialogue on Data, Report of the Ipsos MORI Social Research Institute, commissioned by UK Office for National Statistics and Economic & Social Research Council 2014 https://adrn.ac.uk/media/1245/sri-dialogue-on-data-2014.pdf

O'Keefe CM and Rubin DB, Individual Privacy versus Public Good: Protecting Confidentiality in Health Research, Statistics in Medicine 34 (2015), 3081-3103.

Appendix 1 – EUDM Work Findings

The following general set of findings from recent EUDM work is of relevance and is pinned to contemporary stakeholder engagement. Note that it provides insight relevant to both Government and private datasets.

In November 2015, EUDM ran a series of workshops across Australia with more than 30 participants from across research, government and industry sectors. The workshops focused on building an understanding of energy-data users, the role of energy data in their work, and the issues that are holding back progress in energy-data research. Key themes emerging from these workshops included:

- Privacy and confidentiality requirements limit the capacity for data holders to share key datasets.
- The provision of data is often slow and ad hoc, frequently requiring extended negotiation with data holders.
- Locating data is problematic and often relies on personal networks and connections.
- Existing data collections may be biased, poorly designed or lack statistical value.
- One-off trials, different project methodologies and variable data formatting limits the opportunities for data integration and linkage between datasets.
- Spatially and temporally fine-grained data is a priority, but is seldom available.
- A lack of transparency regarding data quality limits the confidence with which data can be used.
- Participants are enthusiastic about the potential added value that may result from better energy data, particularly in evidence-based forecasting, analysis and decision making.

If we want more detail, we engaged Energeia to consult with ten data holders from government, regulatory, retail, network and research sectors. We have highlighted 8 key findings from this engagement:

Legislation and policy

Data holders are subject to a host of legislative and internal policy requirements that can complicate data provision, and ultimately limit the breadth and type of data provided. Data on individuals is generally perceived to be the most fraught with challenges and considerations regarding privacy laws. As a result, many organisations are reluctant to provide such data. With relevant Acts and laws often carrying significant penalties (including imprisonment), this hesitation is understandable. The typical response from across the sector is to rely on coarse aggregations of data as a basic form of anonymisation and identity protection, despite potentially reducing the usefulness (utility) of the data. Technologies in confidential computing and encryption are currently under development to deal with these issues

Data quality

Quality across datasets is highly variable, with some organisations noting that the true quality of their data is ultimately unknown. Documented data quality assurance and validation processes also vary. The consequence is that while data may appear rich on first glance, its reliability and value are less certain. A public sector wide data infrastructure with clear format and standard guidelines would greatly reduce these issues.

Data format

Data is commonly provided in spreadsheet format. To do this, data holders often need to run queries in Oracle, MySQL, Access or proprietary databases, which inevitably complicates data provisioning. The resulting spreadsheet files may be massive, complicating data transfer (e.g. one 500-household spreadsheet was more than 350 MB in size).

The business case for data provision

It is clear that, for many organisations, the business case for data-provision needs to be clearly stated before it can become viable. In the simplest case, the business case amounts to cost recovery for efforts directly associated with providing data. However, some implicit costs in releasing data are more difficult to quantify, and may otherwise discourage data holders. Such costs include greater public scrutiny of data (and, implicitly, of the data holder) and the potential loss of competitive advantage.

Custodianship

The sector has vastly different approaches to data custodianship and management. These include dedicated teams with deep data knowledge, through individuals with significant personal investment in the data, to a complete lack of formal custodianship with ad-hoc data provision on demand. Custodianship is further complicated by discontinuous ownership of data and loss of institutional knowledge when data experts leave the organisation.

Subsequent use of data

The provision of data is often bound to specific uses or organisation types. For example, data holders may only provide particular datasets for non-commercial uses, or only allow access by government bodies. Wider distribution of data (with more permissive usage terms) often remains untested.

Processes and methods for managing data

Data-provision practices are far from standardised across the sector. They range from ad-hoc processes with limited oversight through to well-structured and documented institutional systems. Decision making may rest with a single individual, a custodianship team, or be subject to strict legal and organisational review. The consequence is that approaches to engagement and expectations of delivery times can vary considerably and result in underestimation in planning.

Value of data

Organisations frequently expressed a strong interest in sharing data, with a clear move towards a wider acceptance of open data practices. However, many organisations still have concerns about the burden of data provision.

Appendix 2 – Data about Australian Businesses

Regulatory data on Australian businesses

The scope of these datasets relates to information about individual businesses and related organisations (e.g. charities, non-government organisations, etc.). The primary datasets include the ASIC Company Register and Australian Business Register (ABR). There are however other important secondary datasets relating to Australian companies and related organisations (e.g. large taxpayers, government procurement, research funding, patents and trademarks and many more related to the regulatory functions of agencies such as employment, environment, etc.). These datasets are different to confidential survey information collected by the ABS for official statistical analysis.

The Australian Government has made good progress with releasing a limited part of the ABR for widespread industry use but could release further information that would be of high value to external users. With regard to the ASIC Registry, the current sale process being managed by the Department of Finance does not appear to have addressed the opportunity to create greater open and free public access to these datasets.

Australian Business Register (ABR)

The Australian Business Register (ABR) is the program that manages the establishment and operation of business identities through the Australian business number (ABN) system. The ABR also supports the Standard Business Reporting (SBR) program to standard the way businesses interact electronically with Australian government agencies. The ABR is administered by the Australian Tax Office (ATO) in partnership with other agencies such as the Department of Industry, Innovation and Science.

The ABR includes over 7.2 million ABNs covering companies, sole traders, partnerships, superannuation funds, trusts and government agencies. The ABN Lookup Service, that provides both website and API based searches, was used for just under 517 million searches during 2014/15.26 This high level of use is because the ABN Lookup service has become embedded in digital business systems such as a finance, invoicing, superannuation and procurement systems where the ABN is used to verify the identity of a company or other organisation.

The ABN Lookup Service, has become a major success story for government online services and a good example of Government-as-a-Platform, where government data about companies and other organisation can be used to enable new and improved services in the general economy.27

The ABN Lookup Service currently provides open access to a limited set of the data held by the ABR about a company or other organisation. This includes information such as entity name, ABN status (e.g. whether current and year of registration), entity type (e.g. public company), main business location (by postcode) and related business and trading names.

The ABR also includes other non-personal information about an entity such as standard industry code, description of main business activity, other business location. The ABR is also complementing ABN data with other information about an entity such as business size (based on revenue and employment numbers).

Access to this non-public ABR data is currently restricted to authorized government agencies (covering national, state/territory and local government jurisdictions) and authorized researchers. This access has

[26] Report of the Australian Business Registrar 2014–15, Australian Taxation Office, https://abr.gov.au/uploadedFiles/Content/Download_Files/ABRreport2014-15.pdf

[27] NICTA, New models for digital government: the role of service brokers in driving innovation https://www.nicta.com.au/content/.../02/2014-NewmodelsforDigitalGovernment.pdf

proved beneficial to other government agencies for their customer information systems and for emergency management where contact with businesses in a specific location is required. Access to the non-public ABR data is also used increasingly to inform regional economic models about the rate of business creation and growth of industry sectors.

There is an opportunity to open up more of the ABR information for wider public use as open data. The release of data such as an entity's standard industry code, description of main business activity, other business locations and business size would be extremely valuable for reuse in a range of business and research activities.

If there are concerns about the sensitive nature of some of this data, it can be modified to release it at a more aggregated level. For example, the release of public data on business size could be based on standard categories such as micro (less than 5 employees), small (5 to less than 20 employees), medium (20 to less than 200 employees) and large (200 or more employees).

ASIC Registry

The ASIC Registry is the technology based data management and processing business that supports the operations of ASIC's various business and related registers. It includes the Companies Register that maintains the filing and notification of business registrations, financial records and ownership details for Australian companies and related organisations.

It forms a critical part of Australia's economic infrastructure and is essential to the efficient operation of Australia's economy. The Australian Government describes it as "a critical part of Australia's economic infrastructure and is essential to the efficient operation of Australia's economy.[28]

While some high level information is available for free, access to the details of the ASIC Registers is currently managed by a number of information brokers that charge users to view records based on a fee structure determined by ASIC.[29] In 2014/25, ASIC's Registry business generated $725.9 million in revenue on a cost base of $64 million. The bulk of the income comes from the business registration and filings, with $58.2 million from public searches (approx. 8% of total income).[30]

As part of the May 2014 Budget, the Australian Government announced it would commission a scoping study into future ownership options for the operation and ownership of the registry functions of the ASIC Registry. In May 2015, the Minister for Finance announced that the Australian Government would proceed with the sale and "undertake a competitive tender process to test the market on the capacity of a private operator to upgrade and operate the ASIC registry'. The Government noted it would "maintain ownership of the base data". A set of high level objectives were outlined for the sale but these did not deal explicitly with key issues regarding access conditions, pricing or competition issues.[31]

The first stage of the competitive tender process was a registration of interest in June 2015 that attracted over 30 potential bidders to buy and operate the ASIC Registry. The second stage was an Expression of Interest process where the registered parties were invited to submit their proposals for the operation and upgrade of the Registry. The third stage, which commenced in December 2015, was an Indicative Bid process where bidders could submit proposals based on the Australian Government's proposed operating, legal, regulatory and economic framework for the ASIC Registry. A fourth stage involves short-listed bidders providing their final bids in August 2016.

[28] Department of Finance, "Australian Securities and Investments Commission (ASIC) Registry – FAQs", http://www.finance.gov.au/procurement/scoping-studies/asic-faqs/ [accessed 13 July 2016]

[29] Selected high level ASIC Registry data is available at both ASIC website via search and data.gov.au as a csv file.

[30] ASIC Annual Report, 2014-15, p.11 & p.134. http://download.asic.gov.au/media/3437945/asic-annual-report-2014-15-full.pdf [accessed 13 July 2016]

[31] Department of Finance, "Australian Securities and Investments Commission (ASIC) Registry – FAQs", http://www.finance.gov.au/procurement/scoping-studies/asic-faqs/ [accessed 13 July 2016]

The Department of Finance have not released any details about the conditions of the sale such as the proposed operating, legal, regulatory and economic framework for the ASIC Registry nor the names of the bidders. As such, it is not clear whether the sale process has adequately addressed key public policy issues about access conditions, pricing and competition issues.

The important public policy issue for the future operation of the ASIC Registry is how open and accessible is its information to other users, rather than whether it is operated by a public or private sector organisation.

The option of providing free and open access to all relevant company information on the ASIC Registry should be properly evaluated before the sale process is concluded. Similar company registers in the UK and New Zealand provide free access to full company records.[32] A recent study by the Open Data Institute identifies that only 3% of countries provide free access to ownership information about companies.[33]

Several groups have proposed that the detailed information on Australian companies in the ASIC Registry should be provided free of charge. This would make the ASIC Registry an open data repository, allowing far greater access to researchers, journalists and the general public. It would also enable other businesses to use this data in innovative ways to create new value-added services.

Jeffrey Knapp from the UNSW Business School has said the sale is likely to “entrench the current exorbitant fees for information” and goes against the notion of informing the public and investors.[34] The Australian Centre for Financial Studies has advocated an alternative model that would involve the future operator setting and receiving fees only from those required to lodge company and director registrations and notices, thus making access to company and director information free and open.[35]

Research and Development Grants and Offsets

Data providing details of R&D grants to or support involving specific research organisations, businesses and other organisations is more available as open data compared to other grant data. This data however could be made available in a more structured manner to make identification of organisations and individuals easier and more consistent. This includes the provision of ABNs as a unique identifier for businesses and other non-research organisations and Open Researcher Contributor ID (ORCHID) as a unique identifier for researchers.[36]

Data providing details of the R&D Tax Incentive to specific businesses is not currently available to the public. Given that these tax offsets amounts to just over $2 billion per year in Australian Government support for businesses (approximately 25% of the total support for R&D), there is an argument that such information should be made available as open data, in a manner similar to R&D grant data.[37]

[32] New Zealand Government Companies Office, www.business.govt.nz/companies
Companies House website, United Kingdom Government, https://beta.companieshouse.gov.uk/

[33] The Open Data Barometer Global Report – Second Edition: World Wide Web Foundation, January 2015, p. 6.
http://opendatabarometer.org/assets/downloads/Open%20Data%20Barometer%20-%20Global%20Report%20-%202nd%20Edition%20-%20PRINT.pdf

[34] UNSW Business School, "Privatising big data: Will the ASIC treasure trove be sold off or sold out?", September 17, 2014

[35] The Australian Centre for Financial Studies, Kevin Davis
Privatising Public Information: The Sale of the ASIC Business Registers, Friday, 4 September 2015
http://australiancentre.com.au/publication/privatising-public-information-the-sale-of-the-asic-business-registers/

[36] NHMRC and ARC Statement on Open Researcher and Contributor ID (ORCID)
NHMRC website www.nhmrc.gov.au/grants-funding/policy/nhmrc-and-arc-statement-open-researcher-and-contributor-id-orcid

[37] National Commission of Audit, February 2014. 10.2 Research and development
http://www.ncoa.gov.au/report/appendix-vol-2/10-2-research-and-development.html

The three main sources of R&D grants relevant to Australian businesses and other organisations include:

- Australian Research Council (ARC)

  The Australian Research Council (ARC), that administers the National Competitive Grants Program (NCGP), provides on its website open data on grants approved for research organisations. This includes a spreadsheet partner organisations associated with the grant program dating back to 2001.[38] This spreadsheet could be improved with the inclusion of ABNs to assist correctly identify the partner organisations, as well as information about the level and nature of industry support for the Linkage grant projects.

- National Health and Medical Research Council (NHMRC)

  The National Health and Medical Research Council (NHMRC), that administers Australia's medical and health research grants, provides on its website open data on grants approved for research organisations. This includes a spreadsheet showing recipients of the grant program dating back to 2000.[39] This spreadsheet could be improved with the inclusion of ABNs to assist correctly identify the partner organisations.

- Innovation Connections Grants

  The Department of Industry, Innovation and Science has funded a series of grant programs to support the placement of researchers in businesses, including the Researcher in Business, Research Connections and the current Innovation Connections program.[40] It would be valuable if the data on the companies that have been recipients under the program were provided as an open data set (including ABNs as a unique identifier).

[38] Grants Dataset, ARC website, http://www.arc.gov.au/grants-dataset

[39] Research funding statistics and data, NHMRC website, www.nhmrc.gov.au/grants-funding/research-funding-statistics-and-data

[40] Innovation Connections Programme, www.business.gov.au/assistance/innovation-connections