CIFR
Centre for International Finance and Regulation
Towards Financial System Integrity

Submission to the Productivity Commission

Data Availability and Use

July 2016

DATA AVAILABILITY AND USE

Author: Dr Kingsley Jones, CIFR Research Fellow

Contents

1 Forward

CIFR congratulates the Federal Government on initiating the Productivity Commission inquiry into the availability and use of public and private sector data by individuals and organisations in Australia. CIFR's submission is authored by Dr Kingsley Jones, a CIFR Research Fellow, and an expert in this domain.

We live in an information-based age that is characterised by rapid growth in the widespread use of technology. Associated with these developments is an exponential rate of growth in the creation of data.

Data can contribute to the wellbeing of society through improved decision making in a broad range of areas. For example, empowering consumers with more detailed information enables them (with appropriate tools) to make better-quality decisions in relation to all aspects of their lives. The Productivity Commission Issues Paper identifies the social benefits associated with increased availability and use of data. In particular, there are potential efficiency gains associated with the application of data-driven processes, as opposed to traditional and intuitive approaches. Moreover, increased access to data can serve to promote competition by facilitating the creation of new business opportunities. Also, increased data availability addresses information gaps, which, in turn, lays a foundation for increased innovation. Similarly, there are value opportunities inherent in the linking of datasets. Specifically, a creative ability to re-use data by linking it to other sources represents a potential opportunity to maximise its value. Importantly, any data capture must ensure that the data is collected in a format that is machine readable so that it can be utilised in a flexible manner preserving high integrity.

Although these prerequisites for efficient data usage necessarily involve development costs, these may be mitigated by, among others, the efficiency benefits inherent in a well-designed data architectural model.

At CIFR, we are vitally interested in data and its availability for research purposes. We are an independent research organisation that promotes and facilitates evidence-based research conducted with the aim of contributing to the development of a first class financial regulation and policy regime. High quality data is critically important in ensuring that stability of the financial system can be achieved by Australia's financial regulators.

There is a widespread view that Australia presently lags other developed countries in terms of a high-quality data architecture regime. It is vital to the national interest that we establish a best-of-breed data architecture system. This will further help to enhance the quality of research and will encourage innovation. It will also further encourage domestic researchers to use their skills and talents examining research questions that are applied to the Australian context, furthering our understanding of issues in our own backyard.

A best-of-breed data architecture system would be of immediate benefit to both Australian and international researchers, particularly academics, who are always on the lookout for new and innovative datasets that can maximise their research impact for publishing in the very best scholarly

journals. Usage of Australian data by such researchers promotes Australia and Australian assets in a very positive manner.

We recommend that consideration be given to carrying out an inventory or audit of current public and private data architecture, capture, storage and accessing arrangements that span Australia's financial regulators with a view to determining any linkages and networks they share. Due consideration of international best practice should also be applied here. This process may then be a step towards a solution that allows the safe and efficient sharing of data and intelligence within the broader regulatory arena. Independent bodies that effectively harness the research skills within the academic sector and practical nous of industry participants are well placed to make a useful contribution to evidence-based policy making if they were able to access such data.

Professor David R Gallagher
Chief Executive Officer
Centre for International Finance and Regulation

2 Executive Summary

Government policy recognizes that digital technologies for collecting, processing and analysing data pose very significant opportunities for enhanced innovation, efficiency and productivity across the economy[1]. However, while significant progress in Public Sector data management is being made[2], the policy implications are widened by the need to consider interactions between and among the private sector, the public sector and citizens in the dual roles of *data consumer* and *data contributor*. The role of Open Data and policies to promote data sharing with appropriate protections on privacy and security, are potentially of great significance in transforming the economics and efficiency of service delivery and discovery across the economy. This submission to the Productivity Commission Public Inquiry into Data Availability and Use[3] sets out responses to the specific questions posed, along with supporting background considerations deemed relevant to the subject of the Inquiry.

For economy of presentation, the front matter of the submission sets out the basic stance we take towards consideration of public and private Data Use and Availability as it relates more specifically to the conduct of research in finance, and the practice of effective financial regulation.

2.1 Origin of this submission

The Center for International Finance and Regulation (CIFR) has a particular remit:

> *To promote financial sector vibrancy, resilience and integrity, supporting Australia as a regional financial centre through leading research and education on systemic risk, financial market developments and market and regulatory performance*.

With this mission, CIFR has direct experience of the challenges facing the constituency of financial data consumers, data contributors and data custodians across commerce, regulation and research.

In our view, Australia has significant pockets of data and analytics excellence embedded within a vibrant and robust commercial finance sector supported by high quality regulation and research. The principal challenge we perceive is to make this data capability *better integrated* for purpose.

Government policy appears fully cognizant of the future economic value that might accrue from the encouragement of innovation in the emerging digital economy. However, it has been recognized that Australia has a relatively poor record of R&D commercialization and less than ideal patterns of collaborative effort between business, government and the research sector. This familiar past is best viewed as a prologue to a more joined-up future view of innovation as an economy-wide activity.

One metaphor for the current state would be to liken the data assets of the nation as residing upon well-tended continental plates that occasionally bump up against one another. The points of contact are sometimes points of friction because the necessary collaborative frameworks and policies have yet to be fully hammered out. However, the opportunity is huge because the underlying quality of the human and data resources is high. With this perspective, the principal challenges of policy are to smooth out some of the points of friction for: privacy protections; data licensing; data linking and sharing; data skills development and the agility of policy consultation and development processes.

[1] https://www.dpmc.gov.au/sites/default/files/publications/aust_govt_public_data_policy_statement_1.pdf
[2] https://www.dpmc.gov.au/sites/default/files/publications/public_sector_data_mgt_project.pdf
[3] http://www.pc.gov.au/inquiries/current/data-access

2.2 The Agile Path to Policy Development

We think it appropriate to best discover effective policy through *active doing* of the target activity: to promote *Data Use and Availability*. When a suitable collective goal is identified and a program to deliver against that is enacted, one soon discovers the material legislative, social, commercial and technology blockers of progress. Iteration of effort against the goal helps surface solutions and has the benefit of providing well-informed parties to the consultative process of framing new legislation.

One mechanism to propel such effort is to pose *Challenge Questions* to the community:

How might we reconcile the need for personal privacy with the benefits of ubiquitous data?

Community consultation can be an effective means to progress such questions, but it seems that such effort must be targeted. The question of privacy is very broad for the community at large. If it is not reduced in scope then the answers given are likely to be framed for the worst case scenario.

A more agile path to policy development might be to tackle more concrete questions:

How might citizens best decide who gets to read the data generated by them which is about them?

These are more concrete questions that can be usefully narrowed in scope to specific data items on a scale of accessibility. Who gets to know my phone number? Who gets to see my credit card details? In what circumstances must the consumer positively permission each data access instance?

This leads to even more targeted questions for consideration:

How might transactors permission read-only access to their transactional data?

Such narrower questions surface the attributes of data in a context of social behaviour. Data can be written, read, transmitted, reproduced, sold, erased, lost, stolen and breached. The importance of any one such activity with data depends greatly on the *context of trust* in which it occurs.

In summary, we think policy development can be framed around an experimental and consultative approach which pushes forward in those areas *where the above questions are clear*. One example of this transitionary policy approach is the Australian Securities and Investments Commission (ASIC) idea of a *regulatory sandbox* for emerging digital finance firms. The policy philosophy is to limit the scope and scale of activity during the experimental phase while a business idea is tested and proven.

On the other hand, there may be signs of potential market dysfunction due to opposing commercial interests adopting strategies that are clearly sub-optimal for the consumer. An example of this is the unclear status of bank customer ownership and rights to transactional data. While that data exists on service-provider systems, it is generated (authored) by the customer entering into a transaction. There are third-party services that might add value to this data when shared, but the terms imposed on data systems access might effectively defeat that without resort to the third-party impersonating the customer via access to their account passwords. This is clearly a situation fraught with risk. It would seem better to acknowledge the conflict and resolve to rectify the confusion at law. Proactive policy development of financial services data access and sharing protocols might facilitate the aims of efficiency and innovation without abnegating responsibility to the chaos of the marketplace[4].

[4] The alternative is to allow the sub-optimal solution to become an industry de-facto standard of behaviour.

2.3 Data Use and Availability: Costs and Benefits

Digital technology continues to drive the marginal cost of data acquisition and distribution towards zero. However, the hidden cost of data, in practical terms, arises from the *risk* which attaches to any sensitive data becoming public[5]. This might be reputational risk for personal data or competitive risk for data that is commercial-in-confidence. It follows that many of the core principles which attach to regulating the access rights and distribution rights to data are already familiar *within* organisations.

The new policy issues concern the linking and sharing of data *between* organisations. Mitigating risk in this area relates to establishing circles of trust, clear data licensing and rights to use, and the social conventions expressed in law through the Privacy Act 1988 and associated Privacy Principles.

In the area of financial data gathering, the Australian Prudential Regulation Authority (APRA) and the Australian Taxation Office (ATO) have well-established programs to the standardize reporting of data from contributors through initiatives such as the Small Business Reporting (SBR) standard. Emphasis on such standards improves data quality and potentially reduces the cost of regulatory compliance.

However, there are also costs to data gathering and reporting when they depart significantly from line of business activity. For this reason, it would likely be advantageous to strengthen the present efforts for standardizing whole of government digital information exchange with new developments in line-of-business commercial reporting systems, such as accounting, payroll and transactions. The nature of standards is that new standards compete with old standards, and so there is never any one perfect solution. However, the growth of internet standards for data exchange through Application Programming Interfaces (API) has made ubiquitous data interchange more feasible and could well replace the inefficiency of paper forms processing or their unstructured electronic equivalents.

The risk of data breaches is perhaps the key factor behind the need for privacy principles. Mitigating this risk can be addressed through better defined circles of trust that determine who needs to know a datum, at what level of granularity that data is released, and on what terms of use.

The cost of data acquisition has much to do with how well the information architecture is designed. Where electronic data acquisition is mandated, there are system costs of implementation, but also data quality and re-useability benefits, provided that the chosen data formats are widely employed.

Since particular industries and government departments have different data types and legacy data investments it seems unlikely that any one standard can be applicable. However, there are likely some broad data interoperability principles which can be established and Public-Private Partnerships formed within natural areas of interest such as: geospatial data; health records and financial records.

Within the finance sector, the SBR and its reliance on Extensible Business Reporting Language (XBRL) has provided interoperability benefits for those firms and agencies which have adopted it. Since SBR directly targets Data Availability and Use, in a set of standards, it offers further promise for innovation and efficiency gains in financial reporting[6]. Uptake of SBR has been limited due to the constraints of legacy accounting systems. However, the shift to cloud-based accounting systems that are SBR compliant may well make a general economy-wide transition easier at this time[7].

[5] The Ponemon Institute estimated the cost of data breaches at over $100 per record: http://ibm.co/1LmrS9q
[6] http://www.sbr.gov.au/about-sbr/news-listing/august-2015/digital-transformation-sbr-strategic-reports
[7] Resistance to the adoption of XBRL seems largely driven by the constraints of legacy accounting systems.

2.4 Vision for an Integrated Information Architecture

The information architecture of the Australian financial system is largely organized around four key elements: the payments system, as regulated by the Reserve Bank of Australia (RBA); the registry and licensing systems as regulated by ASIC; the tax reporting and Self-Managed Superannuation Fund (SMSF) systems, as regulated by the ATO; and Approved Depositary Institutions (ADI), the insurance sector and the non-SMSF superannuation funds, as regulated by APRA. The Australian Bureau of Statistics (ABS) collates economy wide data series, while the Australian Transaction Reports and Analysis Centre (AUSTRAC) is Australia's financial intelligence unit, having a regulatory responsibility for anti-money laundering and counter-terrorism financing investigations.

The information flows within this system represent an ideal model case for surfacing the key policy issues that naturally arise when considering these elements:

1. Appropriate circles of trust and confidence for access to data
2. Appropriate levels of granularity for sharing data
3. Protocols for anonymising and aggregating data
4. Protocols under which data may be linked
5. Data items of particular value for financial intelligence
6. Data items that pose particular risks for public disclosure or breach
7. Standards by which data is best gathered and stored for ease of use
8. Standards by which data access is best secured and protected when in transit.

It may be noted that finance data has particular attributes which make the appropriate design of the information architecture arguably less complex than health records and amenable to sharing within a model of anonymisation and aggregation of unit-record transactional data.

The decision value of high-frequency transactional data is potentially high, especially if the payments system were to be viewed as a real-time source of economy-wide statistical information. The public benefit of such high-frequency information, when widely reported in statistical aggregate, would likely be high, but the public discussion of this possibility has been limited. Presently, transactions in excess of $10,000 AUD are reportable to AUSTRAC for financial intelligence purposes.

Consideration of this possibility leads to a natural question:

How might we better use transactional systems in place of system-wide statistical survey activity?

The nature of digital transaction systems is that they capture almost all of the non-cash economic activity for planning and forecasting purposes. The design of the financial information architecture now naturally encompasses the question of how national statistics are gathered, how the system is regulated, and how privacy is to be respected. The payments system is a natural area of focus, through standards for permissioned access to customer records via Open Banking APIs[8]. Other examples include the reporting of Linked Employer-Employee Data (LEED) in New Zealand[9], and the US EDGAR and 13-F filing systems for financial statements and mutual fund holdings[10]. There seems to be scope for community discussion of the benefits of integrated information architecture.

[8] See UK HM Treasury report on *Data Sharing and Open Data for Banks*: http://bit.ly/1w5tO0U
[9] http://www.stats.govt.nz/leed.aspx
[10] See https://www.sec.gov/edgar.shtml and https://www.sec.gov/answers/form13f.htm

3 Recommendations

3.1 Recommendation: Revisit the status of public data linkage and proofs of benefit

In broad terms, the goal of enhanced efficiency and innovation appears to run into two main areas of policy friction. On the one hand, the existing Commonwealth Public Sector policies on data linkage adopt the base policy position that data should *not* be linked absent clear positive demonstration of public benefit. There are obvious privacy issues which vote in favour of this stance, but it might be usefully re-thought in those cases where the data type, or the protocols of collection, enables low risk data linkages and sharing, without requiring a proof of benefit. This approach would be more nuanced to the realities that some data types, such as spatial and registry information, are not particularly sensitive. It also opens the door to a more systematic approach to the development of durable national data assets, for research and policy development. Examples include longitudinal studies of health, welfare and employment, which are of significant value in achieving better social outcomes. The New Zealand *Integrated Data Infrastructure* is one example of this approach being applied in practise. The consultation procedures that were used in connection with developing this policy seem instructive for the Australian circumstance. The prior co-operation between both countries in developing and releasing important spatial data sets is also instructive of those policy development pathways which seem to have led to significant progress over short lead times.

3.2 Recommendation: Emphasize the value inherent in longitudinal data series

The principles of data-driven policy development and evidence-based regulatory intervention are strongly supportive of developing strategic longitudinal data series that measure the appropriate items for both public and private decision makers. The historical development of census and other time-series of economic performance speak to the practical merit of this stance. Inventories of data, as it exists today, might be usefully lined up with areas of perceived gaps. In finance and regulation, these include items such as: Linked Employer and Employment Data (LEED); Retirement Incomes and Requirements; Longitudinal studies of Business Formation and Entrepreneurship.

The creation of new longitudinal data series from existing data stores might be greatly facilitated by attention to the first recommendation. The current policy restrictions on data linkage appear to pose an unnecessarily high bar on projects that might seek to experiment with creating longitudinal data by matching records held by separate public and private custodians. The visible policy imperative in this area is effective public consultation and communication about the positive social benefits, with due attention to mitigating risks. The established activities of the Australian Institute of Health and Welfare (AIHW), in acting as a data integration agency and data governance partner for projects that involve sensitive health and welfare data is one possible best practise model to consider.

3.3 Recommendation: Promote Public-Private partnership on data standards

Contemporary trends in data infrastructure highlight a generational opportunity to re-consider how the public and private sectors exchange data in the normal run of administrative activity. There are presently unmeasured benefits and costs of moving to preferred digital modes of interaction where data is collected and validated at the source of service delivery and interaction. Web standards are increasingly the de facto choice for interoperability and this represents a natural focus for policy.

In simple terms, the world of filling in paper forms, posting forms and re-keying data from forms into electronic storage systems is now fading. Revisions to disclosure practices, communication practises and the interaction between these data gathering activities poses great *potential* benefit. However, if investments in systems and processes are not standardized in some key areas then there is a risk of data balkanisation. This may appear through a plethora of data channels and regulations which simply re-establish old areas of existing inefficiency through lack of coordination. There are obvious benefits from attempting the re-design of common touch points between the private and public sectors. Registrations, tax reporting, regulation and compliance are key examples[11].

3.4 Recommendation: Consider mandating key electronic financial disclosures

Once information production happens within an electronic environment, there are clear efficiency benefits from re-purposing data to new questions when provision has been made for that. However, in Australia there are numerous examples, such as financial reporting by public companies, where there is no clear legal status to the electronic filing, nor any stipulations on machine readability.

One best practise example is the US SEC EDGAR filings database, where corporate reporting is done in the machine readable XBRL format. Furthermore, the electronic filing is the legal filing document. Examples of this approach are likely to become more frequent as digital technology gains ground across most line of business applications. In Australia, the SBR standard is an example of this approach, but was not mandated. Through the growth of cloud accounting services, Public-Private partnerships between the ATO, ASIC and digital accounting firms might usefully promote micro-economic efficiency gains from electronic accounts processing. Where this touches the financial system, the promotion of safe and secure third-party read only access would assist digital entrepreneurs to develop new advisory services for business and consumers.

3.5 Recommendation: Promote skills development and data literacy

Government initiatives such as Gov 2.0[12], GovHack[13] and the Digital Transformation Office (DTO)[14], are consistent with private sector digital incubators and research sector activity to develop skills in service design, data science, data literacy, data architecture, data visualisation and communication.

There is community-wide interest in developing skills for innovation and entrepreneurship. This is reflected in a range of University courses in Data Science[15], Innovation[16] and Data Analytics[17]. While educational fads and fashions come and go, the element within the Data Science skills base which seems important to emphasize is a confluence of: coding skills; data modelling and analysis skills; and the skills to effectively communicate key findings to stake-holders. Over time, the transition to a digital economy might well see these skills as being separate aspects of specialization. However, in the early stages of building effective organizational cultures for data-driven analytics it is helpful to grow individuals with multiple over-lapping skillsets. One way to do this is via collaboration between teams with different domain knowledge, possibly from different industry backgrounds, and across academic disciplines. The US Agency 18F is an example of such team composition in practise.

[11] MyGov is a step in this direction: https://my.gov.au/LoginServices/main/login?execution=e1s1
[12] https://www.finance.gov.au/archive/policy-guides-procurement/gov20/
[13] https://www.govhack.org/
[14] https://www.dto.gov.au/
[15] http://sydney.edu.au/courses/master-of-data-science
[16] http://www.uts.edu.au/future-students/find-a-course/courses/c04293
[17] http://programsandcourses.anu.edu.au/2016/program/MADAN

4 Response to Inquiry Questions

The following sections provide detailed responses to the inquiry questions. Additional background material is relegated to two technical appendices. This material simply provides some conceptual background to the tensions we see active in policy development and the relationship between the technology trend and policy development. This may be read as an adjunct to the answers given.

4.1 Public Sector Data

QUESTIONS ON HIGH VALUE PUBLIC SECTOR DATA

What public sector datasets should be considered high-value data to the: business sector; research sector; academics; or the broader community?

There are a number of datasets that may be considered *foundational*, or platform datasets, since they constitute key reference data upon which data analytics and business processes are based. The obvious examples are geospatial reference data, describing the natural and built environment. There are many such datasets already exposed on data.gov.au and the recent release of the G-NAF address file[18] and Administrative Boundaries dataset[19] is an excellent example of high-value data that has been released in recognition of: community requests; overseas best practise and open data policy.

Other examples are natural resources datasets which have various degrees of openness. The South Australian minerals resources portal, SARIG[20], is an excellent example of best practise in releasing large quantities of survey and resource industry data in a rich cross-referenced mapping portal. There are many possibilities to build centres of excellence and collaboration in minerals and energy data archiving, analysis and distribution. Given the high natural endowment of Australia and the close possible co-operation with global centres of excellence, such as the US Geological Survey[21], the British Geological Survey[22], Geosciences Australia[23] and state-based bureaus of minerals and energy, a policy focus on driving release and integration of high value data could accelerate innovation.

Other examples of key platform data items, that have already been released, include registry and entity reference data such as business names and registrations, financial licensees and professional registers. These have significant value to business when integrated into workflows to anchor creation of effective industry reference indices and data for search, marketing and geo-analytics.

The value of public data in the above applications is clearest where it: has wide use; serves as a gold copy of truth for entity recognition and record linkage, and has an obvious public interface for ongoing collection through regulatory or supervisory activities.

There are other datasets which provide *research opportunities*, both public and private, since they contain measures of activity, relationships or survey items that would be difficult to gather on a private basis. Examples include census data, although some elements of this may be open to regular refinement using electronic survey methods. The key value of government-led comprehensive data gathering efforts is that they provide ground or reference truth for private efforts.

[18] https://data.gov.au/dataset/geocoded-national-address-file-g-naf
[19] https://data.gov.au/dataset/psma-administrative-boundaries
[20] https://sarig.pir.sa.gov.au/Map
[21] https://www.usgs.gov/
[22] http://www.bgs.ac.uk/
[23] http://www.ga.gov.au/

Examples include the development of high-frequency private indicators of economic activity using lower-frequency government survey data to make appropriate adjustments for differing sample methods. The Medicare Benefits Schedule (MBS)[24] items, and expenditures associated with those[25], provide valuable data to assist in developing health informatics and other managed care solutions. There are cost-benefit trade-offs between: prevention and cure; age and occupation-related risks; and services siting. With such datasets there is a clear and recognized need to preserve privacy and release only de-identified data. However, there are obvious opportunities for improved public health outcomes alongside the possible improvements in service efficiency and health-system productivity.

Educational outcomes data, labour and income studies[26] and statistical measures of social mobility and progress have important public policy uses alongside private benefit in providing ground truth for innovative measures of consumer demographic attributes from digital activity.

The payments system itself is a very promising source of data for public and private purposes. Digital transactions define two-way relations between payer and payee. These have important supervisory and regulatory functions, particularly for reportable transactions that are used to police the system.

However, when suitably de-identified, the transactional data (point-to-point) can provide incredibly rich insight into the patterns of economic activity. Two-sided digital marketplaces are showing the possible richness of the future data ecosystem in tracking: goods flows; labour hire patterns; and the readership and interest patterns of social media properties. Recognising these trends, it would seem prudent to consider pro-active policy to facilitate opening aggregated transactional network data for public, private and research use. Rich network data might assist all four areas: business; government; research and community. The issues are around technical feasibility for existing networks, the protocols for de-identification, levels of aggregation and privacy protection.

The challenge and opportunity of digital data collection is to refine the sampling, sample adjustment, aggregation and de-identification methods to enable reliable real-time indicators. This may provide future public health and consumer choice benefit to complement the obvious benefits to come from the sharing sensor data in areas like *weather*; *transport telematics* and *utility network operations*.

What characteristics define high-value datasets?

High-value datasets typically have situational or decision-making value. They are data items where the *price-of-ignorance* is potentially very high. For geospatial, geotechnical and hydrology data[27] the price of ignorance might involve construction activity in a poor location, or with heightened climatic or disaster risk alongside the more mundane reliable delivery of parcels to street addresses. Health data that captures pre-cursor patterns to potentially expensive health interventions might lead to lower hospitalisation rates, better consumer quality of life and lower system costs. Predictions of consumer demand may lead to higher operating efficiency in goods and services supply chains due to lower inventory, less wasted stock and a more consumer-responsive marketplace. Regulatory data on past enforcement actions may help guide effort towards the more objectively risky behaviour, whether in private sector compliance or public sector agencies.

[24] http://www.mbsonline.gov.au/internet/mbsonline/publishing.nsf/Content/Home
[25] http://medicarestatistics.humanservices.gov.au/statistics/mbs_item.jsp
[26] https://www.melbourneinstitute.com/hilda/
[27] http://www.anzlic.gov.au/

What benefits would the community derive from increasing the availability and use of public sector data?

Examples of benefits are tracked across many countries in annual indices by the Open Knowledge Foundation (OKFN)[28] at their Open Data Index survey website[29]. In Australia, one notable example is the improved transparency of government spending since the annual Budget data was released[30]. The evidence of enhanced innovation activity seems clear from the release of transport timetables and geospatial data. These have supported state-based transport apps for mobile phones, and also the creation of the National Map initiative[31]. This permits free and easy linkage of spatial data.

Another noteworthy example is the South Australian government SARIG data-hub[32], which exposes a rich linked database of mineral tenements, land boundaries, infrastructure and geophysical data. In the hands of exploration geologists, the tool enables desktop research on existing tenements, those up for lease, prospectivity from state survey data, logged drill holes, seismic and other data.

QUESTIONS ON COLLECTION AND RELEASE OF PUBLIC SECTOR DATA

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

It would seem to be a combination of skills-gaps and inertia due to fears over privacy and security, or other internal policy tensions. The Open Data Institute has identified *skills development* as a priority area which led to the development of the Web Foundation Open Data Labs[33]. In a global context, it is clear that education, community building and skills development are key to maximizing the potential of open data[34]. It is worthwhile to note, Australia rated #2 overall in Government Readiness, for use of Open Data, in the most recent Open Data Barometer[35], but #10 in Entrepreneur Readiness and #13 in Social Readiness. It is therefore important to develop the public-private system connectivity and knowledge of available data resources. The GovHack[36] competition, running each year since 2012, has helped to develop interest in using public data to develop new private sector applications.

The other limiting factor is existing revenue streams from data sales through infomediary firms. The prime example of this is Land Title data. It is common for State-based titles offices to charge for this data[37]. Wherever the revenue generation is substantial, such as Australian Securities & Investment Commission company searches[38], or Australian Federal Police (AFP) background checks[39], there will be resistance to opening data. In the case of police checks this seems warranted. The nominal fee probably prevents over-use of criminal background checks when they are not really required. For the ASIC registry, some items are already free, such as the Financial Advisers register[40]. There are widely different patterns of charging for data items due to a range of legacy publishing arrangements.

[28] https://okfn.org/
[29] http://index.okfn.org/dataset/
[30] http://theopenbudget.org/
[31] https://nationalmap.gov.au/
[32] https://sarig.pir.sa.gov.au/Map
[33] http://labs.webfoundation.org/
[34] http://opendatabarometer.org/2ndEdition/summary/index.html
[35] http://opendatabarometer.org/2ndEdition/analysis/readiness.html
[36] https://www.govhack.org/
[37] http://www.lpi.nsw.gov.au/land_titles/access_titling_info
[38] http://asic.gov.au/
[39] http://www.afp.gov.au/what-we-do/police-checks/national-police-checks
[40] https://www.data.gov.au/dataset/asic-financial-adviser

How could governments use their own data collections more efficiently and effectively?

Through using the federated data capability[41] of the existing CKAN data-hubs to enable API-driven data sharing between agencies. The systems in place support this capability but it is not widely used. This may be due to unfamiliarity with how to architect and operate federated data networks. Global best practise is indicated by the World Bank data-hub[42] which federates many data sources.

The opening of public research data sets has considerable potential. For instance, the academic data repository SIRCA[43] contains extensive price and volume information for the Australian equities market, but this information has historically been available only to academic researchers. Policies that encourage shared public-private data services might help fund facilities, while enabling greater innovation in financial trading. The Australian high-frequency trading industry is small, and this may be in part due to the low levels of access to tick data in private sector firms that are not already part of a global operation. This would appear to have limited broad skills transfer in automated trading.

Should the collection, sharing and release of public sector data be standardised? What would be the benefits and costs of standardising? What would standards that are 'fit for purpose' look like?

CKAN open source software has become the global de-facto standard to power data hubs. It is in use at the Commonwealth[44], most States[45], the UK[46], the EU[47], the US Government[48], the NOAA[49], and the US EDX Energy Data Exchange[50], among many others. There is great benefit in such standardization. It provides standard set-up routines, a global pool of skilled data infrastructure integrators, and the guidance of existing data cataloguing, indexing, search and API solutions[51].

Not all types of data are suited to CKAN repositories but they are a good general purpose solution. In other, more specialized purposes the use of web-standards, like RSS feeds and push notification, or Content Delivery Networks (CDN) for files, video portals and other tools can be useful. Very large data items, such as seismic data sets, or extensive imaging datasets are best stored within computer facilities where they are likely to be used. There is a national infrastructure of high-bandwidth fibre backbone, through the AARNet[52] and other research networks such as VERNet[53]. Historically, these facilitated high-bandwidth data exchange where the public network was insufficient. Now that there is the National Broadband Network (NBN), and private inter-cloud routes such as Amazon Web Services Direct Connect[54], and Microsoft Azure Express Route[55], it makes sense to think of national

[41] Federated data means a method to interlink data-hubs: http://ckan.org/features-1/federate/
[42] The World Bank federates a very large number of data sources: http://data.worldbank.org/
[43] http://www.sirca.org.au/products/
[44] The Commonwealth uses CKAN: http://data.gov.au/
[45] NSW: http://data.nsw.gov.au/, Victoria: https://www.data.vic.gov.au/, Queensland: https://data.qld.gov.au/, South Australia: https://data.sa.gov.au/, and Western Australia: http://data.wa.gov.au/ all use CKAN.
[46] https://data.gov.uk/
[47] http://open-data.europa.eu/en/data/
[48] http://data.gov/
[49] https://data.noaa.gov/dataset
[50] https://edx.netl.doe.gov/
[51] https://okfn.org/
[52] https://www.aarnet.edu.au/
[53] http://www.vernet.com.au/
[54] http://aws.amazon.com/directconnect/details/
[55] https://azure.microsoft.com/en-us/services/expressroute/

freeways for data[56]. Some are private, some are public and they interlink as part of the internet. The sectors of interest are likely: health and medical imaging; Internet of Things (IoT) sensor traffic, remote sensing data, and geophysical, mineralogical and petrological data[57].

In addition to CKAN, there is a Drupal version termed DKAN, which has been used in the USA for those sites which already had Drupal content management systems in place. Understandably, there is a wide array of commercial and open source content management systems that can be and have been used to manage data portals[58]. In Australian government circles, the CKAN system appears to be the de-facto standard, but it does not suit every use case. For example, in the scientific and research data space, it is common to have very large data sets stored with computational facilities that generated them. The CSIRO Data Access Portal[59] and National Computational Infrastructure[60] sites are good examples, along with SARIG[61] for minerals and energy data in South Australia.

What criteria and decision-making tools do government agencies use to decide which public sector data to make publicly available and how much processing to undertake before it is released?

This question is best addressed by the agencies in question. However, from a policy standpoint, one of the clearer examples of the criteria, process and decision-making is the Australian Institute of Health and Welfare (AIHW)[62]. Among other features of the AIHW protocol are guidance items on the policies and procedures attending data linkage for health and welfare data items[63]. The procedures described are in accordance with the 2010 Australian Declaration of Open Government[64], according to a developed and publicly available Data Governance Framework[65]. There are stipulations relating to public interest, privacy and best practise in accordance with established guidelines[66]. The AIHW is a good example, in our view, of international best practise for record linkage of sensitive data, and follows the Commonwealth practise of supervising governance through an Integrating Authority[67].

The other noteworthy example is the Australian Prudential Regulation Authority (APRA), which has a well-developed data governance framework that addresses data gathering, data quality and public consultation protocols[68] prior to the release of data items collected under their mandate. APRA data are widely shared within other government agencies such as the ABS, RBA, OECD and IMF[69].

One takeaway from the observed behaviour of AIHW and APRA is that transparency in process and a consultative mechanism does appear to promote trust and confidence in the release of data. When there is confidence in the protocols by which data is gathered and released it is more likely, in our view, to be actively used. This is because effective governance confers confidence for data usage.

---

[56] https://researchdata.ands.org.au/
[57] http://www.ga.gov.au/
[58] http://www.digitalgov.gov/resources/content-management-systems-used-by-government-agencies/
[59] https://confluence.csiro.au/public/daphelp/data-access-portal-functional-overview
[60] http://nci.org.au/
[61] https://sarig.pir.sa.gov.au/Map

[62] http://www.aihw.gov.au/home/
[63] http://www.aihw.gov.au/data-linking/
[64] http://www.finance.gov.au/blog/2010/07/16/declaration-open-government/
[65] http://www.aihw.gov.au/data-governance-framework/
[66] http://www.nss.gov.au/nss/home.nsf/NSS/085988C74EAFAB58CA2577F20017118A?opendocument
[67] http://www.aihw.gov.au/data-linking/integrating-authority/
[68] http://www.apra.gov.au/AboutAPRA/Pages/Policy.aspx
[69] http://www.apra.gov.au/statistics/Pages/default.aspx

What specific government initiatives (whether Australian Government, state, territory or local government, or overseas jurisdictions) have been particularly effective in improving data access and use?

As indicated above, the widespread adoption of the CKAN software has enabled some standards to be put in place for data-access, even though the software permits extensive look-and-feel tweaks. The state and federal CKAN data hubs may look different, but the data storage and access layers are on common technology. This assists interoperability between Australian hubs and also global ones.

QUESTIONS ON DATA LINKAGE

Which datasets, if linked or coordinated across public sector agencies, would be of high value to the community, and how could they be used?

The minerals and energy industry operates on a state based system of leases and tenements, with the support of national bodies such as Geosciences Australia. Developing methods and protocols for better linking and exposing such reference datasets could lead to global business opportunities. The Australian research agencies have significant digital assets and expertise in compiling and using highly technical datasets for construction, mining, oil exploration and remote sensing.

While these datasets are expensive to gather and maintain, they form valuable digital capital on which to build higher level data analytics. This seems to be an area of significant future potential.

The AIHW health and welfare datasets clearly have significant value for longitudinal studies, as is evidenced by the National Mortality Database[70], and the Cancer datasets[71]. There are public health statistics through Medicare[72], and many other sources such as the Medical Benefits Schedule[73]. One of the larger issues confronting effective policy development in this area is developing effective means for *discovering opportunities* for effective linkage between public service departments.

It is our impression, that the traditional mindset has treated data linkage as a task requiring proof of benefit. This may have retarded publicly useful data linkage efforts which have no great privacy risk, but where the ultimate benefits are hard to discern *up front*. The value of releasing data items into a reasonably standardized data portal infrastructure is that both public and private organizations can very easily discover the data and experiment, at lower initial cost, in discovering value within data.

In the USA, the federal government established the 18F team[74] as an internal digital services strategy consultancy to assist with making sense of data integrations, migrations and digital services. There are now many other such organisations such as Civic Actions[75], which assist governments. The UK has mirrored this pattern with the UK Government Digital Service[76] and the Open Data Institute[77]. In Australia, this trend is apparent with the establishment of the Digital Transformation Office[78].

Ultimately, the question of what is of highest value to the community is likely evident from three main sources. The first is what specialised data the community demands to be released. Here, the February 2016 release of the G-NAF address file conforms to this principle, having been ranked

[70] http://www.aihw.gov.au/deaths/aihw-deaths-data/
[71] http://www.aihw.gov.au/cancer-data/
[72] http://www.health.gov.au/medicarestats
[73] http://www.mbsonline.gov.au/internet/mbsonline/publishing.nsf/Content/Home
[74] https://18f.gsa.gov/
[75] https://civicactions.com/
[76] https://gds.blog.gov.uk/
[77] http://theodi.org/our-network
[78] https://www.dto.gov.au/

highly in requests on data.gov.au. The second is the obvious areas of public benefit: better health, wealth, welfare and education. The last data items are the common concerns of planners: transport data, a clean environment, use of land and natural resources and the value of security.

Which rules, regulations or policies create unnecessary or excessive barriers to linking datasets?

Privacy regulations impose limits on the linking of datasets.

How can Australia's government agencies improve their sharing and linking of public sector data? What lessons or examples from overseas should be considered?

The Open Data Institute in the UK has championed Linked Open Data as a creator of public and private value. A good example of linked data in action is the firm Open Corporates[79], which operates a global search engine across over 90M private firms gathered from public registry information.

The national registry agencies have a logical local focus upon their own jurisdiction. However, this led to a gap in the global linking of companies, parents and subsidiaries, across borders. Through the use of open data linkage, indexing and search technology, Open Corporates is able to provide global search services where previously there were none.

There are valuable lessons in such case studies for how the public-private benefit of open data operates in practise. Energetic entrepreneurial teams can spot hidden value in raw public data and exploit data linkage methods to expose that. However, it is important to have licensing terms that are transparent to such linking and sharing activities. Many of the key issues have been summarised in the recent Public Service Data Management Project Report[80]. The primary blocker at present is the lack of agility in response to rapidly changing opportunities to link and exploit data alongside a traditional, and understandable, risk management process to limit linkage unless public benefit can be demonstrated. This may well be overkill in those circumstances where the release of data is very unlikely to harm anybody. Some streamlining of established protocols and procedures seems to be warranted, especially if this is done in an incremental and agile fashion starting with priority areas where the harm of possible streamlined protocols is effectively minimized[81].

4.2 Private Sector Data

QUESTIONS ON HIGH VALUE PRIVATE SECTOR DATA

What private sector datasets should be considered high-value data to: public policy; researchers and academics; other private sector entities; or the broader community? In each case cited, what characteristics define such datasets?

The value of a dataset depends on the perspective. Firstly, the ability to charge high fees for access to data demonstrates value to the owner of the dataset. Secondly, there is value in the hands of those who creatively re-use data by linking it to other sources, making it searchable or providing aggregates. Finally, there is value in the reference data which forms a matter of public record for making sense of past events. High-value data is clearly present where the cost of ignorance is high[82].

[79] https://opencorporates.com/
[80] https://www.dpmc.gov.au/sites/default/files/publications/public_sector_data_mgt_project.pdf
[81] For example, Data.gov.au, the Australian central repository currently links to 6,700 datasets, compared to 25,461 in the UK and 132,865 in the US. Data is valuable when used. To be used it must be accessible.
[82] Geospatial data has high value for this reason. Digging a trench across an electric cable can prove expensive.

In broad terms, geospatial and telematics data has high value because it saves money and time and avoids costly error. This was recognized with the February 2016 release of the G-NAF dataset, which had previously been sold commercially. While the free data is raw, the value-added data is still sold by commercial providers. This is reasonable since it encourages new entrants to compete on adding value to the raw data release. Other high-value private data is the detailed consumer-graphic data comprising the purchasing behaviour of shoppers online and offline along with social media. Some of this information might (one day) be incorporated into economy-wide statistical surveys.

In the USA, the Center for Disease Control and Prevention (CDC) made use of Google search trends data to track the onset of flu outbreaks and other maladies[83]. The key value of Google data for the CDC was its real-time character, being some weeks in advance of old-style surveys.

Wearable device firms, such as FitBit, are now compiling substantial data on heart rate and exercise regimens for a large cross-section of society across many nations. The development of such datasets represents a real-time experiment in social health and the benefits of exercise. It further contains some valuable behavioural research data concerning which types of feedback to users have proven most effective in promoting healthy levels of physical activity. There are clearly many opportunities for public-private collaboration in making full use of such "natural social experiments".

In Australia, there is public policy interest in private sector innovation and the performance of firms in commercialising innovations, both public and private. There is also an emerging private sector focus on venture capital, university commercialisation programs and intellectual property. Locating the centres of activity and excellence becomes easier over time, but there remain stark gaps in data linkage across the different touch-points of the innovation economy. Thoughtful cross-linkage from the publicly released registry data of corporate entities into IP Australia, the private sector, and the university research grants system would likely aid research into early-stage business activity. There is not likely a requirement to gather new data, but there may be significant benefit to linking data.

For the private sector, the linkage of their own data systems into such a public backbone of registry information would promote release of the appropriate and valuable insights. Where is innovation activity gaining most traction? What are the typical funding sources and their capital structure? Which skills are most in demand? What rates of return are being experienced on new investment?

Reported *financial data* and *corporate entity information*, such as *registration details*, *directors* and *capital structure* are important high-value data items. This may be judged from the price that is now charged in the market for access to digital records. A simple data request from an ASIC infomediary may cost $50 or more, for a single PDF file. In the age of paper records, with printing and distribution costs, this pricing made sense. However, in the digital age, such high unit record costs impede the creation of company analytics at scale, to analyse the financial performance and credit worthiness of large sections of the economy. Inappropriate digital pricing is likely impeding the development of new services that were not previously possible with traditional print media.

Past case studies have indicated that data access can be highly elastic to price changes. For instance, the BEV in Austria experienced a 7000% increase in data requests when the price of data requests was reduced by 97% for an overall revenue increase of 46% during the course of the study.

[83] Joint research between Google and the CDC, see: Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski & Larry Brilliant , *Detecting influenza epidemics using search engine query data*, Nature 457, 1012-1014 (19 February 2009) doi:10.1038/nature07634

Pricing access to high-value data requires a balance between maximising the utility of data in active use, and re-use, versus the disutility to the data owner or generator of lower fees per access event. The important principle which seems to drive natural outcomes is the elasticity of demand. Greater attention to the standardisation of licenses for use and re-use may help in this area.

In the case of Open Data, the creation of Creative Commons licenses[84] has given considerable clarity to the distribution and use of open data. For important verticals, such as finance and insurance, the creation of industry standards in licensing may help accelerate value creation from data in use. An excellent example of this dynamic playing out in the market is the case of Pete Warden[85], who built a web-scraper to crawl the Facebook network back in 2010. Over the course of some hours, the bots accumulated a large trove of data amounting to around 200 million Facebook users, at minimal cost. The intention behind this exercise was to release an anonymized version of the data for research purposes and to promote Warden's skills in data collection. Facebook responded with a cease and desist order, and eventually Warden relented and agreed to delete his data set.

The mechanism Facebook now uses to monetise social network data is Facebook for Developers[86]. The low cost to scrape the web is matched by a low cost to serve an API. However, the resulting micropayments can add up to a very significant revenue stream, for Facebook, while enabling an ecosystem of innovative analytics firms.

What would be the public policy rationale for any associated government intervention?

Intervention might be appropriate if the data-gathering of private firms violate privacy principles. Another case is when the data collected has some clear public interest purpose. Airliners carry flight recorders, by law, to better determine the cause of accidents. There may come a time when similar requirements are expanded from trucking fleets to private motor vehicles. In general the rational for such mandated release of data is competition policy, public safety, or a risk mitigation purpose.

What benefits would the community derive from increasing the availability and use of private sector data?

The value of data is generally traceable to improved decision making. Private sector data may assist the community to make better decisions: where to live; what to study, what to eat, buy, and so on. The greatest public-private benefit likely accrues in just those circumstances where the data is most often personally sensitive: transactional data; health data; insurance risk data and financial data. The development of suitable policies to protect against misuse and breach of confidentiality is therefore of key importance in realising the benefit of greater use and access to private data. For example, the use of screen-scraping of financial data poses potential data breach risks that could be mitigated by the wider availability of secure Application Programming Interfaces (APIs)[87].

QUESTIONS ON ACCESS TO PRIVATE SECTOR DATA

[84] For the Australian site see: http://creativecommons.org.au/, where there are two versions available. There are the standard international licenses and a "ported" Australian version of the same licenses.
[85] This case study is from the Bloomberg Law article: *Use of Online Data in the Big Data Era: Legal Issues Raised by the Use of Web Crawling and Scraping Tools For Analytics Purposes* (Snell and Care 2013)
[86] https://developers.facebook.com/
[87] For opinion on the API stand-off in the USA, see Brian Peters in the American Banker (January 4th 2016). http://www.americanbanker.com/bankthink/if-banks-fear-screen-scraping-why-are-they-fighting-the-alternative-1078565-1.html?zkPrintable=true

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

There are some patterns of data gathering and dissemination where restrictive rights to use the data are imposed at the *point of distribution* without clear ownership. One example is the Australian Securities Exchange (ASX) distribution of Market Announcements. The Terms of Use that are associated with the release of that information restrict secondary use of the information[88]. Listed companies submit market announcements under their continuous disclosure obligations. It would seem that the filing company is the owner of the copyright, in so far as they produce the document, but that the ASX imposes a restrictive right of use as part of its legislated role to disseminate release of market sensitive information. In other jurisdictions, like the USA, company filings are a matter of public record and are maintained as free open data in a public repository, EDGAR[89].

While there seems no reason to insist on a public repository, the law might usefully address the fair framing of Terms of Use where there is a *public interest archival and research purpose*. For instance, the ASX might offer unrestricted dissemination of market releases after a delay, of say fifteen minutes. This allows the ASX to continue to generate revenue from low-latency news feeds while allowing mining of documents for financial statements, directors, dividends, and such like.

Similarly the slow progress to date of the Small Business Reporting standard[90] calls attention to the lost opportunities in productivity from a reluctance to mandate certain data standardization efforts.

The framing of data Terms of Use is particularly important for comparison shopping services that must gather up-to-date pricing information from different websites. Such services typically use data mining via automated web-scrapers and spiders. There are web standards in place, such as the use of Robots.txt files, to set out appropriate terms of access for robots. However, there are many issues of competition policy which surface when balancing the rights of web property owners to restrict use of data and the public benefit of transparent price comparison in otherwise opaque markets[91].

While does it cost time, money and effort to assemble a data service, the Terms of Use ought not, in our view, to restrict creation of alternate *non-competing* derivative uses of the data. In such a framework, there may be some legitimate (small) charge levied for heavy usage of a web-service, but that needs to fit with the economics of the service delivery on a marginal cost basis. There is likely a role for competition policy in developing a policy framework for *Commercial Terms of Reuse*, over the terms of use. Ideally, these might be standardized.

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

The provision of line-of-business data presents legitimate concerns of loss of competitive advantage. This may be outweighed in circumstances where greater visibility and availability of data can lead to higher operating efficiency across the economy as a whole, or the avoidance of risks.

The two obvious areas of private-sector data availability where we consider increased exposure and transparency may be beneficial are: customer credit records in banking and fund industry portfolio holdings data. In the first case, the Comprehensive Credit Reporting (CCR) regime was put in place

[88] See for example, the ASX terms of use: http://www.asx.com.au/about/terms-use.htm
[89] https://www.sec.gov/edgar.shtml
[90] http://www.sbr.gov.au/
[91] Of course Google depends on web-crawling robots to produce the index which powers search.

for the express reason of encouraging better credit risk pricing for consumers. There are unarguable benefits from a fairer and more accurate pricing of individual credit risk which we consider to outweigh the short-term competition issues for those institutions with large data holdings.

If CCR data were to be made generally available, we see potential benefits for incumbents in the development of a competitive market-place for credit-risk pricing tools. The financial benefit of superior credit risk pricing is shared between the quality borrower, in better interest rates, and the lender, through the reduced default rate and lesser need for cross-subsidies. This benefit will be captured by the lender no matter which agency is the source of the credit risk assessment.

In the case of funds industry portfolio holdings, the US industry has filed these quarterly in form 13-F for a long period of time with no obvious detriment to the filers. However, the increased portfolio transparency enables the development of improved fund manager research services. Transparency of fund ownership is good for the end investor, particularly given that Australia has a compulsory retirement incomes savings scheme. It seems anachronistic that savers are compelled to invest into funds that are not obliged to regularly report their own portfolio holdings except in summary form.

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

Private sales of data services are commonplace and generally fit the pattern of data extracts. When these are in non-machine readable form the data is not readily re-purposed. Private sector concerns to preserve existing competitive advantage are not likely, in our view, to be met by legislation.

Private interests will often choose to release sensitive data in forms that discourage active use when it is in their interest to do so. For instance, it is quite common for fund managers to enter notices of substantial shareholdings in barely legible handwritten script to discourage mining of the data. This seems to be against the spirit of the intended disclosure requirements. Policy discussion might include efforts to mandate workable standards in data format. One of the key considerations of such standards is durability and transformability against inevitable changes in future technology[92].

Legislated data release requires some focus on consultation with data contributors and attention to the useability of the information for data consumers. The Australian Prudential Regulation Authority (APRA) has powers to compel data collection, and a long-standing set of protocols and procedures to regulate release of the data items gathered. They conduct regular public consultations to seek input on the case for and against the public release of contributed data. The organization is also involved in the regular release of research datasets to the academic community with controls on disclosure and dissemination of that data to third parties. This seems like an effective policy model to emulate.

Should the collection, sharing and release of private sector data be standardised in some way? How could this be done and what would be the benefits and costs? What would standards that are 'fit for purpose' look like?

There are some identifiable circumstances where legislative change for standardisation would help. Financial reporting of public listed companies is typically via non-machine readable PDF files. This is inefficient for re-use, search and archival purposes. It would be better to mandate a scheme similar

[92] In financial reporting, the Extensible Business Reporting Language (XBRL) is one such standard. While the standard is based on Extensible Markup Language (XML), which is less favoured today over lightweight text protocols like Javascript Object Notation (JSON), these are inter-convertible https://www.xbrl.org/xbrl-json-making-xbrl-easier/ . The key considerations of data archiving are an open standard and transformability.

to EDGAR in the United States, where the legal filing is the electronic machine-readable file in the Extensible Business Reporting Language (XBRL) format. Regulatory filings of all kinds would be easier to process and archive if submitted in this form. However, there are costs involved in re-tooling form entry that suggest a highly targeted program of standardization where the volume and value of data release is sufficient to justify the costs of transitioning to new systems of data collection.

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

Third party intermediaries are already active in buying and selling data. The primary restrictions on the availability and use of data are: cost of access; useability of format and clarity of data licensing.

Once data is released, it is the *licence* for use which governs uptake. To maximise availability and use open data licenses like Creative Commons are effective since they provide indemnity to the data user against possible legal risks due to the otherwise unclear provenance of data in the wild. Firms that release open data, voluntarily, will see it used, so long as the data has value in use.

The primary risk to business appears to be that of third-party data breaches from data sharing[93].

Would such voluntary arrangements raise competition issues? How might this change if private sector information sharing were mandated? Is authorisation (under the Competition and Consumer Act 2010) relevant?

If there are clear competition issues attending the voluntary release of data then it would seem that those who have data of competitive value will voluntarily choose *not* to release it.

We have difficulty comprehending why voluntary programs to release data could ever have serious uptake when there are major competition issues involved. In such circumstances, mandated release would seem to be appropriate if there is a clear public interest reason to do that.

What role can governments usefully play in promoting the wider availability of private datasets that have the potential to deliver substantial spill over benefits?

Governments can lead by example. The recent release of the G-NAF address file is a great example.

How can the sharing and linking of private sector data be improved in Australia? What lessons or examples from overseas should be considered?

The internet exposes a great deal of naturally shared and linked data by private sector firms. Efforts to promote Linked Open Data (LOD) repositories have generally found better traction wherever the data to be linked can be identified as having clear referential properties. In simple terms, two data items that are said to be linked are so because they share something in common. They are data that are “about” the same thing. The sharper the sense of “identity” to the data items being linked, the easier is such data linkage. This means that linkage is most effective, in practise, for data concerning places, people, products and organizations such as nations, companies and institutions.

Australian initiatives to release bulk registry data such as the Australian Business Names (ABN) file, the Administrative Boundaries file, and the G-NAF address file greatly assist with record linkage. For privacy reasons, it would *not* seem appropriate to bulk release the Births, Deaths and Marriages file.

[93]See the recent Ponemon Institute report “Data Risk in the Third-Party Ecosystem”: http://bit.ly/1SNgaHJ

The rapid growth of social media, plus the preponderance of internet device identifiers, has had the effect of creating *de-facto* private identity tokens for most individuals having mobile phones and an internet presence. Sharing of this data is already widely employed through the internet advertising industry. It is unclear to us that there are major impediments in place to the linking and sharing of data about citizens. If anything, the issue is how to reconcile current activity with privacy principles.

Where there are options to share and link data which are not being commercially exploited already it is likely due to an absence of knowledge and skills in the business community about how to do it.

Who should have the ownership rights to data that is generated by individuals but collected by businesses? For which data does unclear ownership inhibit its availability and use?

This question is fraught since the ownership of data is typically defended through copyright and terms of use agreements. The essential difficulty is that copyright theory generally depends on the idea of *creative authorship*. Machine generated data is hardly a creative work. Therefore, ownership of the physical recording device is unlikely to confer ownership of the data which comes from it. The unclear status of data ownership is apparent in markets for financial data where it is common to sell "subscription services" and to restrict the rights of re-distribution for data through user agreements.

Public education is likely necessary about the terms-of-use contracts employed for data-access and digital rights management services. The legal question of who really has, or should have, ownership of sensor data seems very unclear. In our view, ownership of "data" is unclear since it comprises a list of literal values, or facts. If these facts have no discernible "creative author" they do not seem properly to be owned by anybody, as facts. Nonetheless, a private business may have spent a great deal of money gathering data. While they may have unclear rights of ownership, they certainly have a right to protect privately gathered data as a trade secret, via encryption. Since data encryption at the source is always an option, unclear ownership does not seem to be an impediment to business.

4.3 Consumer Access

QUESTIONS ON CONSUMER ACCESS TO, AND CONTROL OVER DATA

What impediments currently restrict consumers' access to and use of public and private sector data about themselves? Is there scope to streamline individuals' access to such data and, if there is, how should this be achieved?

The great bulk of data of interest to consumers about themselves is their *transactional data*. In our view, it seems sound to suppose many consumers could make more informed economic decisions if they had better access to the transactional data they generate on a daily basis. This proposition does not seem controversial to us from any productivity, efficiency or public-interest standpoint.

Needless to say, the organizations who are involved with consumer transactions already report to the consumer on each of their financial transactions, and in many cases their phone calls and other records of utility consumption. There is very significant public and private value in such data sets.

Recognizing that the consumer is properly the "author" of their own voluntarily executed economic transactions we think it feasible one might mount the case that a consumer owns their transactional data. Furthermore, the consumer's clear knowledge of their own private transactions surely does not detract value from the counterparty to those transactions, or their financial intermediary.

Here is a simple thought experiment… A consumer equipped with an infinite photographic memory would certainly be able to make more sense of their decisions and expenditure patterns. With digital technology, such data is available, but not generally in the hands of the consumer. Why not?

Clearly the present payments system already provides the means to collect rich data on transactions but this is not generally shared or pooled. This seems like a missed economic opportunity.

Pooled transactional data in a central government repository would provide statistical agencies and regulators with real-time data of high quality and relevance to economic management. Similarly, the use of standardized APIs for consumer access and download of transaction data would facilitate the development of entire new categories of consumer advisory services.

We can see few arguments against such a proposal, particularly in an era where many consumers are already conducting much of their personal financial management online. The ideal path forward would be an industry standard API to expose rich transactional data for use in service development.

Are regulatory solutions of value in giving consumers more access to and control over their own data?

Previous attempts to introduce APIs such as the Open Financial Exchange (OFX) standard failed due to active resistance from industry incumbents. In our view, such behaviour is short-sighted since the incumbents are well-placed to develop their own advisory services, and have been for some time. In view of this behaviour, it seems likely that a regulatory enforced solution will prove necessary.

Are there other ways to encourage greater cultural acceptance amongst businesses of consumer access to data about them?

If the problem is cultural then regulation seems the best path forward. Where transactional transparency through voluntarily shared consumer data enhances economic outcomes, transactional intermediaries are likely to *actually benefit* from the changes through increased economic activity.

What role do third party intermediaries currently play in assisting consumers to access and use data about themselves? What barriers impede the availability (and take-up) of services offered by third party intermediaries?

The development of the Comprehensive Credit Reporting framework has seen consumers able to access their own credit ratings and scores. Typically there are free offers to share scores on a once per year basis supported by a fee-based service where more active credit monitoring is needed. This seems like a model innovation of the kind this question invites. Limited access on a free basis serves to help educate customers while also protecting those who have limited means. Charging for more frequent access recognizes the value of the data gathering effort but also the merit of consumers taking an active and participatory role in managing their own credit scores.

In our view, this principle has merit and can be observed in so-called “freemium” online data service models. At one level of detail or intensity access to data is made free. At a greater level of detail or service intensity there is a fee for service. This seems like a rational market-based mechanism for the promotion of consumer data access and transparency while also enabling growth of intermediaries.

What datasets, including datasets of aggregated data on consumer outcomes at the product provider level, would provide high value to consumers in helping them make informed decisions? What criteria should be used to identify such datasets? What, if any, barriers are impeding consumers’ access to, and use of, such data?

Transactional data on consumer purchases, behaviours risk and preference profiles are likely to be of high value even when aggregated. The central question is which data items profile and segment any given consumer base in ways that are predictive of their purchasing behaviour. Informed consumer choice on the basis of data is likely a new business model for service providers who rate products and services. There are not any evident barriers to sharing aggregate data other than competitive strategy for those who possess it, or wish to limit access. Comparison shopping engines are a good example of a service that aggregates, pools and ranks pricing on consumer product. The barrier to the success of such businesses is the desire of product and service companies to *avoid* comparison.

QUESTIONS ON RESOURCE COSTS OF ACCESS

How should the costs associated with making more public sector data widely available be funded?

The data.gov.au platform employs a technology base called the Comprehensive Knowledge Archive Network (CKAN) which has an open source basis and is highly scalable. Furthermore, this system is capable of federating data exchange with other similar platforms as used by State Government. That provides interoperability for free.

The principal element of cost is likely the reformatting or connection of existing government data repositories to expose their data within the open government portal. There is a personnel cost that is associated with that and also the time and overhead of data conversion or transformation. Longer term, such efforts are likely to lead to efficiency dividends. One approach is to prioritise the release of those items of highest value with a commensurately high efficiency dividend. The large statistical agencies such as the Australian Bureau of Statistics (ABS) and the Australian Taxation Office (ATO) are likely those parties best placed to evaluate usage of web-scale data platforms such as CKAN as a replacement internal solution where this makes operational sense.

From a private sector standpoint, there is much to be gained by developing data transformation tool chains for the purpose of lowering data acquisition and storage costs within government. One way to accelerate innovation in this area might be for selected government departments to run hacking competitions for the purpose of parsing and transforming hard to access data such as PDFs though use of modern image processing and machine learning techniques. The use of a competition setting provides a beneficial exposure for innovative firms who may have new methods to pioneer, and can also save government resources on framing research and development tasks in data mining.

To what extent are data-related resources in agencies being directed towards dealing with data management and access issues versus data analysis and use?

One rule of thumb in data analytics in business is that 80% of time and effort is typically spent in collecting and marshalling data for later analysis. This rule is most likely true of government also. The most effective way to mitigate against high data preparation and governance costs is through rigorous application of data checking and data cleaning at the *point of collection*. For this reason, it is highly desirable that important “facts” such as names, addresses, geographical locations, registration numbers and other “markets of truth and identity” are readily available for data validation at those points where data is first ingested. In practical terms, the largest area of cost associated with data for government is likely that of *forms use* in collecting data from citizens without checking at source. When considered as an element of “data management” cost, it would appear sensible to measure, or estimate, the cost of forms and data-keying, data correction and false data when framing costs.

What pricing principles should be applied to different datasets? What role should price signals play in the provision of public sector data?

The marginal cost of disseminating data in digital form is extremely low. When access prices are lowered the demand increases by a larger amount. Therefore, the marginal cost of *data acquisition* is likely the highest cost and this is defrayed over many subsequent uses. The maximisation of re-use reduces the effective amortised cost of the data per data service request. The pricing structure must therefore consider the primary cost of data acquisition against the likely benefit of making that data available. This stance pushes policy development towards *metrics of efficiency in data acquisition*.

In addition, it may be possible to measure the elasticity of demand directly where online distribution channels are used. The cost of data acquisition is very important, but the monitoring of this can be used to raise efficiency of collection. For example, a current ASIC Company Extract costs $9 AUD for online delivery. It seems doubtful that it costs ASIC anything near that amount to collect that information from registered companies. In effect, the price signal deters use of the extract, except in those circumstances where access to it might be a legal necessity. For the Productivity Commission, it would seem that measuring the full life-cycle costing of data acquisition methods is the clear place to lean when seeking a virtuous feedback loop of cost reduction, efficiency gains and innovation. The cost of paper forms and the re-keying of data from PDF forms is an obvious area for improvement.

Is there a role for government in improving the skills base in this area?

The active engagement of government data custodians with private sector firms and academics who are researching data cleaning, data transformation, record linkage and machine learning at scale, might help to promote skills development and cross-fertilization between centres of excellence.

QUESTIONS ON PRIVACY PROTECTION

What types of data and data applications (public sector and private sector) pose the greatest concerns for privacy protection?

Data that relates to personal identity and health or finances are the most critical. Since digital service channels depend completely on digital authentication of identity, this is the weakest link. Flaws in the design, or practise, of digital security, access and authentication pose the greatest concerns for the protection of digital privacy since they are the primary vector for breaches or cyber-crime.

How can individuals' and businesses' confidence and trust in the way data is used be maintained and enhanced?

Public trust and confidence in the way data is used and maintained is determined by one factor: the performance of the data custodian in ensuring it is used and maintained well. In matters of public confidence, there is only the *actual performance* against community standards and contract.

The corollary to that assertion is that sound performance that builds trust and confidence is assured when there is close attention to the data governance: quality; security; access; usage and privacy.

What weight should be given to privacy protection relative to the benefits of greater data availability and use, particularly given the rate of change in the capabilities of technology?

The public risk of data becoming widely known is partially bound up with the value attached to facts. There are some areas of risk which are frankly created by poor security architecture.

Facts of an inalienable personal nature, such as: birthplace; date of birth and eye-colour, have been embedded into many digital security protocols. This seems like bad policy and bad practise since there is no reason why such obvious items should not be widely known or easy to discover.

Development of robust digital identity solutions that do not depend on inalienable personal facts are the likely solution to this problem. The concept of trusted parties to store such private personal facts has inherent weaknesses given that, once widely known, the personal inalienable facts cannot readily be unknown. When securing financial systems, for single transactions, it would seem that a better way is to create tokenized one-time use identity keys based on more secure biometric data. There are always weaknesses in any security system, but the widespread use of birthdays and other easy to guess facts like favourite colours, model of first car, and so on, seems weak, at best.

Strong passwords seem like a good idea, in theory, but countless studies have shown that human beings are careless with passwords and are prone to foolish choices or insecure storage habits. The key point to remember is that we rarely have cause to "remember" who we are – except when faced with any digital access or authentication system. This seems like poor security architecture.

Are further changes to the privacy-related policy framework needed? What are these specific changes and how would they improve outcomes? Have such approaches been tried in other jurisdictions?

The New Zealand Integrated Data Infrastructure[94] provides a positive example of innovation in public data linkage that is sensitive to community privacy concerns. The system integrates unit-record level data with attention to privacy protection for the purpose of illuminating public policy questions.

How could coordination across the different jurisdictions in regard to privacy protection and legislation be improved?

The principles and legislation in this area seem widely disseminated at both state and federal levels. The areas for improvement are more likely in the practise of data release. There are two opposing angles to this practise. On the one hand, privacy principles and legislation is sometimes invoked to sequester data against public release. This may be unwarranted. However, from another perspective some forms of data release might permit apparently anonymized information to be de-anonymized.

Of these two challenges, the first appears to be fading from view as the default policy is increasingly to make government data *open by default*. This shifts emphasis to the second area, which is really a question of skills, standards, and education about what forms of identity proxy may prove harmful.

The tracking of an unidentified individual, as representative of a cohort, to gather health, behaviour, or consumption preferences has undeniable commercial value. However, if this happens where there is an obvious proxy data item for the true personal identity then it may be a non-consenting use of the data in question. There are two obvious standard examples. Home internet connections are often assigned with a static Internet Protocol (IP) address with close resolution to a street address. That means the personal identity of home web-browsing activity is often discoverable. The same is true of mobile phones, especially in the hands of a telco, but through wireless internet. This means that some combinations of data-set are richer when joined than in the hands of just a single custodian. Understanding this issue, in the context of compliance with privacy principles, is likely the key problem in ensuring that legally compliant but trust-eroding practises are not commonplace.

[94] http://www.stats.govt.nz/browse_for_stats/snapshots-of-nz/integrated-data-infrastructure.aspx

How effective are existing approaches to confidentialisation and data security in facilitating data sharing while protecting privacy?

There is guidance from the Australian Bureau of Statistics (ABS) on principles for confidentialisation of data prior to release[95], primarily with regard to aggregates, but also for micro-data[96], which is data at the single-record level. The practise is to release Confidentialised Unit Record Files (CURFs), which are processed to make re-identification extremely unlikely. The advent of internet data gathering has changed the nature of the problem in degree but not in kind. The primary risk with internet data capture is the existence of proxy identifiers for individuals, such as static IP addresses, mobile phone numbers, and physical device identifiers like MAC addresses[97], phone IDs and IMEI numbers[98]. The major change with new technology is that communication networks with identifiable physical devices proxy for the location and activity of individuals. Stitching such data together can enable much richer visibility onto individual personal behaviour. Data at this level of granularity needs to be managed very carefully, in our view, since it has clear privacy and security implications. One way to address this issue is to employ "consumer persona profiles" which contain a demographic snapshot that is informative, but not too identifying of the real individual. This might be an age range, gender, postcode, income range, broad occupational category or credit score, for example. The principles used by the Census Bureau to de-identify survey data can be updated to reflect contemporary digital practise, and the types of information that can be reasonably shared to preserve privacy.

From another perspective, there is the question of best practise for the storage of highly granular data that has privacy and security implications. The standard remedies are secure access control and encryption of data at rest and in motion between data centres. There are record-level access controls employed in commercial databases, such as Sqrrl[99], which are used for record-linkage. This is based on the National Security Agency (NSA) database Accumulo[100], which limits data access by the identity and role of the interlocutor at the cell or record level of the data store.

What lessons from overseas jurisdictions can Australia learn from regarding the use of individuals' and businesses' data, particularly in regard to protecting privacy and commercially sensitive or commercially valuable information?

The commercial practise of major cloud-computing players such as Amazon Web Services (AWS) and Microsoft (Azure) is informative. The issues involved in maintaining a public cloud infrastructure are

[95] http://www.nss.gov.au/nss/home.nsf/pages/Confidentiality+-+How+to+confidentialise+data:+the+basic+principles

[96] http://www.nss.gov.au/nss/home.nsf/pages/Confidentiality+-+Managing+the+risk+of+disclosure+in+the+release+of+microdata

[97] This is the unique number on a network card which is generally only visible in a local network.

[98] https://www.telstra.com.au/coverage-networks/check-imei To obtain the IMEI on a Telstra mobile phone enter *#06#. This number is a unique identifier for the physical phone which is sometimes accessible to developer applications. Device numbers can proxy for identity in the same way that a mobile phone number proxies identity. Through linking the IMEI to the mobile phone number to the IP address of a fixed location, such as the home address of the person, it is possible to track individuals as they move through different web-browsing and calling activities. The level of detail depends on how a person chooses to camouflage their activity through use of virtual private networks. It also depends on the position, within the system, of the data gathering agent. Compromised network architecture, such as a router, can enable siphoning of such data or more elaborate Man-in-the-Middle (MITM) attacks which attempt to patch secure communications

[99] https://sqrrl.com/

[100] http://www.pcworld.com/article/2060060/nsas-accumulo-nosql-store-offers-rolebased-data-access.html

naturally the same as those faced by public and private institutions that must share data across organisational boundaries[101]. Arguably, the considerations of a public cloud provider require a higher level of attention to the data privacy and security architecture. So-called multi-tenant clouds have a range of customers co-existing in the one physical infrastructure. Strong encryption is used for data at rest and in motion within the cloud. This is supplemented by Public Key Infrastructure (PKI)[102] and Digital Rights Management (DRM)[103] to ensure identity and access protocols are respected.

In this regard, the transnational public cloud infrastructure of major trusted cloud providers surfaces the key issues and guides best practise development for data storage and sharing protocols. There is no reason, in our view, to consider effective policy and standards to have a unique national basis. It would therefore seem reasonable to develop public policy in parallel with the development of cloud security, trust and regulatory compliance in the private sector. The problems are identical.

What are the benefits and costs of allowing an individual to request deletion of personal information about themselves? In what circumstances and for what types of information should this apply?

The costs are clearer to state. Digital data once stored and made searchable, can spread with no real practical limit of reach or access. Therefore, erroneous digital data, about an individual, can be very damaging, when not corrected. There are obviously costs to the redaction or correction of a false data item. However, architecting systems for this purpose is simply good practise wherever the data is perceived to have value. Evidently, not all data is of equal value when judging the cost of data errors. However, it seems that errors relating to *digital identity*, *credit records* and *health records* are of particular importance. For this reason, some types of data might have either gold records, like a government registry, or an addressable channel to establish truth. Since new data systems are more than likely distributed rather than centralized, distributed shared-write systems, such as blockchain, may provide the technical basis for establishing truth, redacting truth, or correcting error. Problems of this kind do have technical solutions, but these exist alongside important social, legal, political and economic questions. The appropriate place to establish sound policy by example is likely through the existing channels the state has to maintain "gospel truth" data — such as birth records.

What competing interests (such as the public interest) or practical requirements would indicate that the ability to request deletion should not apply?

There is likely a policy principle lurking here in the concept of *alienable* versus *inalienable* data. The private biometric signature of a person (their DNA, iris pattern and such like) is inalienable to them as a matter of fact. However, the data record of association between them and an iris pattern can be alienated from truth, which is to say, it could easily be erroneous. For instance, a successful hack on a digital identity database might replace the real signature of a real person with the real signature of a different actual physical person. That would constitute a corruption of data, which by substitution amounted to an effective theft of identity. Examples of this are already occurring in the credit industry via "card not present" transactions. Lasting damage to a credit record is possible when such errors are not detected, not remediated, or otherwise allowed to spread in the wild as "bad data".

[101] https://www.microsoft.com/en-us/trustcenter/Compliance
[102] https://www.humanservices.gov.au/health-professionals/services/medicare/public-key-infrastructure
[103] This can protect access to a document, thereby making it a trade secret. There are natural policy issues that arise when the courts seek access to such protected documents. Since strong encryption is available to all, it is difficult to compel a party to divulge the encryption key. A standard response would be: "Sorry, I lost it."

In such cases, where the fact of the data (identity) is inalienable but the data has been corrupted, the right to correction would seem particularly strong and uncontroversial (however inconvenient).

However, there are other circumstances which are more concerned with opinion and comment. Legal systems already have well-developed theories of slander, libel and misrepresentation. There are grounds to approach the development of digital standards in this well-established spirit.

For example, it may be difficult in practise to "correct" all false or misleading statements on the web. However, where cause is mounted, under existing laws, it would seem reasonable, and less onerous, to reflect any "public retraction" of such statements a suitable rank in any search engine. There will be arguments for and against, running to cost versus benefit, but it does seem, to us, that methods such as the publication of public notices encounter no real obstacle to wide dissemination. The news media regularly publish such correction notices. With suitable metadata the publication of corrected data by a "trusted" or "authoritative" source can be made an effective communication method.

One simple example of this in practise is digitally signed personal metadata. If individuals wish some facts to be known about them, and be clearly authored by them, a metadata standard to package the data in digitally signed form is one way to achieve that. Using the resulting personal *micro metadata standard* would enable any number of authorities to assert truth in connection with such data. The role of courts would then be via a Truth Resolution Protocol (TRP), to determine what data is actually correct about an individual. Since data is just 1s and 0s, we cannot really place faith in data to be correct when it purports to be the gospel truth about an individual.

QUESTIONS ON OTHER RESTRICTIONS

Having regard to current legislation and practice, are further protocols or other measures required to facilitate the disclosure and use of data about individuals while protecting privacy interests?

The Office of the Australian Information Commissioner (OAIC) has set forth privacy principles. The use of data about individuals is most fraught where there is no clear understanding of *what* data has been gathered and *how* it is being used. The issue is really the status of *informed consent*.

What form should any such protocols or other measures take?

Statements of privacy principles might be extended to include recognition of some standards and protocols concerning the informed consent for data collection about individuals. The use of website cookies and statements about the use of cookies are examples of such protocols in practise.

New data collection channels are likely to proliferate: in device cameras; wearable devices; monitoring of communications and such like, as the Internet of Things and connected devices expand. In our view, the principle issue is the existence or otherwise of informed consent. There is a need to foster industry best-practise disclosure so as to educate consumers and mitigate the risk of failing pubic trust in custodians of personal data. The emergence of a community standard, backed up by legislative standards seems likely as data collection touches more aspects of daily life. The key problem to be avoided is that encountered by the media industry with invasions of privacy[104].

Is there need for a more uniform treatment of commercial-in-confidence data held by the Australian Government and state and territory governments?

[104] The legislative tension is between the public "right to know" and the personal "right to privacy". These two are intertwined with the search engine related "right to be forgotten" and the social "right to be forgiven".

The Australian Bureau of Statistics (ABS) has developed a guide for data integration projects involving Commonwealth data for statistical and research purposes[105]. This involves seven high-level principles[106]. In paraphrase, these principles are:

1. Data is a strategic resource
2. Data custodians are responsible and accountable for security and confidentiality
3. Data integrators are responsible to manage the project in agreement with custodians
4. Integration should only occur where it provides significant overall benefit to the public
5. Statistical data integration must be used for statistical and research purposes only
6. Policies and procedures should preserve privacy and confidentiality
7. Data integration should occur in an open transparent and accountable fashion.

There are existing programs whose aim appears to be standardization and to promote best practise at Commonwealth, state and territory levels. Agencies such as APRA run programs that collect sensitive commercial data with established protocols to publicly consult with stakeholders prior to the introduction of any planned statistical release. There appear to be well-established protocols and procedures, but there may be a benefit to identify where there are government centres of excellence and facilitate knowledge sharing.

Are there merits in codifying the treatment and classification of business data for privacy or security purposes? What would this mean in practice?

The process of aggregating data benefits from codified and standardised treatments. It also requires protocols for the release of data to ensure that the size of the dataset or other attributes do not allow it to be re-identified where there are privacy or security implications. The standard ABS style categories may need to be adapted where the nature of the data-gathering process promotes a novel approach to classification. For example, certain industry classifiers that are commonly used in the financial markets are different from those used in regular economic and trade surveys. Creating conformance tables has been a common solution to this problem over time. Rarely do parties agree on the best aggregation category to use for data, but some standard categories do help. The merit of flexibility in this area is that it may reduce compliance costs, set-up costs, or delivery costs of data in cases where several possible categories exist but only a few choices are readily useable or deliverable. Whatever classifications are chosen, it seems important that they are open data licensed classes. The use of proprietary classifications has the undesirable effect of limiting uptake.

For instance, the release of the Administrative Boundaries data provides a convenient public domain de-facto standard against which to aggregate and report public and private sector spatial statistics.

QUESTIONS ON DATA SECURITY

Are security measures for public sector data too prescriptive? Do they need to be more flexible to adapt to changing circumstances and technologies?

Data security is a common problem for both the private and the public sector. The private sector has clear concerns for loss of intellectual property and commercial value alongside reputational damage due to privacy and data breaches for clients. Technologies developed in this area, such as multifactor authentication, data encryption at rest and in transit, as well as public-key infrastructure based upon

[105] https://statistical-data-integration.govspace.gov.au/
[106] http://bit.ly/1TeZLcC

web standards seem adequate to both public and private sectors. The area where government seems to lag is an acute focus on *data security policy* versus *implementation effectiveness*. The most important element to modern cyber-security is *defence-in-depth*, meaning attention to securing all aspects of the design, implementation and operation of secure systems. Prescriptive policy with weak attention to implementation is a waste of effort and leads to a false sense of security. The best practise approach combines flexible teams who are charged with probing and testing systems rather than asserting standards of compliance. Cyber-criminals, state-actors and other threat sources look for weaknesses in the overall security posture via a combination of behavioural, hardware, software and procedural gaps. In such an environment, a slow-moving bureaucracy is likely to be penetrated without realising this has happened due to over-reliance upon "policy as defence". The best analogy for this is the Maginot Line concept of an "impregnable perimeter barrier". Determined hackers simply go around such systems by exploiting other weaknesses in the attack surface. History shows one of the more reliable attack modalities is to exploit the "human element": over-reliance on single responsible persons; faith in engineered systems; run-of-the mill carelessness and other foibles. In our view, effective cyber-security requires agile, well-resourced and skilled teams with the mindset to actively break systems. The best cyber-defence is to cultivate the mindset and skills of an attacker.

How do data security measures interact with the Privacy Act?

The security of digital identity is the key interaction point with the Privacy Act. There are emerging new hazards in the development of "data as a proxy for physical identity". From a human behaviour standpoint, the concept of digital identity carries great benefits of convenience (such as tap and go payments), but also risks of a new class of lucrative crime: *digital identity theft*. Digital theft and fraud are as simple as stealing a small sample of data that proxies for identity (passwords, personal information; digitised biometrics etc). Whereas the digital identity appears more secure superficially, the effective compromise of digital identity is potentially more damaging since it carries intrinsic high levels of trust. The introduction of biometric identity does not ensure absolute security since a system of this type can be compromised through classical *man-in-the-middle* attacks that intercept data through an untrustworthy computer masquerading as a trustworthy one. These attacks are the equivalent of cyber-phising to fool a computer into thinking another computer can be trusted.

Provisions of the Privacy Act might establish a clearer recognition of identity theft, cyber-phising, digital fraud, the falsification of records and impersonation as serious crimes. There likely needs to be more attention to effective law enforcement for cyber-crime as the focus shifts from encouraging compliance with privacy principles to prosecuting criminal activity. Today it is easier to rob a bank digitally than it is to crack a bank vault or hold up an armoured car. There is a poor understanding, in our view, that the rigour of digital identity as a proof of identity does not make the theft of identity (which is simply a parcel of data) less feasible or likely. Arguably, the digital world greatly simplifies crime as entry to the target premises need not require any physical presence at the scene[107].

Enforcement that is serious in intent to detect, pursue and prosecute cyber-criminals shifts the focus from law-making to cyber-security teams in government, the private sector and law enforcement. In our view, any nation state that wishes to have a thriving digital commerce sector and reap the efficiency benefits of digital government must have a mature digital security sector. With that goes

[107] Sometimes a physical presence, in the form of a compromised computer, is the key to penetrating a facility. There are many methods to do that, but the use of removable storage media to spread malware is popular.

the legislative requirement for balance between personal privacy and the prevention of overreach by agents of government or carelessness on the part of commercial interests.

How should the risks and consequences of public sector and private sector data breaches be assessed and managed? Is data breach notification an appropriate and sufficient response?

In our view, the key public benefit of data breach notification is to share intelligence concerning weaknesses and emerging threats. Notification of a data breach can be very important for those whose data has been compromised. Credit card data provide a simple example. The timely notice of a breach can alert those whose cards were compromised to contain the potential financial impact. The damage can extend beyond monetary loss to personal credit ratings. Without effective breach notices, the consumer may not realise that their "personal identity" has been sold and is circulating in the criminal market for use by digital fraudsters. If not detected early, this can lead to creation of a fraudulent account trail which is extremely difficult to unravel long after the fact of the breach.

Developing a sufficient and effective response to data breaches should follow established principles of accident investigation. Simple breaches involving commonplace flaws should be avoidable. Reporting of these, along with an effective analysis of why they happened is simply good policy. For small and medium enterprises, which do not have a large internal team, an effective response would likely be to note that the data breach belonged to a common category with a mode of failure.

The value of notification is the ability to detect patterns that signal a need to improve security, operating practise or training. The software industry has the *Common Vulnerabilities and Exposures* register of common cyber-threats along with the US *National Vulnerability Database*. In Australia, there is the National Computer Emergency Response Team (CERT). A connected strategic innovation policy would see the legislation of the Privacy Act as statements of *intent and principle* backed by the common practise of sharing information on the proximate cause of any "digital accident". When viewed this way, the framing of legislation must ideally avoid the trap of discouraging the reporting of breaches, due to the fear of overly punitive sanctions.

The key legislative balance is likely a commitment to enforce laws in prosecuting cyber-criminals, combined with a coherent policy to mitigate and reduce "digital accidents". Information about how accidents happen is useful to organizations, both public and private, that are not cyber-security experts. The field is complex. Very simple software configuration errors can easily lead to breaches where there was no criminal intent and any negligence is down to ignorance. This feature of digital crime suggests that the *perpetrators* really should be pursued vigorously. There is a role for the state in providing such community protection since state actors also feature prominently in the use of cyber-incursions for espionage and offensive military operations. Malware, network taps, viruses and Trojans are the future of espionage and criminal activity, with security implications.

5 Appendix A: Policy Tensions and Dichotomies

The policy questions posed by the *Data Availability and Use* inquiry can be answered in several ways. In this submission, we pursue both the obvious *direct response*, that offers answers to each of the specific questions raised, but we also provide a *contextual response*, which serves to capture the philosophical ground to our thinking, from which the specific answers are constructed.

It is a matter of taste as to how much weight the reader may give to these different approaches. Those who attach weight to our view out the window may read the context section. Those who simply want input on the specific questions may read the questions section. The sum of both perspectives amounts to a national opportunity, for which we provide recommendations.

5.1 Public vs. Private Goods

Public policy is invariably an exercise in balancing competing interests between the individual and the collective benefit. Wherever data is involved, the properties of data as a commodity tend to be mixed inextricably with the *rights to use*, specifically the rights to *deny access to use data*. Several recent contributions to the economics of public data have emphasized this point. We defer to the more detailed enquiries on the subject by Gruen at al, but call out a clear point of agreement, for those who have studied data economics in depth – data has clear public-good properties.

Data is *non-rivalrous* in that the possession of data by one party does not reduce availability to other parties, and is also *non-excludable*, in that once made available its use is not readily limited.

> *"If you have an apple and I have an apple and we exchange these apples then you and I will still each have one apple. But if you have an idea and I have an idea and we exchange these ideas, then each of us will have two ideas." — George Bernard Shaw*

Nonetheless, data has great value, in broad circumstances, for informing decisions and facilitating the discovery of relationships, patterns and probable future outcomes. Data is not simply a static commodity but, once in active use, can promote the generation of new knowledge and insight.

In this way, the question of what to do with public data in particular, and data licensing in general, has another dimension: namely that of the value of data-in-use for *creative discovery*. There are possible products and services derived from raw data that create significant additional value.

In this time, rapid innovation in digital technologies such as: cloud computing; big data analytics; and the growth of agile micro-services architecture, promote policies that work to take data out of public and private siloes and to put that data in motion to serve the growth of an information economy.

5.2 Open vs Closed Licensing

This starting point clearly promotes consideration of *data licensing*. Since data is non-rivalrous, but has private decision making benefit, there is clear value in securing private information. Society has always maintained a public-private benefit tension in the management of *creative product*. It must also be remembered that data about entities, and most especially natural persons, has another dimension of value in connection with *personal privacy* and *security of identity*.

There is not only *positive value* in the release of data, but also clear *negative value* where there is a breach of commercial confidentiality, personal privacy or the security of digital identity.

It follows, that discussions of data access and policy must necessarily involve a balance between the poles of positive value creation and negative value destruction. This is not new, and there is a rich body of law regulating such activities. In our view, the clearest construct for policymakers is the concept of a *licensed use* (sanctioned as acceptable) versus an *unlicensed use* of data.

Due to the properties of data being non-excludable by nature, once released[108], the only conceivable check on unfair or inappropriate use is really the legal sanction of punishment and remedies. There are many ways to pursue such policy, but we believe the preferable and clear alternative is the use of clear and unambiguous *licensed terms of use and ownership* on an up-front basis. Such clarity of licensing can be directly approached by any party to use of the data without fear of third-party actions. This is in the tradition of simplifying contract as an agreement between two parties.

Needless to say, there will also be *data crimes*, meaning the acquisition and use of data outside of the terms of any contract, or for purposes that the state purposefully excludes. The theft of private data for impersonating identity and making fraudulent transactions is one such example. With this understood, the policy discussion can be framed around *terms of licensing* versus *prohibited use*.

The Open Data Institute has introduced a concept called "The Data Spectrum" which helps classify the access, assignments, sources and custodial roles for different data items spanning open public data to completely closed private data. This concept can be easily specialized to verticals having a clear need to protect and regulate the release of sensitive data such as the banking industry.

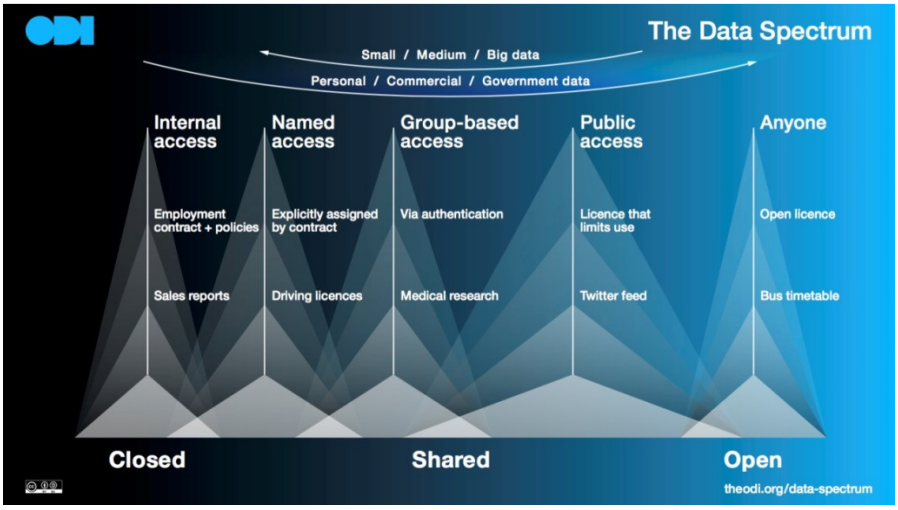

Figure 1 The Data Spectrum concept from Open Data Institute (Source: theodi.org)

The concepts of *Open* versus *Closed* data help organise the discussion of data access, terms of use, and data licensing. The Creative Commons licenses for differing degrees of openness are emerging as a de-facto global standard. In our view, the use of standardized and clear data licensing might be usefully extended to the private sector, as well as voluntary personal disclosure.

A common feature of digital markets is that *standards promote interoperability*, which grows total network value for community benefit. This principle has been demonstrated by public and private interests through the Word-Wide Web Consortium (W3C), to promote internet standards for the effective *technical interoperability* of internet-based data systems.

[108] It follows that licensing must be clear up-front. Once data is released, it is too late to argue terms of use.

However, the mere exchange of data is insufficient to promote aims of productivity, innovation and efficiency in the digital era. We need also clarity of the legal basis for use and re-use of data, along with the development of effective standards for acceptable versus unacceptable use. Since digital technology now enables relatively frictionless gathering and exchange of data, along with search against that data, the primary remaining constraint is the *legality or otherwise of data usage*.

In our view, the promotion of standards-body partnerships between the public and private sector might assist in developing industry standards for data licenses and disclosure in complex verticals like: telecommunications; healthcare; finance and education. Data licensing rests at the interface between technology, public policy, legal interpretation, and entrepreneurial innovation.

Australian Federal Government policy has stated that public data is a "strategic national resource" and further stipulates that non-sensitive public data be "open by default" to contribute to greater innovation and productivity. We support this policy in the interests of transparency and innovation. The nature of the federation places a potential additional burden of compliance with different state and federal licensing formats. In the interests of productivity, standardisation around Creative Commons licensing is a visible trend and one that might be usefully encouraged[109]. The Australian Governments Open Access and Licensing Framework (AusGOAL) was designed for this purpose[110].

5.3 Privacy vs Openness

According to the above, the central avenue open to framing policy from a legal perspective is the use of data licensing to make clear the acceptable terms of data release, use and re-use. In addition, we mentioned the issue of unacceptable usage that properly supervenes through rights and principles that attach to individual freedoms. These are covered through extensive privacy regulation at both state and federal levels along with the Office of the Australian Information Commissioner (OAIC).

In our view, the development of hyper-scale information gathering and distribution systems, such as modern social media networks, has not altered the principles at work in privacy. However, it does increase the *scale* of possible public detriment through security breaches and privacy invasion.

This is probably of greatest importance in connection with persons who are minors and thus may not be fully cognizant of the future consequences of present actions. It also applies, we think, in cases of possible abuse of private information of an ethnic, religious, sexual orientation or genetic nature.

Existing legal frameworks are probably adequate to deal with particular circumstances but, in view of the large scale of data gathering by private interests, there may need to be some attention given to remedial protocols to deal with damaging data at large. In the interests of productivity, and the principle of least interference, the encouragement of industry protocols to sequester or remove erroneous data records or to otherwise facilitate "forgetting history" may prove beneficial.

In purely legalistic terms, this debate represents the tensions between: the *personal right to privacy*; the *public right to know*; the *personal right to be forgotten*; and the *public right to forgiveness*. The principal areas where this attention seems justified are: protection of minors; the avoidance of any culture of public vilification; protection against irreparable reputational damage from identity theft; and variations on the old themes of protection from public harassment, slander, and libel.

[109] See UK standard license: http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/
[110] For the AusGOAL program aims see: http://www.ausgoal.gov.au/overview

The legal system has already many deliberative processes and remedies in place, but the hyper-scale nature of modern private data-gathering processes, and the potential for unintended breaches, has amplified the social risks in play. In our view, an appropriate policy response, from a productivity and efficiency stance, would be to encourage Public-Private Partnerships for workable standards and protocols for redaction, remediation and potential limitation of damage from data breaches.

This probably extends to allowing individuals who were compromised through chance events, folly of youth, or via cyber-crime, to receive some redress and remedy. There are existing principles and laws concerned with circumstances whereby personal data may be corrected to guide this.

5.4 Security vs Agility

Related to the question of data breaches, through misadventure, poor practise or cyber-crime, are the balance and trade-offs between the pace of innovation and the quality of data security. For digital business, and government, the opportunity set for technological innovation is vast. The public benefit aims of innovation and efficiency can be usefully balanced against security.

One contemporary example of this trade-off informing public policy is the creation, by the Australian Securities and Investment Commission (ASIC), of a “regulatory sandbox” to balance the agile testing of new digital financial advice products and business models against a prudential risk limitation on scale of activity, until proven. This approach seems very sound where data policy is involved.

This is one possible design pattern to experiment with in the development of protocols for sharing, linking and exchanging data across the boundaries of governments and private enterprise.

One excellent example of such policy in action is the use by the UK Government of private hackers to screen-scrape government websites for the purpose of re-factoring data in more accessible formats. Hackathons and challenge problems present an opportunity for governments and private firms to make a limited release of sample data for the purposes of discovery and engagement. Unlike the negative scenario of a privacy breach, the intentional release of anonymized data may promote public innovation in solving difficult problems of data access and transformation.

There is an old historical precedent for such activity in the release of US Census name-related data describing the frequency of personal names to facilitate research in record matching. Recognizing likely matches between individuals on the basis of names, abbreviations and nicknames occurs frequently in both commercial and government problems, particularly in health records.

Creating suitable “corpus data” which is anonymised, and thus privacy neutral, but rich in content for the development of better health outcomes, seems like a natural area for policy development. Security is met through anonymisation, while agility is fostered through quality publicly available research data-sets, in the spirit of census and other socioeconomic surveys.

5.5 Cost vs Benefit

The social dimension to cost associated with privacy and security has been mentioned. Where public data is involved, there is a clear cost-recovery mechanism for government via taxes on activity. This policy stance is becoming the norm worldwide, as acknowledged in the G8 Open Data Charter, the Open Data policies of the UK, the US and policy research by the Open Data Institute, the Omidyar Network, the World Bank and numerous other promoters of open public data.

For such a policy approach to be effective in practise, it is important that the data released into the public sphere is: of high intrinsic value; of a high quality; readily accessible and capable of being processed by both machines and humans for maximum value in use. It is also important that data sets be updated regularly, and provided with interoperable access methods. The use of industry standard Application Programming Interfaces (APIs), and data packaging formats such as XML, XBRL and JSON, helps support automated analytical processing at scale. This policy maximises the benefits of low marginal cost digital processing. When coupled with suitable open data licensing, public data can then be refined, combined, analysed and re-processed for improved economic efficiency.

The policies mentioned above facilitate value creation, both public and private, from data in motion and in active use to inform public and private decision-making across the entire economy. There are benefits to be captured from improved efficiency, higher activity and the mitigation of risks.

That is the theory behind the release of public data under open data policies. There is supporting evidence from a wide range of previous studies. A 2014 study by Lateral Economics, commissioned by the Omidyar Network[111], estimated the GDP benefits of open public data to be around $15B AUD per annum, or approximately 1% of GDP. This number is broadly consistent with a McKinsey global study, once scaled for the relative size of the Australian economy.

The key point to note is the *low cost to serve* data of a digital model. Typically, the serving of data via a public internet portal is an economic micro-transaction that costs only a tiny amount per interaction. We can estimate these costs through the example of paid for web-search.

Wherever the decision value of a data analytic served on a single micro-transaction basis rises above the threshold of (perhaps) $0.002 to $0.005[112] there is clear net economic benefit to the system at large. When public data is gathered for public purpose the cost of data gathering must be defrayed through tax revenues, or direct charging for use. Since the cost to serve data is very low, the principal cost is generally the data gathering process. Wherever the gathered data is of general utility, such as search data, the cost of gathering can be amortised over many queries. In this case, the limiting cost to serve is the power and energy requirements per query, in volume[113].

In our view, the primary opportunity of digital government is to reduce the costs of data gathering and digital service delivery. The benefits are likely to accrue through higher aggregate taxes, on a larger and more innovative digital economy. Reasoning in this fashion, we take the view that government most likely has greater present visibility on data gathering costs. Effective innovation policy might then engage public-private partnerships to reduce this element of cost.

---

[111] See "Open for Business How Open Data Can Help Achieve the G20 Growth Target" a Lateral Economics report commissioned by the Omidyar Network, by Gruen, N., Houghton, J. and Tooth, R. (June 2014). https://www.omidyar.com/sites/default/files/file_archive/insights/ON%20Report_061114_FNL.pdf

[112] This number is a rough estimate from the volume rate for search engine API use. For example, the Bing search engine charges $2,000 per month for up to 1,000,000 search transactions ($0.002 per transaction). http://datamarket.azure.com/dataset/bing/search. The Google custom search API charges $5 per 1000 queries over 100 per day, up to 10,000 per day, or $0.005 per transaction. https://developers.google.com/custom-search/json-api/v1/overview#pricing

[113] There are few "official numbers" available, but a 2009 Google blog post suggested that the energy cost of a typical query was around 1kJ or 0.0003 kWh. At (say) 8c per kWh that is 0.0024c. Clearly the pricing at work for search engines is the (higher) cost of powering human software engineers and the needs of the shareholders https://googleblog.blogspot.com.au/2009/01/powering-google-search.html. Nobody is suggesting that the pricing of data services should obey the Labour theory of value, but clearly the cost to serve is pretty low.

This suggested approach recognizes that government can most likely not drive any private sector innovation beyond the natural appetites of the for-profit motive. However, a clear policy aim to reduce the cost of data acquisition through adoption of digital technologies is synergistic.

The quid pro quo, which goes with such a policy stance, is that cost reduction is always easy when the decision is to *stop providing a service*. This would be inconsistent with the view that digital methods offer broad positive transformational impact across the economy.

To summarize this standpoint, the low marginal cost of data analytics delivery ought to promote attention on efficient and effective means of gathering high-value data at low aggregate cost. A concrete example of this might be environmental sensor data through networks of public and private sensors that were aggregated and shared on a suitable collaborative basis.

Numerous policy models are possible for exploration: *shared systems*; *systems for tender*; *private data* escrowed and then released; and other variations on the data gathering theme. The areas of policy experimentation with least privacy impact, but high-value in use are likely: geospatial data; minerals and environmental data; and transportation systems and telematics data. Pursuit of this policy theme is likely to yield more and higher accuracy data on the benefits of digitisation.

In some limited cases, there may be a firm public policy interest in compelling the release of private datasets. The Privacy Act (1988) has been amended to introduce Comprehensive Credit Reporting (CCR). The purpose of such changes was to improve the visibility of consumer credit histories to add *positive* aspects of credit history, rather than simply negative events such as missed loan payments. This data initiative ought to promote improved credit risk pricing and credit allocation. While the scheme is voluntary for now, credit data release by Australian banks has been slow.

The Murray Financial System inquiry stated (Recommendation 20): *If, over time, participation is inadequate, Government should consider legislating mandatory participation*. This recommendation may need to be enforced if the private sector interests prove uncooperative.

5.6 Distributed vs Centralised

The Internet naturally promotes distributed data storage and cross-linking. The natural problem that arises in distributed architectures is a lack of coordination and control for data governance and data provenance. These issues can be addressed by a combination of best practises, policy and standards that promote tested patterns for solving common problems.

In some cases, where security is especially important, it makes sense to centralise the management within a single body with appropriate controls and backups. In other cases, the nature of the data is such that it poses no great risks, in which case direct open release is a natural default.

For the development of skills to promote best practise it likely makes sense to centralise some of the utility functions related to advocacy of best practise solutions and skills dissemination. The latter strategy seems to be the preferred option of most governments[114].

In Australia, the development of the Gov 2.0 initiative and the Digital Transformation Office (DTO), along with the identification of the Whole of Government Centres of Excellence (WGCoE), is consistent with best-practise overseas[115].

[114] See the UK services guide: https://www.gov.uk/service-manual

6 Appendix B: The Interaction of New Technology and Policy

New technology is rapidly re-shaping the data architecture of most organisations. Some familiarity with these trends may prove helpful in a policy context. The principal areas where technology has enabled qualitatively different types of organizational behaviour, and thus risk, have to do with the Volume, Variety, Velocity and Veracity (4Vs) of data when used to inform decisions. In the following we summarise the key elements that seem important: the capability to build integrated knowledge by combining different sources; the driving capability for data linkage to support this activity; the risks posed to privacy and identity disclosure; and the hazards of poor data governance practices.

6.1 Knowledge Graphs

A Knowledge Graph is a database whose structure reflects the "Web of Relationships" natural to the internet and the more general problem of describing and storing related facts. Simple examples can be found with basic facts about an entity: a person; place; or thing. For instance, a parent company may have many subsidiary entities, each of which may have its own senior executives, registered offices, business establishments, employees, products, trademarks, and so forth.

In any realistic scenario, the complexity of the total picture across organizational boundaries creates problems in maintaining any system-wide "schema" or pattern for data-storage. The traditional form of database uses a table or file metaphor for storing facts, wherein the important relationships are assumed to be knowable in advance. This is generally not the case for any real-world system.

The idea of a *graph database* is to store information in the form of simple relationships between two entities, without any upfront restriction on the type of relation to be stored. This makes it easier to combine disparate databases which may store different types of relationships. So long as there is high enough quality on the "identifying" characteristics of the entities stored in the database, linking of data to discover new facts across databases is made much easier.

Web search and social network firms such as Google, Facebook and Microsoft have begun using such technology to discover or mine information from the public internet. This can be an effective means for refining data from disparate sources to produce a higher-level mosaic of combined data. There are obvious productivity and efficiency benefits from more accurate and better linked databases.

However, there are also important privacy concerns about the linkage of data relating to individuals. Arguably, the advent of social networks such as Facebook, Twitter and LinkedIn has rendered discussions of public acceptance something of a moot point. Clearly, the public does broadly accept the idea of private organizations holding significant quantities of personal data, possibly more so than government organizations. However, it seems likely that case law and policy will develop side by side in dealing with "edge" cases. These are circumstances where there is clearly a tension between the stated purpose for which data has been gathered and the actual purpose to which it has been put, possibly by third parties. The legal concept in question is whether the customer agreement to the stated *Terms and Conditions* of data gathering constitutes *informed consent* to all purposes to which the data has been put. Due to rapid developments in data science, probable inference makes possible the discovery of *probable* associations and *likely* relationships. While such inferences may be useful for decision making, they are not, strictly speaking, facts.

[115] See the US agency 18F: https://18f.gsa.gov/

In our view, data which drives decision processes based on statistical likelihood carries with it the implicit propensity for decision outcomes due to statistical error. The law recognizes legal doubt and the rights of commercial entities to manage risk, but where public perception has caused an *incorrect likelihood* to be viewed as *factual*, then remediation in the public record seems reasonable and prudent. Governments are familiar with this problem, but the world of private business may not yet be fully prepared for the variety and volume of such corrective actions. Where erroneous data, or flawed processes of inference, lead to actual lost economic opportunity it seems likely that new policies and remediation protocols and processes will prove necessary[116].

6.2 Identity Protection and Anonymisation

There are many benefits to the use of knowledge graph technology in organizing the vast amount of data now published to the internet. However, there are also important policy implications regarding those circumstances where unrelated facts might be mistakenly linked to persons or companies. The best way to mitigate against such unforseen hazards is through identity protection *at source* via a range of anonymisation, aggregation and tokenization methods. In Australia, the official policy on government data linkage sets out protocols, and a range of suggested solutions to such problems[117].

However, one can foresee a range of edge cases, such as "identity theft", leading to the propagation of erroneous facts about an individual or a corporation. Disentangling such circumstances may well become more difficult with distributed linked databases. When there is no gold copy of "truth" it is hard to know where to start in reversing bad data introduced through fraud, crime or error. There will be technical remedies for such problems, but policy needs to be supportive of the remedy. This aspect of the problem is most evident in the case of knowledge bases alluded to earlier. Credit data, insurance claims data, health data, education and employment data are all examples where a fake or stolen identity may wreak havoc with the economic circumstances of an individual.

The real "problem" of such circumstances is likely to be the creation of robust unique identifying facts for an entity. In this regard, registry data and the individual non-forgeable biometric data are clearly of high importance. However, one should note that a data record of biometric facts can still be wrong. Identity theft in a digital identity world may be as simple as overwriting a biometric record using a false data record describing the *true* biometric facts about a *different* individual.

Personal *data* is disembodied from the individual that it may purport to be about. Once economic value is attached to the data, then there are economic incentives for misuse of that data. Among the largest of these perverse incentives are theft and fraud related to payment systems. Whereas it has been customary to use personally known facts (such as birthdays and favourite colours) as tokens of identity, this practise is now clearly failing as a means to guarantee positive proof of identity. In the example given, even the *possession* of biometric data is no positive proof of identity if the system of gathering or interfacing such data has been compromised (fingerprint scanners can be hacked[118]).

Recognizing the mounting public and private costs of data breaches and transactional fraud, there is a global trend towards mitigating damage at the source. This includes mechanisms for anonymizing the individual profile, to focus on the important stylistic features of a customer (a preference profile of a demographic profile which is not personally identifying), to tokenization of credit transactions.

[116] It is too early to imagine specifics, but the terms *data court* and *bad data damages* are at least evocative.
[117] https://www.oaic.gov.au/agencies-and-organisations/advisory-guidelines/data-matching-guidelines-2014
[118] See the Chaos Computer Club iPhone hack: http://www.ccc.de/en/updates/2013/ccc-breaks-apple-touchid

In the latter method, losses for fraudulent transactions are limited by using one-time disposable tokens of identity. These are encrypted forms of a real identifying datum which can be verified easily when the true secret is known, but which are cryptographically hard to decode for a bad actor.

Necessarily, the policies, protocols and technologies to protect identity and anonymize private facts into lesser shareable "customer profiles" and such like, lie at the intersection of legislative practise, public and private data custodians, cybersecurity experts and data architects and engineers. There is obvious private and public benefit to encourage better standards and practice in this area.[119]

6.3 Data Matching and Record Linkage

Data matching of names and other identifying characteristics can lead to "false matches" wherever there is ambiguity of naming, spelling, duplicates, data errors or other sources of confusion. Some private data records are already subject to release protocols for the remediation of data errors. In future, similar protocols may well be needed to remediate errors of erroneous data linkage. This is clearly related to the earlier commentary on knowledge bases and identity protection.

While the above issues with data matching are pretty obvious, and have been with us since the first government census[120], the important changes today are largely driven by advances in computation. For the record, the first US Census took about six years to process, by which time the next was ready to be started. That problem led the US government of the day to engage Hollerith to come up with a solution to the processing backlog. He invented and patented the punched card system for data entry that cut the processing time from six years to one year. The company that he founded, with the underlying punched card technology, later became International Business Machines (IBM).

In the 1950s, and 1960s further advances at the US Census Bureau were made in the use of digital computers and the development of the mathematical theory underlying probabilistic data matching. The advances today are in the parallelization of data matching and entity recognition algorithms so that they perform efficiently on very large data sets. While the underlying computer speeds are no longer improving, there is an industry-wide move to using multi-core microprocessors. The means to exploit this new scaling of computer power is via *massive parallelism*. Problems of graph search, the indexing of large document stores, and record matching are amenable to such solutions. However, the importance of the algorithm design has not diminished. In light of this fact, there is considerable effort devoted to towards improving the accuracy and speed of data matching on unstructured and semi-structured data sources such as webpages and human generated text, such as PDF files.

The practical upshot of these developments, alongside ubiquitous cloud computing, is that private firms can now process quantities of data akin to those of a national census in short timeframes. The best current example is Facebook, whose social graph contains around a billion individuals[121]. Queries against graph data can have prohibitive computational complexity[122], but matching of data has, at worst, a power-law complexity. For the practical cases, brute force is not typically required, since there are many avenues to block data in ways that enhance the probability of positive matching in a manageable amount of time. Hence it is possible to do a *much better job* of matching data across large data sets than it was previously.

[119] Due to the interdisciplinary nature of the problem, we think it does require Government sponsored effort.
[120] According to the ABS, the earliest census were likely taken in Babylon around 3800 BC http://bit.ly/1VfbJIW.
[121] The May 2001 estimate was around 720 million active users: https://arxiv.org/pdf/1111.4503v1.pdf
[122] The so-called NP-Complete family of problems are computationally hard and contain many graph questions.

The upshot of these technical developments, along with the explosive growth of big data from web and other customer records, is that record linkage for knowledge discovery is becoming much easier. Even small firms, armed with a good understanding of the appropriate technologies, are able to do a reasonable job of linking publicly available data to construct useful datasets. This can certainly lead to a significant enhancement in productivity for: sales lead generation; market research; customer segmentation; and demographic profiling. In our view, the most significant opportunities for public and private efficiency gains likely come from better use of linked data to inform decision making.

6.4 Data Provenance and Governance

There are many benefits to the use of knowledge graph technology in organizing the vast amount of data now published to the internet. However, there are also important policy implications regarding those circumstances where unrelated facts might be mistakenly linked to persons or companies. The larger question is how to establish the credentials of a datum in respect of its ultimate source. In the field, this is known as data provenance, and forms one part of wider data governance.

In the past, it was customary to store organizational data in record form within a structured database, such as a Relational Database Management System (RDBMS). Less structured methods of storing information, such as paper and electronic documents, are also typical, but have existed in a state of natural tension with the more structured solution. There is a very natural spectrum of data in use that spans the scruffy to the neat. Very often, the usage does not justify close attention to how accurate the data might be, when it was last changed, by whom and using what source.

The essence of good data governance is to know how loose or how tight to make controls upon data that are both a reasonable imposition on those involved and a reflection of the rights of those who may be affected by incorrect data. This is simply good organizational practice.

However, there are some changes to policy emphasis that are likely to represent better paths towards good data governance going forward. The essential change wrought by the web and newer unstructured and semi-structured data is best described as requiring *distributed data governance*. The data may be shared actively across organizational boundaries, with the potential for governance to fall between two stools. Once data is in motion, the very transportability of electronic data can soon make the origin and veracity of that data unclear. Provenance is a term that refers to electronic metadata which attests to the source of the data and some basic versioning and audit trail data. The practical need for that is represented in simple paper-based protocols of labelling shared documents as versions in a sequence, with a reference number and from a certain source. Technologies to make data provenance more seamless involve automating the creation of such provenance metadata.

6.5 Encryption and Rights Management

Supporting the above elements, the critical technology enabler to effective data policy is encryption. The use of encryption represents the only effective way to keep that which is now known to some as a secret to be kept from others. Encryption technology elevates a natural contest between those in whose interest it is to know a secret and those who choose to keep it. In the policy context, it is most fraught in connection with rules and conventions surrounding legal discovery and evidence. What likely makes a difference to the questions of efficiency and productivity is the legal clarity brought to bear on important edge cases that are related to due process, national security and privacy controls.

Centre for International Finance and Regulation

CIFR

Towards Financial System Integrity

CONSORTIUM

GOVERNMENT PARTNERS

Australian Government
The Treasury

NSW GOVERNMENT Department of Industry

AUSTRALIAN UNIVERSITY PARTNERS

UNSW AUSTRALIA

MACQUARIE University SYDNEY·AUSTRALIA

UTS UNIVERSITY OF TECHNOLOGY, SYDNEY

THE UNIVERSITY OF SYDNEY

Australian National University

THE UNIVERSITY OF MELBOURNE

RESEARCH CENTRE PARTNERS

CMCRC ASSURING MARKET QUALITY

sirca

INDUSTRY PARTNERS

CommonwealthBank

KPMG cutting through complexity

MACQUARIE

KING&WOOD MALLESONS

Centre for International
Finance and Regulation

Towards Financial System Integrity

Level 7, 1 O'Connell Street, Sydney NSW 2000 • Phone +61 (0)2 9931 9342 • www.cifr.edu.au