Submission to Productivity Commission Inquiry into Data Availability

Open Source Industry Australia Ltd

Amplifying the voice of the Australian open source software industry

Lodged 29 July 2016

About OSIA

OSIA represents & promotes the Australian open source industry by:

- Ensuring that the Australian business, government and education sectors derive sustainable financial and competitive advantage through the adoption of open source and open standards;
- Helping Australian Governments to achieve world leadership in providing a policy framework supportive of open standards and of the growth and success of the Australian open source industry; and
- Ensuring Australia's global standing as the preferred location from which to procure open source services & products.

OSIA's members are organisations in Australia who invest in or build their future on the unique advantages of open source software. For further information, see the OSIA website at http://osia.com.au.

Authors

Paul Foxworthy & Jack Burton.

Contacts

For further information in relation to this document, contact:
OSIA Director, Paul Foxworthy
OSIA Chairman, Jack Burton

Copyright

Contents

1 Executive summary

OSIA welcomes the opportunity to comment on the present Inquiry and we thank the Productivity Commission for that opportunity. OSIA has contributed to several of the Commission's inquiries and we have always been impressed by their rigour and consideration of the issues. We welcome evidence-based consideration of issues of importance and hope that the Commission's general approach is replicated more widely in the Australian public sector.

The present inquiry on the availability of data is important for OSIA's members and the Australian community. Good data underpins good decision making and good evaluation of programs and projects undertaken by our members, Australian businesses and government agencies. Good data leads to a more informed and empowered community. Better availability of data will also encourage a more transparent and responsive government.

Data isn't information; information isn't knowledge; and knowledge isn't wisdom. But each can underpin the next[1]. The promise of access to good data is that we can use, integrate and interpret it to advance the lives of Australians.

This submission reflects the experience and interests of our members. We hope our expertise in this area can help to inform the Commission and improve data availability for the Australian community.

[1] https://en.wikipedia.org/wiki/DIKW_Pyramid

2 Public Sector

2.1 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?*

Location and GIS data

Any dataset that includes a location is immensely valuable. Two datasets with good location data can be correlated and used in mashups.

For example, we congratulate the Australian government on making the Geocoded National Address File (G-NAF) and Administrative Boundaries open data.[2]

Transport services

Schedules and routes for transport services.

Utilities

Locations and alignments of electricity, gas, water, sewerage and telecommunications.

Health

Without breaching privacy, detailed health datasets will allow independent scrutiny of health procedures and outcomes.

Education

The NAPLAN data could be used for improving teaching techniques and evaluating the ”value add” of schools, teachers and programmes over several years.

Economic activity

Economic data on employment, unemployment, trading conditions.

Democracy and voting

While not strictly an issue of data availability, the EasyCount case is illustrative of a general malaise in access to public sector information.

The Australian Electoral Commission denied access to the ”source code” of EasyCount, the software that distributes preferences in Senate elections. In other words, there is no independent scrutiny that EasyCount implements the rules required by legislation.

The AEC denied a Freedom of Information request on the grounds that it raises revenue from the sale of EasyCount to other bodies that conduct elections, for example trade unions. In other words, the right of the Australian people to know and verify that an election is conducted fairly has been trumped by the fact that the AEC makes some income from the software.[3,4]

> *What characteristics define high-value datasets?*

Accurate

The data has minimal error rates.

Complete

The data is comprehensive and there is little or no missing data.

Significant

The data is relevant to an area of importance to the Australian community.

[2] https://blog.data.gov.au/news-media/blog/geocoded-national-address-data-be-made-openly-available

[3] http://www.theregister.co.uk/2013/11/07/electoral_commission_turns_down_source_code_foi_request/

[4] https://mjec.net/talks/lca2015/

Timely

The data is recent and not outdated.

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

A better informed community will be empowered to make better decisions. There will be better research and better evaluation of programs and projects. In the absence of good data, many decisions are based on hearsay, gut feel and appearances.

Such decisions are uninformed and possibly harmful, but perhaps worse, without good data it's difficult to evaluate whether a decision has been effective or not.

Better access to data will lead to innovative correlation and mashups that might not have occurred tothe original publisher of a dataset.

2.2 Collection and release of public sector data

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

Risk aversion

When there are no benefits to risk-taking, and considerable downside, quite understandably government agencies will tend to avoid those risks.

For example, NAPLAN data on the achievement levels of primary and secondary school students is difficult to obtain, and is approved on a case-by-case basis by ACARA.[5]

The summary data is being used very poorly for ”league ladder” grading of schools. It should be being used to to identify gaps in students' understanding, and thus help improve teaching techniques.

There are political sensitivities to this data, like many other public sector datasets. But there is immense potential for public good with innovative use of the data, and the potential benefits should outweigh risks of embarassment to a few politicians.

False sense of ownership

There should be a general culture of ”the taxpayer has already paid for it, and therefore should have access to it”, but this is not the norm. The US does a better job of this, where any work created by a federal government employee or officer is in the public domain, provided that the work was created in that persons official capacity.[6]

Government agencies are all too ready to hoard valuable data that in truth belongs to the Australian people.

Commercial return

Government agencies seek to raise additional revenue from data that the taxpayer has already funded. For example, genealogy enthusiasts love to obtain family history data, and agencies charge prices that would seem much greater than their costs.

One-off thinking and projects

Government agencies are often given one-off funding to set something up for a specific use case. That's understandable for a point solution to a smaller scale problem.

Agencies are often less effective at building sustained national and international data sets. Typical government purchasing practices will need to be varied for long term, sustainable, strategic programmes.

The most effective way to build long term value is to nurture, encourage and support an international community, based on specific goals, and have the community help to build data value. Open Street Map[7] is a good example of this in action.

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

Structured data first

It is straightforward to produce human-readable documents in formats like a web page or PDF from structured data. It is difficult to go in the other direction — to ”scrape” a human-readable source to obtain structured data.

[5] http://www.acara.edu.au/_resources/D12_1573__ACARA_Data_Access_Protocols_2012.pdf
[6] http://fairuse.stanford.edu/overview/public-domain/welcome/#us_government_works
[7] http://www.openstreetmap.org/

The fact that scrapers like QuickCode (formerly ScraperWiki) [8] are so widespread is testament to the fact many agencies are publishing in the wrong format. The irony is that in almost every case the agency holds the data internally in a more structured format.

APIs

Agencies should be publishing data using web Application Programming Interfaces (APIs). This makes it easy for anyone who is interested to integrate the data into another application or web site.

Allow the public to talk back

Open source projects like PHP [9] benefit immensely from their audience being able to talk back. Where information is incomplete, inaccurate, misleading or plain wrong, there is an opportunity to do something to correct it.

Many public sector datasets would benefit from the same thing. Again, family history is one area where this is done very badly in general. Historical data was transcribed from handwritten documents and inevitablt there is a high error rate. The easier it is for people to talk back, the more likely that corrections can be made and the quality of data improved over time.

> *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?*

Benefits of standardisation

OSIA strongly agrees that gathering and publishing data should be standardised.

- **Discoverability:** When standards are well known, it will be easier for those interested to locate and use data.
- **Encourage best practice:** Standards will make it easier to encourage and disseminate good practices among public sector agencies.
- **Training:** Agencies can develop common training strategies in accord with standards. Training will be more effective and more cost-effective.
- **Reduce duplication:** Standards will lead to less duplicated effort. Data produced by one agency would be more likely to be directly usable by another.
- **Privacy:** Standards will help develop robust privacy management strategies that could be used in many agencies.
- **Reduce ad hoc decisions:** Standards will lead to more consistent and considered decision making.
- **Silo busting:** When there are common, standardised datasets and practices, that will ease the building of a common pool of data and ease the exchange of data. When the inevitable reorganisations of public sector agencies happen, they will be easier to implement.

Open formats

All data should be available in open formats, so there is a wide range of software available to read it and so that reader software is not controlled by one closed-source vendor.

Metadata

There is a huge amount of work being done on attaching decriptive information to raw data to make it much more understandable and discoverable.

data.gov.au has information on metadata[10]. We also recommend the Commission and those interested in enabling access to public sector data look into the Semantic Web efforts of the World Wide Web Consortium.[11]

> *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?*

OSIA does not have detailed knowledge of existing criteria and tools. Here we offer some thoughts on what we think agencies *should* do.

[8] http://scraperwiki.com

[9] http://php.net

[10] https://toolkit.data.gov.au/index.php?title=Discovering_Metadata

[11] https://www.w3.org/standards/semanticweb/

Default to openness

The very real impediments and resistance to open data can be reduced by requiring agencies to make the case *against* openness rather than for it.

This has been done in the US[12]. The Office of the Australian Information Commissioner sets out a similar policy for Australia.[13]

We wonder how seriously the OAIC's guidelines were followed when the Office was neglected and threatened with closure for several years.[14] We welcome the more recent highlighting of these issues by the Prime Minister.[15]

Broad privacy principles

As we have mentioned, standards for privacy evaluation will help deliver good, clear, reasoned and consistent decisions.

> *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?*

data.gov.au

We admire the efforts made by data.gov.au and the indefatigable Pia Waugh during her time there. data.gov.au has provided both a starting point for searching for public sector data, and an accessible way for agencies to learn more about doing so.

Prime Minister's support of open data

There is considerable prestige and influence when the Prime Minister encourages open data.[16].

Open Council Data

The Open Council Data initiative led by the Municipal Association of Victoria has real potential.[17]

2.3 Data linkage

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

Co-ordinated location data, combining and correlating data from different sources, has potential.

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

Exclusive commercial publishing

Exclusive publishing arrangements create monopoly barriers to integration. Where a commercial publishing arrangement is non-exclusive, the publisher has an incentive to add value and new publishers can compete where an opportunity is found. The Australian community can go to the original source if a commercial provider isn't adding value.

> *How can Australias government agencies improve their sharing and linking of public sector data? What lessons or examples from overseas should be considered?*

Data.gov.uk is an information portal somewhat similar to Australia's.

The Estonian government has a goal of enabling online access for all citizens and every interaction with the government. Open data is part of that.[18]

[12] https://www.whitehouse.gov/blog/2014/05/09/continued-progress-and-plans-open-government-data

[13] https://www.oaic.gov.au/information-policy/information-policy-resources/principles-on-open-public-sector-information

[14] http://www.itnews.com.au/news/oaic-saved-from-dissolution-418971

[15] https://www.dpmc.gov.au/public-data/open-data

[16] https://www.dpmc.gov.au/public-data/open-data

[17] https://opencouncildata.org/

[18] http://www.opendata.ee/en/hetkeolukord-eestis/

3 Private Sector

3.1 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?

Housing and real estate

Twelve years after Alan Kohler lamented the state of data gathering for real estate sales[19], there is a still a problem with complete and timely data in housing. For the vast majority of Australians, their house is the biggest investment they will ever make and government policies to encourage home building and ownership should be informed by good data.

Privately owned utilities

The electrical distribution system and telecommunications facilities are essential services and data about them should be widely available.

Transport services

Schedules and routes for transport services.

Company financial results

Most Australians are now significant shareholders via their superannuation. Good data on company performance will help Australian make better and more informed investment decisions.

In each case cited, what characteristics define such datasets?

The characteristics are the same we mentioned for public sector data: Accurate, Complete, Significant, and Timely.

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

Where there is substantial significance and benefit to the Australian community, government intervention to require the sharing of data can be justified.

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

Better decision making based on good data can help the welbeing and wealth of Australians.

3.2 Access to private sector data

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?

OSIA has commented before to the Commission on the excessive length of copyright terms in Australia[20]. Reducing the length of copyright would help access to historical data.

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

There are concerns that OSIA believes can be managed.

- *Privacy and other liability* A private publisher of data might be accused of a breach of privacy. Clear guidelines and practices can ameliorate this risk.
- *Competition* The data might help a competitor. Ensuring a level playing field so all organisations are required to release similar data will help here.

[19] http://www.theage.com.au/articles/2004/06/14/1087065081824.html

[20] Burton, J. & Foxworthy, P., *Final submission to the Productivity Commission's Inquiry into Intellectual Property Arrangements*, Open Source Industry Australia, 2016., pp. 22–23.

- *Support load* Once published, an organisation may face questions or disputes about the accuracy of the data. Where data has been gathered in good faith, a company's obligations to answer such questions should be limited.

> *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?*

Everything OSIA has said for public sector data applies here as well.

Benefits include discoverability, encouraging best practice, common training strategies, reducing duplication, effective privacy management strategies, and reducing ad hoc uninformed decisions.

Open formats and descriptive metadata should be encouraged.

> *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?*

Very much. Community based projects will lead to a long tail of contributions from many, many people. Again, Open Street Map is an example of this.

Success breeds success. Where data sharing is demonstrably working and giving benefits to all participants, businesses are more likely to join and participate.

> *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 (Cth)) relevant?*

We leave a judgement on legislation to experts on competition law. However, we believe pooling of data is little different from benchmarking or participating in an industry association. Businesses co-operate with their peers and competitors where it's in everybody's interest to do so. It can be done without anticompetitive practices arising.

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

Governments can promote open data by setting a good example. Many tools and policies that are developed for the public sector would be useful for the private sector as well.

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

There has been interesting work on data philanthropy, where private sector data has been used to help enable development in less developed countries. [21,22,23]

> *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?*

Where data is generated by an individual and clearly their own work, they should retain rights over it. There have been ongoing struggles where online services have appropriated the words and images of their customers.

We can encourage the sharing of useful aggregate data without impinging on the rights of individuals.

Unclear ownership of data is similar to the problem of orphan works in copyright law and needs to be handled carefully.

[21] http://www.unglobalpulse.org/RDF-private-sector-data-summary

[22] http://www.unglobalpulse.org/blog/data-philanthropy-public-private-sector-data-sharing-global-resilience

[23] https://hbr.org/2014/07/sharing-data-is-a-form-of-corporate-philanthropy/

4 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 existing provisions of the *Privacy Act 1988 (Cth)* granting consumers the right to access data about themselves held by agencies and businesses are probably sufficient in scope as they stand.

However, the Act's provision on the form in which such data are to be provided could use improvement.

As noted in the issues paper, the requirement is that it "must be in a manner requested by the individual if it is reasonable and practicable to do so". That would seem too vague and far too easy to circumvent.

The ability of the requesting person in each case to define the form may well give rise to unreasonable data conversion costs for the agency or business (since multiple unrelated requesting persons may well each prefer different forms).

On the other hand, the phrase "reasonable and practicable" provides agencies and businesses with a very easy (and highly subjective) path to refuse the form requested and substitute their own form (so long as they still deliver the requested data).

As the Commission correctly noted in its issues paper, data released in a structured, machine-readable form is likely to be of far greater use to the requesting party than data released in an unstructured "document-like" form.

Naturally, the question then arises as to *which* specific structured, machine-readable representation should be preferred? There exist an infinite number of potential forms of representing any data.

The common-sense answer of course is that the best form will vary depending on the nature of the data, but the form selected in each case should be the form which is likely to be most convenient for the requesting party and that will impose the lowest possible (ideally zero) additional processing costs on the releasing party, assuming that the releasing party has followed recognised best practices in designing and implementing its own internal records / information management systems.

A good start would be to require that data released under *Privacy Act* requests (and indeed also under FOI requests) always be released in a form defined by an unencumbered open standard, in any case where such a standard exists. Where a suitable standard does not exist, releasing parties should be encouraged to collaborate on development of a suitable unencumbered open standard.

There has been much confusion generally around what constitutes an "open" standard. It is a question that OSIA have addressed before, and we offer the following definition:

> *To be regarded as an open standard, a standard's published specification* ***must*** *be exhaustive (either in and of itself or in conjunction with references only to other open standards) and the standard* ***must not*** *be encumbered by patents or any other artificial barriers to proliferation.*
>
> *An open standard* ***should*** *be developed and maintained by a process analogous to the open source development model.* [24]

In the case at hand, the principal benefit of requiring disclosure to be in a form defined by an unencumbered open standard is that *anyone* skilled in the art of computing can implement software tools to transform the data to or from that form (or indeed to store or analyse it in that form entirely), and in most cases open source software tools for that purpose are available. The importance of available tools being open source stems from two factors: limiting costs and ensuring auditability – both of which carry benefits both for the requesting party and for the disclosing party.

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

As stated above, the relevant existing provisions of the *Privacy Act* are likely to be sufficient in scope. OSIA only sees a need for more clarity around the form of disclosure.

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

Ideally, that sort of shift should be achieved through market forces, rather than through regulation.

As always however, exceptions to that rule will likely need to be made in case of monopolies, near-monopolies and virtual monopolies, since in those sorts of markets the ordinary commercial imperative for an organisation to meet consumer expectations is conspicuously absent.

[24] Burton, J., *What does "Open Standard" really mean? Any why should I care?*, first presented at *Removing the silos: a future built on unencumbered open standards*, Open Source Industry Australia Ltd, Adelaide, 15 Oct 2014. That working definition was arrived at following a comparative analysis of 22 different definitions of "open standard" in wide use, published by various academics, standard-setting organisations, other private sector organisations and governments worldwide.

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

Where consumers are requesting *personal* data about themselves, third party intermediaries should be avoided as rule, since each additional set of hands through which personal data passes gives rise to an additional, unnecessary risk to privacy, thereby running counter to the purpose of the Act.

In a more general setting, for example where broader Government data sets are being released, third party intermediaries can sometimes add value to the system but can also subtract value from the system.

They key principle should be that any non-confidential data set is *always* released directly to the public, without requiring the use of intermediaries, but at the same time there should not be any prohibition on intermediaries entering the market.

The market is quite capable of deciding for itself whether any given intermediary has demonstrated that it adds tangible value to the raw data set. That decision of course can only be made correctly by a free market, which has not been distorted by Government mandating the use of third party intermediaries (as happens at present, for example, with the data sets on corporations released by ASIC).