AURIN
Australian Urban Research Infrastructure Network

Level 2 West, Alice Hoy Building
The University of Melbourne VIC 3010
E: admin@aurin.org.au
T: +61 3 8344 3212
aurin.org.au

Productivity Commission Inquiry into Data Availability and Use – Australian Urban Research Infrastructure Network (AURIN) submission

Introduction

Since 2010 Australia's urban researcher community have united to build and implement the Australian Urban Research Infrastructure Network (AURIN). The focus of AURIN is on providing meaningful data and knowledge – urban intelligence – as the evidence base for informed decisions about the smart growth and sustainable development of Australia's cities and towns.

Funded by the Federal Government through the Education Investment Fund (EIF) and the National collaborative Research Infrastructure Strategy (NCRIS) schemes, AURIN has implemented a 'portal' capable of facilitating the seamless access to, and use and reuse of public, government and commercial data. Currently researchers and policy-makers across numerous fields use the AURIN portal to access relevant data and undertake insightful analysis to progress their critical programs in the fields of population and demographics, economic activity, well-being and quality of life, housing, transport, energy and water consumption and innovative urban design.

Since the formal launch of the AURIN portal, just over 2 years ago, it has registered 4500+ users nationally and this number continues to grow. Whilst the primary audience for AURIN has been academics and students from Australian research institutions, e.g. universities, the infrastructure developed has been increasingly adopted by federal, state and local government policy makers for the source of data and tools to guide their activities, including regional productivity and planning.

AURIN is uniquely positioned to support research and training of the current and future workforce. An opportunity exists to fully leverage this system to build capacity and skills in all areas of data collection and the practical ramifications for data access and sharing across organisational boundaries. Few other projects have realised a technical infrastructure that tackles the data rich environment represented by modern urban settlements.

This submission covers three main topics: (1) the need for 'high value datasets', (2) improving access to data and data sharing, and (3) managing the costs of data access and utilising technology to improve access to data. AURIN has four recommendations for the Commission to consider:

Recommendation 1: Make **‘**high value’ datasets such as Land Use, Property Value, Business Activity, Health and Education available for the purposes of research, or more precisely ensure that such data sets are collected with data sharing in mind from the outset.

Recommendation 2: Remove impediments to accessing data via a consistent approach to licencing data for sharing across platforms such as AURIN to wider scale communities crossing the academic, commercial and government sectors.

Recommendation 3: Manage the cost of data by improving the documentation of the data at the source, including adoption of common standards and licence models, and thus enabling researchers to use data with confidence.

Recommendation 4: Utilise technology to improve the quality and currency of data (and associated metadata) at a fine scale (resolution), so the data is able to be applied for research, and importantly, leverage technologies that overcome data heterogeneity.

High Value Datasets to AURIN (Recommendation 1)

Since inception, AURIN has brought together eminent researchers and discipline leaders (list of Lens Leaders attached) in urban research to identify ‘high value’ datasets, which have the potential to provide the greatest impact to their fields of work. During AURIN’s development these fields of research were referred to as Lenses. The identified Lenses included topics such as: population health and wellbeing, transport, economic planning, energy and water use and innovative urban design. Across these Lenses, a common requirement was finer-grain spatial data – that is, data relevant to the property or neighbourhood or to the individual. A review of the data lists collated by these research leaders identified the following datasets as being of especially high value:

Land Use Data

Within the AURIN network, one of the greatest limitations identified was the lack of availability of highly disaggregated and accurate land use data. This data exists within various land valuation systems across Australia and is urgently needed for a variety of uses. For example, researchers within the McCaughey VicHealth Community and Wellbeing Unit require this data to understand relationships between health and neighbourhood design. Fundamental to this work is understanding how the land use mix influences walking activity (Giles-Corti, et al., 2014). This knowledge can inform policy makers on how neighbourhoods should be designed to improve health outcomes. Different state and Federal Government departments collect land use data (e.g. state-based registries of property valuations, the Federal Government Australian Business Register (ABR) and commercial providers such as Pitney Bowes, PSMA, Sensis). However, researchers, and AURIN, often find limits on access to the data from government departments, while equivalent data available from commercial data providers may not be as accurate, and prohibitively expensive.

The availability of more fine grained Land Use Data would support robust decision modelling. At present availability of this data is dependent on resources available for manual curation of this data requiring specialist work by a data analyst. An example of the use of this data in AURIN is the "What if Planning Support System", which was used by the Department of Planning in Western Australia to model land use change driven by population growth (Perth and Peel @ 3.5 million). To meet the data requirements for this project an AURIN data analyst spent three months curating and validating this data. The outcome of this project directly influenced the Sub-Regional Planning Framework released in May 2015. However greater availability of this data would reduce time and resources necessary to validate the data for research.

Property Valuation Data

In response to the need for sustainable planning, researchers Newton et al., (2012) have been developing tools to provide state and local governments with the capacity to strategically manage growth and urban renewal within neighbourhoods. Valuation data is critical to this project, as it indicates which properties are likely to be redeveloped and therefore which areas will undergo significant change. To date, project managers have had to acquire valuations data from local governments in manual and/or ad hoc approaches, which is inefficient and generally leads to only partial metropolitan coverage, even with the assistance of relevant state planning authorities.

Some valuer generals' offices across Australia are actively pursuing models for the commercalisation of property data, but this is at odds with the Open Data Government Initiatives underway nationally, and in Qld, NSW, Victoria, and WA. To address the need for geocoded research ready property data, AURIN previously purchased property price data from Fairfax Australia Property Monitors (APM). The cost of the contract in place is $250k per annum. The cost of this data is a major impediment to any research or government activities requiring access to such data.

Economic/Business Data

Across Australia, urban policy makers agree that cities need to adopt multi-nodal metropolitan planning strategies in order to foster local economic development and reduce commuting. Particularly in metropolitan Melbourne, planning strategies have sought to promote non-CBD centres. Work by Jennifer Eve Day and colleagues (2016) has investigated the effectiveness of spatial policies aimed at employment clusters. Unfortunately, a lack of available geocoded business data has restricted this work. The most promising current data source are the Australian Business Register (ABR) datasets. However, ABR non-public data that contains geocoded information with the ANSLIC Code (which defines the business activity) is restricted by Commonwealth tax legislation. The ABR can only release this data to agencies who enter into an agreement with the ABR. This means ABR data can be made available to partners of the ABR, but not the broader research community. An additional issue is that the ABR data does not include crucial data such as firm size, turnover, value added, and capital investment. If such data were made available, they could provide a spatial picture of how zoning policy within cities could facilitate clustering of specialised economic precincts. It would also enable discovery of the specific sector-based economies and constraints that exist within cities and indeed opportunities that could arise, e.g. where is the optimal location to establish a new business opportunity in a particular sector?

Education Data

Data on the education of students are being collected and made publicly available via the MySchools website, but only on the basis of specific requests (e.g. information about a particular school). This presents an issue, which some users are currently addressing by ‘web scraping’ the data from the MySchools website (or obtaining the data from web-sites without regard for the legal terms governing use of that data). This data is highly valuable for researchers wanting to understand the implications of the built environment, soci-demographics, school location on education and would be of great value to AURIN users. Making the data available under an open data licence that allows for data use including research would eliminate the need for web-scraping and encourage data reuse.

Public Transport Data

Accessing and interrogating smart card data offers many opportunities for better understanding of the behaviour of public transport users, including multi-purpose trips. There is a significant amount of data available through systems such as Myki (Vic), Opal (NSW) and Translink (Qld) and myriad opportunities exist in this rich data source to better understand commuting behaviours and improvements that might be modelled – this data, however is not widely available. As a consequence, major opportunities to improve public transport using actual travel data analytics are not currently possible.

Private Sector Data

Whilst the public sector holds the majority of datasets identified as being of especially high value to researchers, private sector datasets in this category include property price data. Currently in Australia there are two main providers of property data: CoreLogic and the Australian Property Monitor (APM). As noted previously, typically this data is only commercially available, at significant cost.

Of rapidly growing importance is the new data being generated through online platforms, such as crowdsourcing, and especially social media. In addition, a new economic paradigm, exemplified by organisations such as AirBnB and Uber, is emerging. These companies are also significant data generators and holders and are becoming of great interest for researchers who are focused on understanding the way we live. Multiple ad hoc approaches have been followed to collect and analyse such data for a rich variety of scenarios of high relevance to urban environments (Sinnott et al., 2015, 2016), but a larger scale nationwide capability does not yet exist.

Remove impediments to data access and availability (Recommendation 2)

As noted in the issues paper, and above, there are a number of challenges surrounding access and use of public sector data, which serve as barriers to data availability and impactful research. This section outlines a number of these issues with respect to geospatial data for the urban research community but may touch on data access issues identified more broadly in the issues paper.

Licensing

Uncertainty surrounding ownership of data and how it can be used is a key barrier for the use and reuse of data. Data licensing can play an important role in making users aware of how different datasets can be used and allowing organisations and individuals who created the data to be appropriated attributed when data is used. In particular, open data licensing regimes like Creative Commons can be used to manage rights that that automatically attach to all creative material (including data) protected by copyright. They provide organisations with a low cost way to make their data available for different purposes.

Currently AURIN advocates for the use of Creative Commons licences to be applied. AURIN believes any data should, by default, be open and available through a Creative Commons Licence, unless clear impediments exist such as privacy constraints. Data licensing and open data licensing can provide certainty for the data provider and user, and unlock data for innovation to occur.

At the same time AURIN's experience is that data licensing can be closely aligned with broader due diligence necessary the release of data. This includes for example data curation, de-identification, clearance for commercial sensitivity and IP issues. AURIN is currently developing a checklist for release of data through the AURIN Portal to address these issues. Allocating costs to data due diligence associated with data release forms part of the costs of managing data addressed in Recommendation 3.

Clear documentation and metadata

The Federal Government can enable the wide distribution of data, by mandating that data collected with public money is accurately described using the appropriate data standards. This requires data to be accurately described with the source, purpose, date and accuracy explicitly recorded. Without this, data is typically not shareable or useable.

Third party data collection

Government often outsources data collection to consultancies and this data is generally presented in a report and thus restricted in its availability for reuse. Similarly, data generated through research funds such as ARC and NHMRC grants are often unavailable for re-use. Currently there are some funding schemes that require data to be deleted after research projects are complete.

In addition, there appears to be a growing number of websites from individual government agencies and industry groups which provide data on the site but deliberately restrict data use to personal/individual use of the data. These agencies need to be encouraged to licence the data to permit the reuse of data for research. When such data is identified by members of AURIN the agencies are contacted and the process of releasing data is explained and ad-hoc assistance provided. One resolution to this issue would be for agencies to adopt common open data licence standards such as Creative Commons for the release of data. Three of the largest sources of Federal Government data sets — Australian Bureau of Statistics (ABS), Geoscience Australia and data.gov.au — are all licensed by default under Creative Commons Attribution licences. These sites provide free access to all of Australia's Census data, official geoscientific information and knowledge, and other miscellaneous government data.

Managing the costs of data (Recommendation 3)

There is a significant cost to the collection, maintenance and distribution of data. AURIN utilises an open source architecture for accessing data from over 60 different organisations. It aims to access all this data through a federated architecture where the data remains with the data custodian and is accessed on demand by the (user /researcher). The current problem with this approach is the lack of data services being maintained by government departments and hence much of the data has had to be hosted locally by AURIN, requiring manual updates whenever the custodians amend their datasets.

Recommendations related to managing the cost of data release include:

- Determining common standards, as currently across Australia there are a number of agencies using different or no standards for the description of data. The current situation makes it difficult to compare and contrast regions across Australia and to undertake national research. AURIN has done project work in this area: eg: consistent geospatial analyses using diverse geo-classifications (postcodes, SA4-SA1, LGAs, SLAs, meshblocks etc and how they evolve over time). This facilitates better economic analysis, urban management and future planning as well as reducing the burden on individuals who are constantly working to improve the data.

- Empowering government agencies to provide data through web services (APIs) and using common standards referred to above to achieve this, e.g. geospatial services and associated standards, statistical data services and associated standards etc. Ultimately this will allow greater reuse and reduced cost through following established standards with technologies that are fit for purpose.

Improve the quality, currency and spatial resolution of the data (Recommendation 4)

Over time technology has evolved and now platforms such as the AURIN portal and associated e-infrastructure have provided a robust platform for the transfer of data between researchers and/or trusted parties using next generation big data technologies. In this section some of the ways that technology has advanced to enable improved data quality, access and security are outlined.

Data Heterogeneity

AURIN has had to tackle completely heterogeneous data scenarios. It could not mandate any technical solution on any given provider. Through advanced computing capabilities including use of noSQL databases, AURIN has shown how ad hoc and seemingly non-comparable data resources could be seamlessly accessed and combined in arbitrary ways by researchers. Such technologies are increasingly prevalent and could help in many government data management scenarios more generally.

Trust, Security and Capability

Privacy is often cited as one of the primary reasons for restricting access to data. In fact, with advances in technology there are a number of mechanisms that can be put in place to protect privacy. These include:

Data Aggregation

In 2006 the ABS implemented the Mesh Block as the smallest geographic building block for spatial data in Australia. The concept was to allow for data to be aggregated to the Mesh Block or a group of Mesh Blocks (based on the user’s request) and released at the appropriate geographic scale as to protect privacy of an individual (Eagleson et al., 2002, ABS 2003). Unfortunately, the uptake of this concept by the ABS has been limited and requires further development. Unit level data is also often required *for use* by researchers, e.g. the address of correspondents in a survey and the distance they live from public transport. A rich range of privacy protecting solutions are now available and this data limiting argument is often no longer valid (Sinnott et al., 2016).

Role-based security

In AURIN authorisation-oriented security has been implemented to restrict system access to privileged users. Currently we are using this security system to restrict access to certain datasets. For example, the ABR data is restricted to partners of the ABR - these partners are authenticated at the source.

Single Sign-on

There is a clear need for single sign-on across government. The AURIN platform can be accessed and used by any academic across Australia through use of their own institutional credentials, i.e. the home institution username and password. Once authenticated, researchers can access and use any data sets that their credentials allow without any further challenge/response for usernames/passwords. Data sharing often begins with security, and further progress can be made from the technologies that support sign sign-on.

Where a dataset is deemed to require additional security then 'new technology' platforms which enable security should be employed. AURIN is one such system developed for urban research data (ie data that is geospatially referenced).

Summary

Meaningful, diverse and accessible data is the key to enabling researchers to provide Australia with new knowledge and innovation – which is essential to this country's successful growth and sustainability. AURIN's visibility across multiple sectors puts it in a unique position to provide the insights detailed in this submission about the challenges and solutions that should be considered. AURIN looks forward to working with government and other stakeholders to embed these improvements in the future

Contributors

Dr. Serryn Eagleson, Ms. Emma Williams, Prof. Ian Bishop, Prof. Richard Sinnott, Dr. Steven Glackin, Dr. Jack Barton, Dr. Jennie Day, Prof. Billie Giles-Corti, Dr. Suzanne Mavoa, Dr. Claudia Pelizaro, Ms. Rachel Lerm and Mr. Andrew Dingjan.

References:

ABS (2003) 1209.0 - Information Paper: Mesh Blocks,
2003 (Available http://www.ausstats.abs.gov.au/ausstats/free.nsf/0/FF3E891FE426F412CA256E5800759B5C/$File/12090_2003.pdf (Accessed: 20/06/2016)

Day, J., Davis, B., and M. Spiller (2015) The Dilemma of Urban Employment Land: An Approach for Assessing the Value of Small Urban Manufacturing in Inner Melbourne in Economic Development. In Economic Development Australia Journal - IMAP Volume 8. No. 1

Eagleson,S., Escobar, F. and I.P. Williamson (2002) Hierarchical Spatial Reasoning Theory and GIS Technology Applied to the Automated Delineation of Administrative Boundaries in Computers, Environment and Urban Systems 26,185-200.

Giles-Corti, B., Mavoa, S., Eagleson, S., Davern, M., Roberts, R., and H. Badland (2014) Transport Walkability Index: Melbourne. McCaughey Centre for Community Wellbeing, Melbourne: The University of Melbourne.

Newton, P., Newman, P., Glackin, S., and R. Trubka (2012), Greening the Greyfields: Unlocking the Redevelopment Potential of the Middle Suburbs in Australian Cities. World Academy of Science, Engineering and Technology. 71: pp. 138-157.

Sinnott, R.O, C. Bayliss, A. Bromage, G. Galang, Y. Gong, P. Greenwood, G. Jayaputera, D. Marques, L. Morandini, G. Nogoorani, H. Pursultani, M. Sarwar, W. Voorsluys, and I. Widjaja (2016) *Privacy Preserving Geo-Linkage in the Big Urban Data Era,* Journal of Grid Computing.

Sinnott, R., P. Chhetri, Y. Gong, A. Macaulay and W. Voorsluys (2015) *Privacy-preserving Data Linkage through Blind Geo-spatial Data Aggregation*, IEEE International Symposium on Big Data Security on Cloud (BigDataSecurity 2015), New York, USA.