Summary

The Productivity Commission released the 'Data Availability and Use' issues paper, in April 2016. This submission has been prepared in response to the issues paper and addresses data availability using environmental data as a case study to highlight issues including:

- Classification and separation of data types (e.g. personal and non-personal data, data devices, etc.);
- Consideration of 'high value' data given:
    - A single problem may require many data sets; and
    - A single data set may serve many applications.
- A requirement for easy to formulate queries, for public to access the data;
- Governance requirements vary across the data lifecycle;
- A requirement for database maintenance funding – post project completion, as data storage requirements may outlast project completion;
- Consideration to data generated from the Internet of Things and associated issues;
- Unintended consequences of increased data availability – process(es) that allow citizens to challenge the scientific data incorporated in agency rule making and permit decisions.

Note, while environmental data has been used as the example to highlight the issues concerned, it is more broadly applicable to data in general.

About the Author

I am a private a PhD scholar at the School of Regulation and Global Governance (RegNet) at ANU. My research looks at how data from the Internet of Things (IoT) (e.g. sensors for pollution monitoring, smart cities and grids, etc.) can be used for environmental regulatory purposes with the aim of determining whether it can be used to improve the responsiveness of (environmental) regulation.

The opinions expressed in the submission document are the authors and do not necessarily reflect the views of RegNet. Any remaining errors or omissions are the responsibility of the authors.

Data Types - Heterogeneity and Integration

The issues paper refers to the need to draw a distinction between data and information. However I would suggest that the Productivity Commission go further and distinguishes data types. Currently as it stands the term data is broad sweeping and does not distinguish between data types (e.g. personal and non-personal data, data devices, etc.).

Given the vast landscape of issues (e.g. security, privacy, data provenance, consent, etc.) associated with data generation, capture and collation, use and re-use of the data; by segregating data, one is able to more clearly determine the ontological traits of data, identify issues, classes of issues, and possible solutions. It is envisaged that with greater granularity may come greater clarity regarding issues associated with data (e.g. privacy, security, access, ownership, control, rights, etc.) and impediments to access to the data.

Given the plurality of data being investigated to be made publicly available by the Productivity Commission, an ontological approach to solve problems arising from using different terminologies should be considered for classification, clustering, extraction, and bridging the semantic gap.

Environmental Data as an Example

As environmental issues have become more prominent over time and communities increasingly wanting a right to know about the quality of their environment, certain types of environmental data is made publicly available and hence can be considered a public asset. However under the current paradigm, both the type of environmental data made publicly available and the aggregated nature of publication of the data, reduces the richness that granularity provides. Hence, whilst data is being made publically available, one can question the efficacy of the public environmental data. The Productivity Commission should therefore consider the type, scale and granularity of the data being made available.

Environment – a coupled system

The environment is a dynamic system of coupled systems (see Figure 1[1]). To address complex real-world (environmental) problems, there needs to sharing of data between communities, regions, sectors, disciplines and applications. However, water data is held many institutions and organisations and is stored in various repositories as unstructured, semi-structured or structured data. To exacerbate matters, data is stored in silos and on legacy systems, resulting in most organisations unable to have a uniform view of data and in some instances not being able to integrate data from disparate data sources. Further, there is limited sharing of data between government departments and agencies, amongst state governments and between government and private organisations. Hence, whilst data may be available at the local and sub-national level, it is difficult to 'harvest', thereby inhibiting its use.

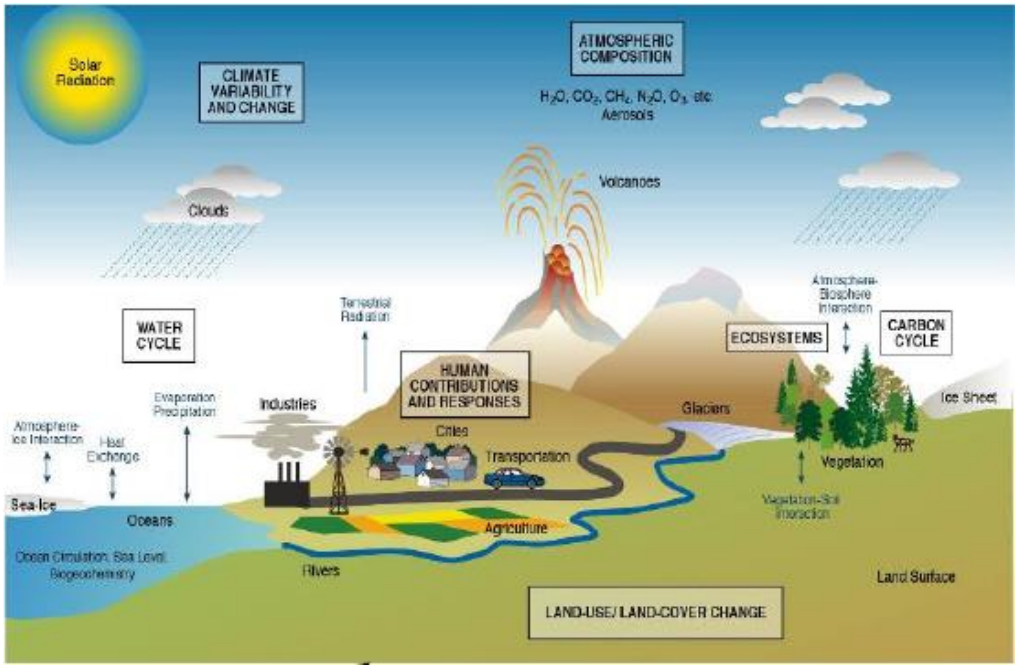

Figure 1 Environment as a Coupled System

Klein (2015) noted that:

- A single problem may require many data sets; and
- A single data set may serve many applications.

[1] http://studylib.net/doc/15380143/thomas-klein--ecds

Hence due consideration should be given to determine datasets of 'high-value', as the linking of diverse sources and types of data can generate value greater than the sum of the parts (pg 6, issues paper).

Internet to the Internet of Things

Additionally, data was traditionally generated by industry and government departments and regulators. With the advent of the Internet of Things (IoT), people and communities will be another data generation source. It is therefore recommended that the Productivity Commission also look at data generated from the IoT, to address the issues and challenges in making IoT data available for use.

The Productivity Commission to consider investigating the data generated from the IoT, as part of the draft report. However, it should be noted that building on policies of the Internet era and applying them to the IoT era (e.g. connected consumer goods) may provide ill-fitting solutions, as IoT data is potentially difficult to anonymise, creating privacy problems, and difficult to secure (Peppet, 2014).

Public Access - Keyword Query

Country specific data about greenhouse gas (GHG) emissions is easily available using keyword queries in search engines, as shown in Figure 2[2]. However a more detailed source of GHG emissions by area, sector, region or obtaining data by year requires more advanced search. This search paradigm assumes the user to be an expert of the underlying data and domain.

Hence, whilst the Productivity Commission is considering how to make data easily available, without easy to formulate queries, ordinary users may find it a challenge to access the data, thereby impeding the availability.

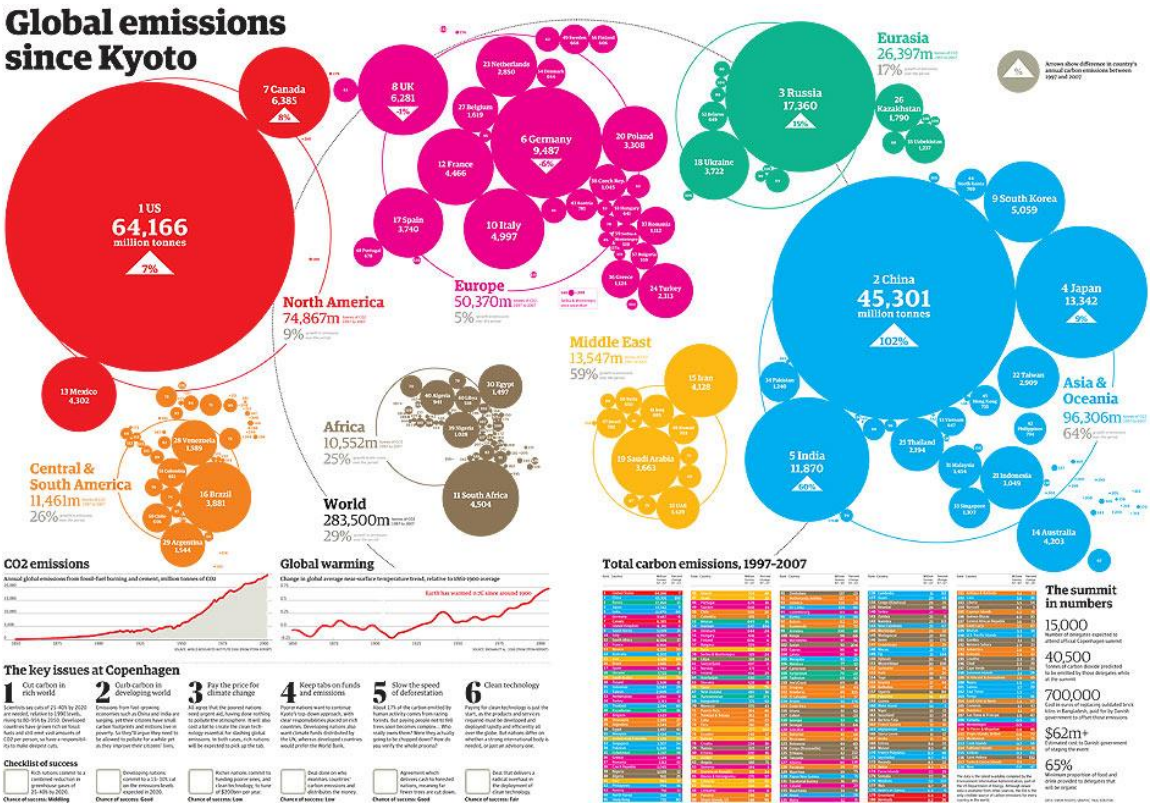

Figure 2 Global GHG Emissions since Kyoto

[2] https://static.guim.co.uk/sys-images/Guardian/Pix/pictures/2009/12/7/1260181048856/Carbon-emissions-graphic-002.jpg

Public Participation

The current paradigm is the regulator/agency develops the regulatory standard with limited public consultation. However, citizens often become involved in the regulatory issue long after the rules have been adopted and there is little incentive to reopen the debate about the choice of inference and / or assumptions selected. As data becomes more available and the velocity of the data generated and being made available increases, the Productivity Commission may have to give thought of processes that allow citizens to challenge the scientific data incorporated in agency rule making and permit decisions.

Data Resourcing Model and Data Lifecycle

As we are currently in the early phases of the big data lifecycle (for public use), current emphasis is on data capture, aggregation and linkage, and analysis to gain insights. Storage and maintenance issues, whilst a concern for organisations, is not a primary concern for consumers / citizens, given the proliferation of cloud based storage solutions. As governance issues for different stages of the data lifecycle vary, due consideration should be given to the governance requirements of the pertinent stage in the data lifecycle.

Goben and Raszewski (2015) provide an indicative list of questions across the various lifecycle stages.

Table 1 Life cycle stages and identified questions

Stages	**Questions**
Identifying	What data is available?
	What is the current audience for the data?
	What potential future audiences exist for the data?
	IS this an isolated data set or could it be combined with other sets?
Digitizing	Are the data in digital format?
	If no, what would it take to digitize the data?
	Is the data in a stable digital format that can be preserved?
Cleaning	How many people have touched or will touch the data?
	What rules have been created to ensure consistent data standardization?
	What tools can be used to standardize the data?
Describing	Is there a README.txt file outlining the project?
	Is there a standard ontology applicable to this data set?
	What information would others need to use the data?

Stages	Questions
Storing and preserving	What access is needed to work with the data now
	Who needs access now?
	What are the best storage options for the future?
	What is the intended duration of preservation?
	Are there any privacy concerns about the data?
Sharing	Who is the owner of this data set?
	What institutional policies apply to the data?
	How can sharing rights be maximized?
Analyzing	What analysis tools are available?
	What are the limitations of the data set?

Note this is an indicative list of life cycle stage and identified questions. For example Chisholm (2015) outlines the data lifecycle stages to include:

1. Data Capture;
2. Data Maintenance;
3. Data Synthesis;
4. Data Usage;
5. Data Publication;
6. Data Archival; and
7. Data Purging.

Given the plurality of data being considered in the Issues Paper, it is recommended that the Productivity Commission decide the most appropriate nomenclature and the associated issues for the data lifecycle phases.

In addition the Productivity Commission also to consider preparing guidelines regarding data quality and assurance of data quality for the data being made available. Note, any analysis from the data made available will only be as good as the data provided as input. Clarke (2015) provides a list of data and information quality factors as an indicative framework within which analysis can be conducted.

As the name of the Issues Paper suggest, the emphasis is on availability and use. Whilst Section 3 of the Issues Paper addresses some of the concerns associated with costs in the section titled 'The Resource Costs of Making Data Available', I would recommend that the draft report also consider end of life issues, such as data archival and purging. Any holistic discussion about data availability and use, will be incomplete without addressing issues and concerns around archival and purging of data, as they have different governance requirements. Furthermore, as the volume of data that is made (publicly) available expands exponentially, issues around data archiving and purging will be increasingly important.

For data purging, the key governance challenge will be to prove the purging has been completed correctly and all data (and copies) in question has been deleted. From a governance perspective, this raises the issues of how one can audit the absence of something? Hence clear guidelines will have to be provided by the Productivity Commission regarding data purging. Governance requirements around data archival (retention and storage) will be extremely varied given the inconsistency for data storage requirements both across firms and government regulations.

Furthermore, whilst funding is available from project inception to completion, given that life of data requirements may require that data to be stored after project completion, it is paramount that project proponents consider allocating some project costs to ongoing maintenance post project completion. Lack of funds may potentially compromise the quality and/or the availability of data post project completion. Hence the Productivity Commission to consider the duration of data availability. Additionally it may want to consider duration of data being made publicly available, based on frequency of use and / or quantum of use.

Cross Regulator Synergies

Whilst the issues paper talks about insufficient data sharing amongst agencies (refer Box 2, pg 8, Issues Paper), the Productivity Commission needs to go beyond data sharing and identify and leverage inter-agency synergies. Failing to leverage the cross-regulatory synergies and duplicating in many cases the data collection, handling and reporting efforts, can result in excessive cost. It is therefore crucial that the Productivity Commission explore opportunities for cross-regulatory synergies when preparing the draft report.

The Productivity Commission to consider conducting a regulatory mapping exercise to determine the current data captured by the various government agencies, to determine duplication and inconsistencies (e.g. data quality, formats, retention period, data purging, etc.), with the aim of identifying synergies, productivity gains in data collection and capture, and consistency across the various government agencies. The Productivity Commission to also consider appointing a coordinating body / organisation with an overarching responsibility to identify and streamline data duplication and overlapping requirements and inconsistencies in data capture, collection and retention requirements.

Open Data to Linked Open Data and the creation of usable knowledge

As we transition from open data to linked open data and real-time data with the IoT, a key consideration for the Productivity Commission is how does one make the data semantically aware, especially when considering the integration of heterogeneous data sources? Also of consideration will be what techniques needs to be developed to convert data to usable knowledge?

Whilst the current issue with data in some instances can be related to volume – i.e. the size of data generated by one institution/organization (e.g. the Marine National Facility can produce 1 TB of data/week, the CSIRO Australian Square Kilometre Array Pathfinder (ASKAP) Science Data Archive (CASDA) has the potential to generate 16 TB/day), it raises the question of how one downloads/transfers such large data sets. Hence for such large datasets, the trend is to provide in-situ data analytics. The Productivity Commission to consider the size of the data being made publicly available, i.e. downloadable and whether analytic services will be made available for very large datasets.

The data generated by the IoT raises the issue of volume and velocity of data generation, however not in the same manner as the examples above (e.g. individual sensor data may be small in size (byte – kilobyte) range, but when aggregated by the number of sensors and the velocity of generation (1sec interval), can become large (megabyte to gigabyte)). Coupled with the plurality of stakeholders that generate the data, lack of clarity of

access to and/or ownership/rights in the data supply chain (i.e. sensor manufacturer, the device/equipment manufacturer, the aggregator, etc.), data provenance issues and the proprietary nature of databases and administrative systems, results in the creation of barriers to the free flow of data. Hence some of the challenges for data interpretation and the formation of knowledge include addressing noisy, physical world data and developing new inference techniques that do not suffer the limitations of Bayesian or Dempster-Shafer schemes (Stankovic, 2014).

Given the plurality of stakeholders involved in the creation, the variety of data being generated, and use and publication of data, the Productivity Commission will need to address the issue of the different types of disciplinarities (refer Figure 3[3]) in making data available.

- Intradisciplinary: working within a single discipline.
- Crossdisciplinary: viewing one discipline from the perspective of another.
- Multidisciplinary: people from different disciplines working together, each drawing on their disciplinary knowledge.
- Interdisciplinary: integrating knowledge and methods from different disciplines, using a real synthesis of approaches.
- Transdisciplinary: creating a unity of intellectual frameworks beyond the disciplinary perspectives.

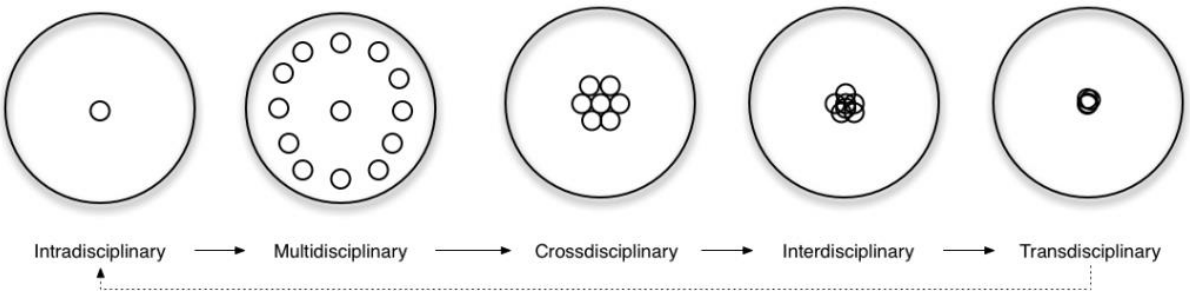

Figure 3 Different types of disciplinarities

Trust, Security and Privacy

Trust is one important aspect of the usefulness of big data, with security and privacy being essential elements of trust. In making data more readily available, care should be taken to not to equate simple data sharing with harmful misuse. Further, while ethical controversies exist in both the technology field and the social sciences; the nature of the issues, however, can be different. Hence the Productivity Commission to consider workshopping to identify some of the ethical issues around making data available with engineers, computer scientists, social scientists and various other stakeholders (government departments and agencies, private organisations and the community) and find possible solutions.

[3] http://www.arj.no/2012/03/12/disciplinarities-2/

Bibliography

Clarke, Roger (2013). Big Data's Big Unintended Consequences. Website: http://www.rogerclarke.com/DV/BigData-1303.html. Date Accessed: 25/07/2016

Clarke, Roger (2015). Big Data, Big Risks. Website: http://www.rogerclarke.com/EC/BDBR.html#BDQ. Date Accessed: 25/07/2016

Chisholm, Malcolm (2015). 7 Phases of A Data Life Cycle. Website: http://www.information-management.com/news/data-management/Data-Life-Cycle-Defined-10027232-1.html . Date Accessed: 25/07/2016

Dou, Dejing, Wang, Hao and Liu, Haishan (2015). Semantic Data Mining: A Survey of Ontology-based Approaches. Website: http://ix.cs.uoregon.edu/~dou/research/papers/icsc15_invited.pdf. Date Accessed: 25/07/2016

Kabmala, Malee, Manmart, Lampang and Chirathamjaree, Chaiyaporn (2006). An Ontology Based Approach To The Integration Of Heterogeneous Information Systems Supporting Integrated Provincial Administration In Khon Kaen, Thailand. Website: http://ro.ecu.edu.au/cgi/viewcontent.cgi?article=1081&context=ceducom. Date Accessed: 25/07/2016

Klein, Thomas (2015). Environment Climate Data Sweden – A Swedish research infrastructure. Website: http://studylib.net/doc/15380143/thomas-klein--ecds. Date Accessed: 25/07/2016

Narayanan, Ravishankar (2016). Leveraging Big Data for regulatory projects. Website: https://www.gtnews.com/articles/leveraging-big-data-for-regulatory-projects/. Date Accessed: 25/07/2016

Oboler, Andre, Welsh, Kristopher and Cruz, Lito (2012).The Danger of big data: Social Media as Computational Social Science. http://firstmonday.org/ojs/index.php/fm/article/view/3993/3269#author. Date Accessed: 25/07/2016

Impact of emerging regulations on the mobile payments industry. July 2016. Website: http://www.paymentscardsandmobile.com/impact-emerging-regulations-mobile-payments-industry/. Date Accessed: 25/07/2016