

APRIL 2016

# Open Data Supply: Enriching the usability of information



# About Phoensight

Phoensight is a data science consultancy group that specialises in employing contemporary analytics to support economic research and policy analysis.

Our diverse combination of skills gives us a unique perspective in gaining valuable insights for business problems. Using sophisticated model development and the latest visualisation frameworks, Phoensight provides a compelling narrative based on innovative solutions.



VISION WITH DATA

For more information about Phoensight or to get in touch, visit us at:

[www.phoensight.com](http://www.phoensight.com)



# Contents

Executive Summary .....	5
Introduction .....	7
Scope .....	9
Open Data Supply .....	11
Open Data Repositories .....	11
Open Data Accessibility .....	16
A Closer Look at Government Open Data .....	18
Estimating Government Open Data Supply .....	21
Measures of Data Usability .....	26
The Four Pillars .....	28
Quantifying Open Data Usability .....	30
The Scoring Criteria .....	31
The Open Data Usability Index .....	32
Open Data Usability Standards .....	36
Industry Impacts .....	39
Case Study: Data-Driven Innovation in Australia .....	40
Future Possibilities .....	44
Conclusion .....	46
Technical Methodology .....	48
Authors .....	52
Bibliography .....	53
Terms and Conditions .....	54

Copyright © 2016 Phoensight  
All rights reserved. This report may not be reproduced or redistributed, in whole or in part, without the written permission of Phoensight and Phoensight accepts no liability whatsoever for the actions of third parties in this respect.





# Executive Summary

With the emergence of increasing computational power, high cloud storage capacity and big data comes an eager anticipation of one of the biggest IT transformations of our society today.

Open data has an instrumental role to play in our digital revolution by creating unprecedented opportunities for governments and businesses to leverage off previously unavailable information to strengthen their analytics and decision making for new client experiences.

Whilst virtually every business recognises the value of data and the importance of the analytics built on it, the ability to realise the potential for maximising revenue and cost savings is not straightforward. The discovery of valuable insights often involves the acquisition of new data and an understanding of it. As we move towards an increasing supply of open data, technological and other entrepreneurs will look to better utilise government information for improved productivity.

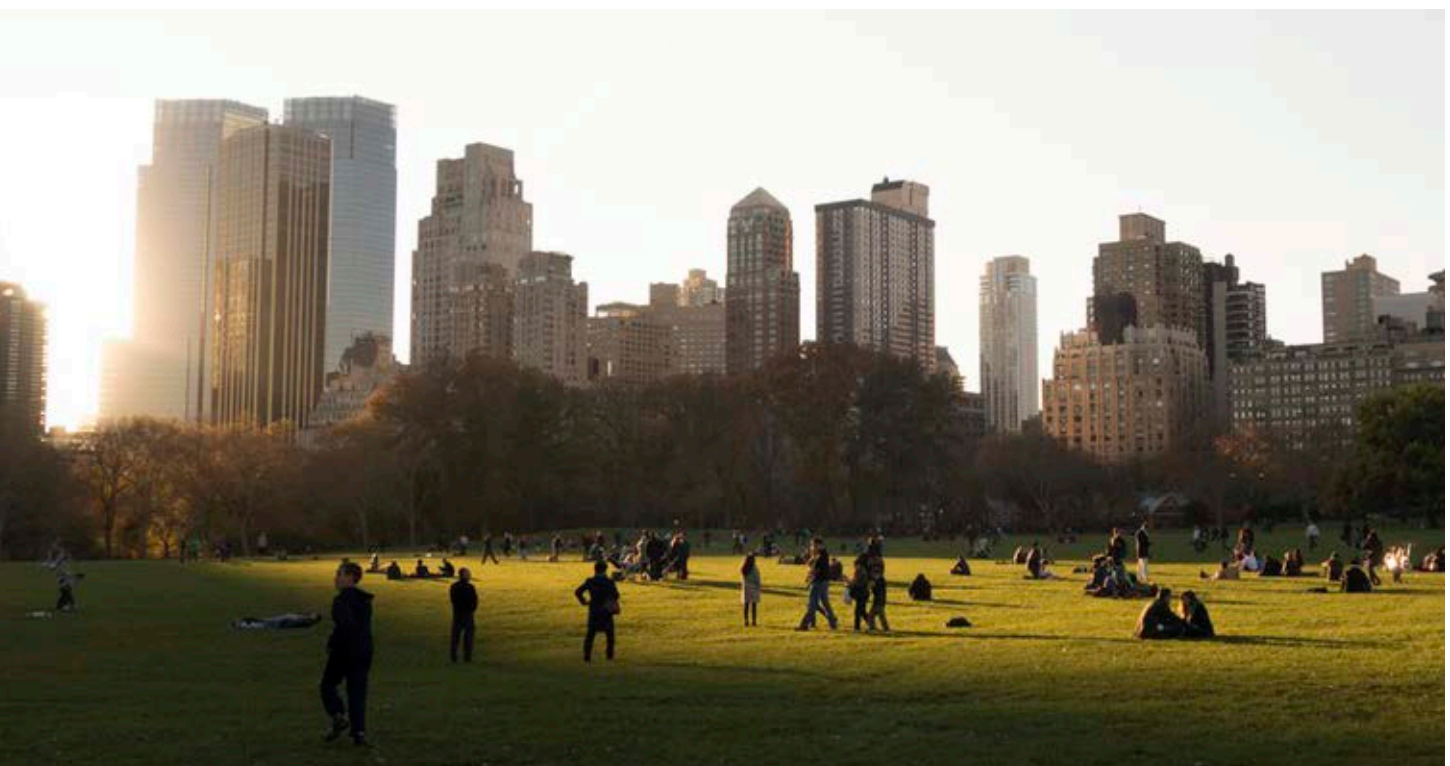
This report uses a data-centric approach to examine the *usability* of information by considering ways in which open data could better facilitate data-driven innovations and further boost our economy. It assesses the state of open data today and suggests ways in which data providers could supply open data to optimise its use. A number of useful measures of information usability such as accessibility, quantity, quality and openness are presented which together contribute to the *Open Data Usability Index (ODUI)*. For the first time, a comprehensive assessment of open data usability has been developed and is expected to be a critical step in taking the open data agenda to the next level.

With over two million government datasets assessed against the open data usability framework and models developed to link entire country's datasets to key industry sectors, never before has such an extensive analysis been undertaken. Government open data across Australia, Canada, Singapore, the United Kingdom and the United States reveal that most countries have the capacity for improvements in their information usability. It was found that for 2015 the United

Kingdom led the way followed by Canada, Singapore, the United States and Australia. The global potential of government open data is expected to reach 20 exabytes by 2020, provided governments are able to release as much data as possible within legislative constraints.

It is important that both Government and Industry are acutely aware of the state of open data supplies to help close the gap between the provision of data and its applications. With Singapore's new government open data platform being a shining example of how enhancing user experience can be an enabler of innovation, the future of open data platforms will play a significant role in ideas creation across a range of industries. Industry impacts for Australia based on 2013 data found that although the Health Care & Social Assistance and Mining Sectors were well placed to utilise the full potential of open data, the data supplies for these could be optimised further. It is also observed that the Financial and Insurance Services sector is well placed to respond to an approaching digital transformation.

It is hoped that the findings in this report will inform the agenda for improving the usability of open data and contribute to best practices across both the public and private sectors. Phoensight intends to update the analysis and the ODUI for each country annually, expanding this work to include additional countries, various levels of government and different government agencies for future reports.





# Introduction

With countries around the world becoming increasingly committed to creating unprecedented levels of openness in the interests of transparency, public collaboration and effectiveness in Government, comes the releasing of previously unpublished information in the public domain.

Governments around the world are endeavouring to facilitate the global digital economy by releasing data to the public so that public-private partnerships and other entrepreneurs are able to use government information to further boost productivity and growth.

Global initiatives to make government “open by default” were formalised in 2015 with the adoption of the *International Open Data Charter*. The Open Government Partnership (OGP), an international organisation with approximately 69 participating countries, aims to promote the expansion and growth of open data by supporting the infrastructure and resources required, and increasing the awareness of open government data issues.

*Open data* is free, public data that is freely available without restrictions, and may be used commercially and for non-profit ventures to develop new products and services. It creates new opportunities to innovate and adds value to business models by enriching their datasets, strengthening their decision making and improving their customer services. In considering the benefits of releasing open data, it is important to also consider balancing potential issues around protecting the privacy and autonomy of individuals and other entities.

Government publication of open data has the potential to unlock large amounts of economic value, with the McKinsey Global Institute estimating this impact in 2013 to be approximately \$3 trillion USD in additional value annually. However, for this value to be realised to its full capacity, governments need to assess their open data policies and the usability of the open data that’s available. This report examines the supply of government open data today and the issues faced by users of this data, both from a technical perspective as well as how this would translate into the wider economy.

## Principles of the International Open Data Charter

The International Open Data Charter was formally adopted by seventeen governments of countries, states and cities at the 'Open Government Partnership Global Summit' in Mexico in October 2015. The main principles are outlined as follows:

### 1. Open by Default

Governments are encouraged to open their data by default, recognising that intellectual property, personally-identifiable and sensitive information need to be protected.

### 2. Timely and Comprehensive

Release data that is comprehensive, accurate and high quality, acknowledging that it may require time and resources to prioritise data for release and / or improvement.

### 3. Accessible and Usable

Release data in open formats wherever possible, allowing for free access with minimal barriers to entry.

### 4. Comparable and Interoperable

Improve data usability and presentation, and support interoperability, traceability and effective data reuse.

### 5. For Improved Governance and Citizen Engagement

Strengthen governance through increased transparency and accountability thereby improving decision-making and the provision of public services and citizen engagement.

### 6. For Inclusive Development and Innovation

Encourage creativity and innovation by allowing access to the data and unlocking economic value by providing the tools and resources to understand and use data effectively.





## Scope

Open data has a number of benefits to society including increased transparency, accountability and promoting good governance within Government.

This report is primarily concerned with the economic benefits of open data from a data supply perspective, and how these may be optimised by improving the usability of open data. Although the concepts discussed in this report may be applied to all types of open data, the results presented focus on *Government Open Data*, and apply to data owned by all levels of government including national, federal, state and local bodies.

Whilst the number of countries adopting open data initiatives are steadily increasing, this report focuses on government open data from Australia, Canada, Singapore, the United Kingdom and the United States. Moreover, the results presented are based on analysis performed on government open data on the *Comprehensive Knowledge Archive Network* (CKAN) management system. This report recognises that although increasing government open data could contribute significantly to economic output, governments also have a need to protect the privacy of their citizens and other entities by providing the required data anonymity to achieve this.





# Open Data Supply



# Open Data Supply

For open data to have a significant impact in the economy, government reforms need to focus on the effectiveness and efficiencies in the open data market.

Recent studies on open data (Davies 2014) have suggested a mismatch of open data supply and demand, stating that released datasets are often not those that are most in demand and that counting datasets is a poor way of assessing the quality of an open data initiative. This section examines the supply of government open data for our five selected countries and provides a discussion on their current data composition and future expected trends.

## Open Data Repositories

The International Open Data Charter aims to publish government data on a national portal so that released data may be found in one location.

The data portal may be a central website from which data can be downloaded, or a website listing all open government data stored at a different location. Ideally, the data portal will include a registry listing all the data and metadata as well as providing Application Programming Interfaces (API<sup>1</sup>) for easy accessibility.

Data portals today are evolving to being platforms where not only can data be accessed, but also a place where open data users can discuss uses for the data, develop applications and build communities. For data-driven innovation to thrive, users need to be able to view data casually, analyse their data and link datasets to discover new business insights. Although the infrastructure and development of technical platforms and open data portals are critical for the provision of open data, they require the support of government open data reforms to ensure that the data available is useful and actively used.

---

<sup>1</sup> APIs let a system provide a simpler set of instructions or requests that can be used in other applications. The responses to those requests are provided in a way that's consistent in terms of content, structure and delivery. Using them means a developer doesn't need to know how to talk to the complex parts of the system and can rely on the result being the same each time. On the system side they mean that developers can change the underlying infrastructure but still provide the same means to ask for data and the same responses to those requests.



**There are many data portals available for open data suppliers today, some of which include:**

#### **CKAN**

Maintained by the 'Open Knowledge Foundation', CKAN is the leading open-source data portal platform for governments around the world making data accessible by providing tools to streamline publishing, sharing, discovering and using data.

#### **Socrata**

A platform providing software solutions designed for digital government, Socrata turns data into a utility that can be discovered, consumed, visualised, analysed and shared.

#### **Junar**

A cloud-based open data platform that is fast, cost effective, focused on powerful analysis and data visualisation, and is easy to use.

#### **OpenDataSoft**

A turnkey platform which makes it easy to publish data, share it as interactive data visualisations and reuse them via automatically-generated APIs.

#### **Accela**

A platform that offers software that streamlines land, permitting, asset, licensing, right-of-way, legislative management, and resource and recreation management.













Singapore's new government open data platform, currently in beta, is an example of breaking down the barrier to entry for idea generation and innovation.

Key visual representations within categories and of specific data sets provides a means for a broader audience to explore and engage with open data.

Complimenting datasets with such an attractive design and ample visual representation helps remove barriers and drive innovation. Focusing on user experience, the Singaporean CKAN reveals data insights and possibilities to non-developers, enabling them to more simply explore and interrogate government data.

Singapore's CKAN also makes very effective use of default metadata fields which further assists with discoverability of datasets. Having "groups" consistently filled for example, helps ensure that when searching on topics, users get the full picture of available datasets in their area of interest.

Singapore is a leader in this space and will only improve further as more of their data is transitioned from their existing open data repository to their beta CKAN repository.

# Open Data Accessibility

Whilst open data repositories go a long way to addressing the issue of simplifying the discovery of data, data accessibility also plays a fundamental role in driving the usability and usage of data.

For data to be truly accessible, it is also important to consider the speed and reliability in accessing the data, as well as issues around data licensing and the conditions under which the data can be accessed or used. In a digital landscape where data proliferation is becoming ubiquitous within government agencies, users of open data also face challenges in addressing requirements for greater transparency and accountability in data-driven approaches, and it is expected that data provenance will become increasingly prominent.

Metadata<sup>2</sup> is critically important in enabling data accessibility by enhancing discoverability and allowing users to determine data suitability without needing to download and inspect the data itself. A robust data repository with comprehensive metadata should assist with ensuring that when data sets are moved online, developers can quickly point their applications to the new locations. Whilst metadata is useful for cataloguing data, it is incredibly important in managing a large body of data and assisting with linking datasets for analytics.

Metadata does not need to be stored with the data itself, however it must be accessible via an API. Where possible, the data itself should also be accessible via an API. Without an API, the process of extracting and ingesting the data becomes both cumbersome and expensive to maintain and considerably impacts the accessibility of the data. Further, if the API is poorly designed or limited in scope or implementation, undue cost is placed on developers, which increases the barrier to accessing open data. While API access is critical for developers, APIs can be complex for non-technical users to understand or utilise. This is where simplified APIs or API request builders can be of great benefit and add to data usability.

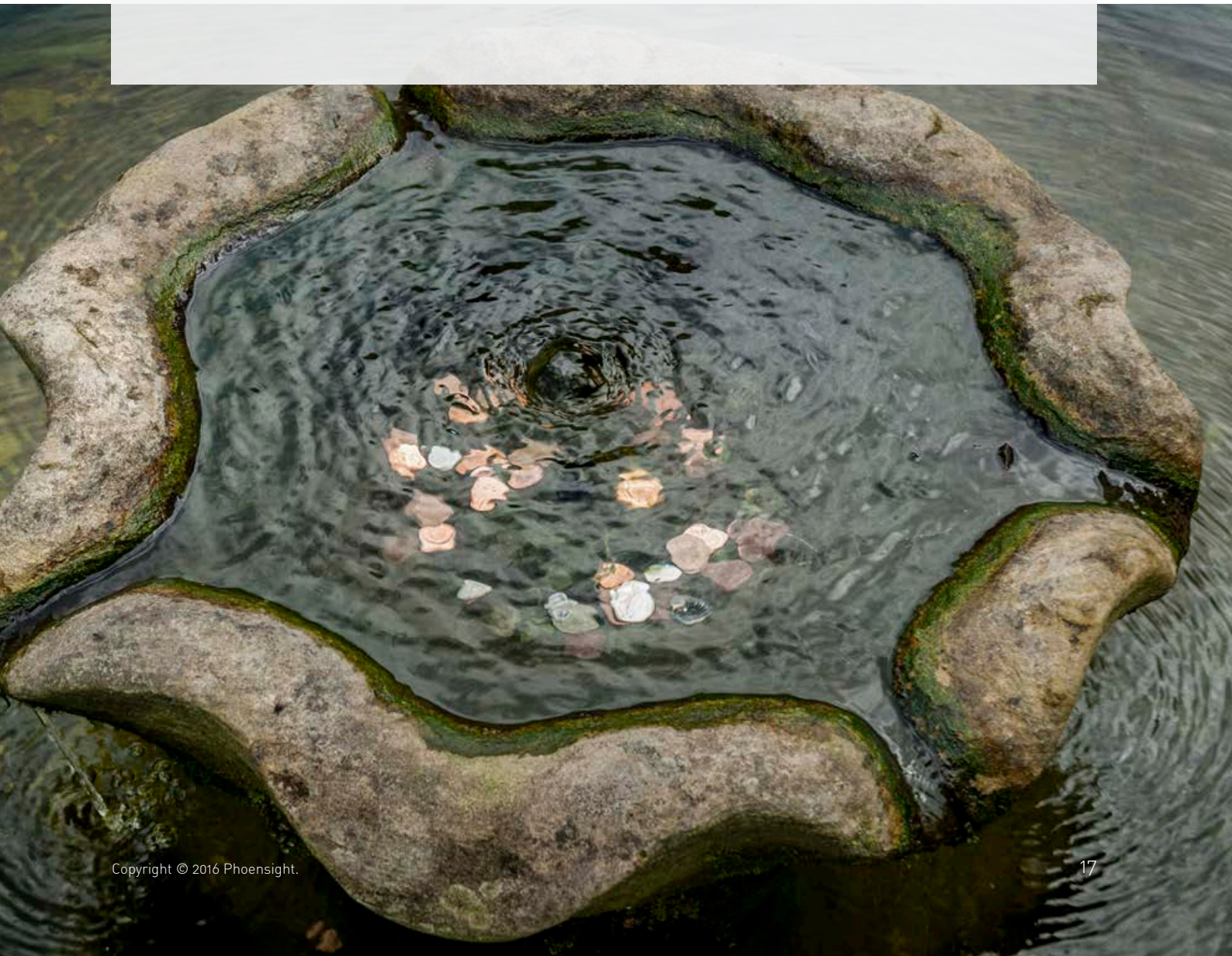
---

<sup>2</sup> Information describing the details about the data such as its content, format, quality, when it was created, the license it is provided under, the owner and whether the data is considered active.



The movement to API enabled data continues to do an excellent job in making it simpler for developers to access government data for their applications.

Further, it assists application development through removing needs for more complex data ingestion, cleansing and transformation. However not enough has been done to apply the same principle to those lacking developer skills. While widely recognised that cross disciplinary groups and those with strong business expertise can lead to impressive idea generation, the barrier to entry for those without developer skills remains high, even with machine readable and API enabled data sets. Simplified API builders such as the trial underway on the UK Governments Open Data platform take great steps in introducing and enabling API requests to non-technical users.





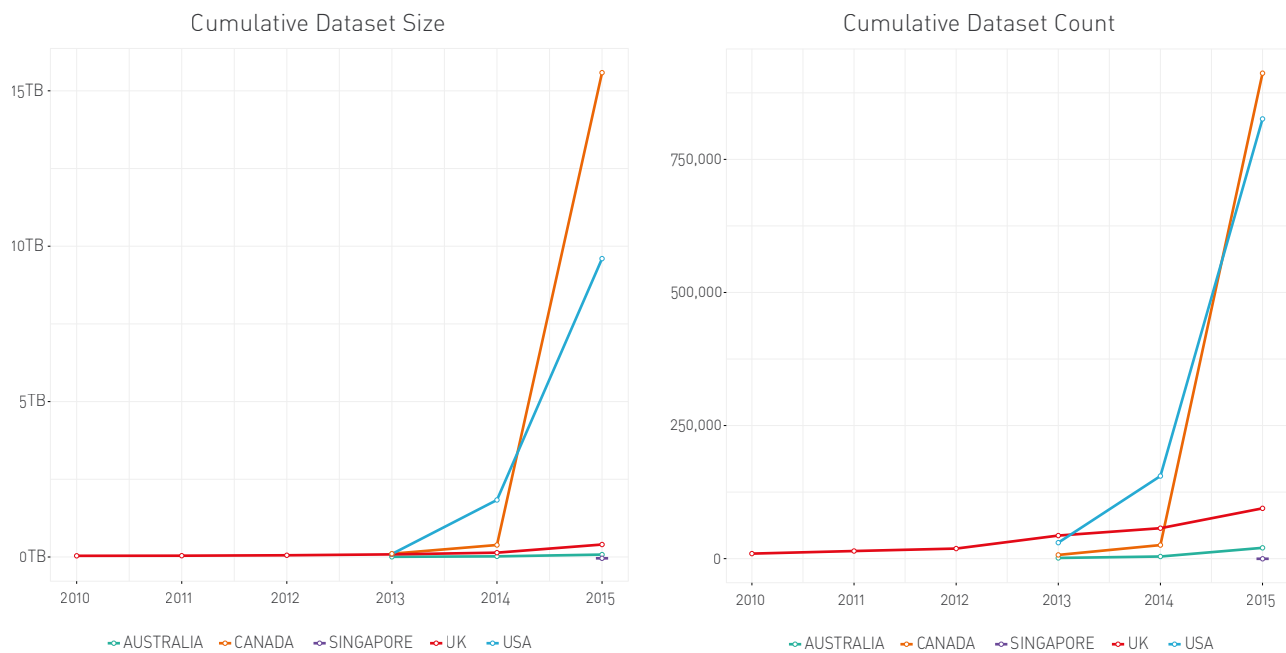
# A Closer Look at Government Open Data

With the launch of the ‘Open Government Partnership’ in 2011, countries around the world have been gathering momentum in a global movement to open up public data that governments collect and generate.

For open data policy to be effective, measuring data release progress is essential and provides feedback to policy makers in addressing any potential issues. Reviewing open data performance and the quality of publication of key datasets is a key recommendation put forward by the *Open Data Institute*. More recently in 2014, the *G-20’s Anti-corruption Working Group (ACWG)* identified that the quality, quantity and content of government open data be broadly examined in an attempt to measure a country’s transparency and as a deterrent to corruption.

The ‘Open Data Index’ is a measure of the state of open government data around the world based on a crowdsourced survey designed to assess the openness of specific government datasets. However, few studies directly examine the open data on the repositories themselves. This section examines the supply of government open data *directly* to assess the current state of the data and how it has changed over time.

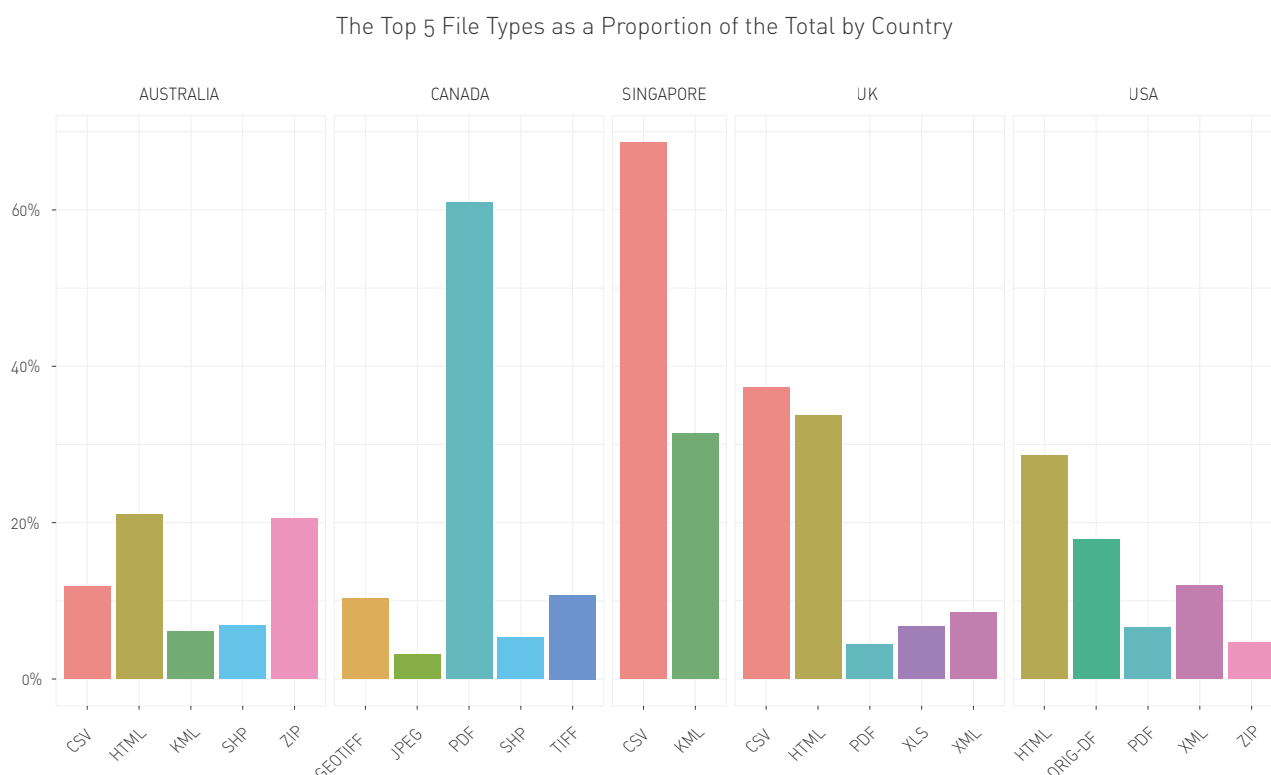
Figure A on the following page shows that government open data across all five countries has initially been released in small quantities, followed by periods of rapid growth since 2014.



**Figure A: The volume of open data sets by counts and size from 2010 until 2015 for each of the five countries**

Canada and the USA have typified this having accelerated the inclusion of data within their repositories during 2014 and 2015. Based on estimates from the CKAN repository, the total file size for Singapore, the UK and Australia are 240MB, 400GB and 80GB and their file counts are at 319, 94,490 and 20,309 respectively. Canada and the USA are significantly higher with file sizes at 15.5TB and 9.6TB respectively, whilst their file counts are at 911,890 and 825,998 files in 2015. The gap between Canada and the USA on file size is significantly higher than would be expected for the difference in the number of files. This is a result of Canada releasing larger files on average than other repositories including the USA during that time period. Moreover, much of this volume of data tends to be spatially related.

Across the USA and Canada, a small number of organizations contribute the bulk of this data, such as the National Oceanic and Atmospheric Administration in the USA and Natural Resources Canada in Canada. For example, Natural Resources Canada contributed over 800,000 files in 2015, over 95 per cent of all files added in the Canadian repository during 2015. While the bulk of these files were PDFs, they were focused on geographic information, and were predominantly images of topographic maps.



**Figure B: The percentage of the top five most common files for each country**

As seen from Figure B, file types across the repositories show similar themes across the five countries, with Shape files, KML and PDFs tending to be the common file types released. Singapore, being in a beta stage of their CKAN repository are currently limited to just two file types, CSV and KML. As both file types are non-proprietary, structured formats, this bodes well for Singapore's long term performance on openness across the repository. However, Singapore currently has a relatively large number of files that are not under an open license.

Generally, Canada seems to have a relatively high proportion of unstructured and proprietary formats. The UK however tends to have a higher proportion of structured and open formats. The USA is more difficult to measure as the repository metadata proved less reliable in this regard. However, it does generally have a higher proportion of unstructured formats. Australia has a high proportion of zip files and falls

somewhere in the middle when looking across structured versus unstructured and proprietary versus open.

With the bulk of data files across data repositories either not structured (such as images of tables or maps), in proprietary formats, or links to web pages that provide further links to data, there is great potential for improvement. While providing API access through repositories such as CKAN has been a major improvement in the accessibility of data, if this is only enabling access to metadata or files that are not easily ingested into data driven applications, then the gap between providing more accessible open data and being able to use it is still significant. Nevertheless, the recent growth in government open data is a promising indication that countries are becoming actively involved in facilitating the supply of open data.



# Estimating Government Open Data Supply

The World Bank (2014) has recommended that Government Institutions need to make publicly visible more details of their overall data holdings, including those datasets not yet available as open data.

Although countries are becoming increasingly committed to initiatives around government transparency, there is almost no information in the current literature as to *how much* data countries should release in order to achieve a certain level of transparency. Whilst it is difficult to assess whether a country has released as much open data as possible<sup>3</sup>, it is nevertheless useful in estimating this value to provide an indication of the extent of the probable government open data reserves that may be tapped into in the future.

The volume of government open data depends on a number of factors such as the resources available to countries to gather and set up open data, governance and administration costs, and the potential size or volume of the data itself. This section aims to quantify the volume of government open data in an idealistic environment where transparency is paramount, administrative costs to governments are minimal and in the absence of future digital revolutions. The components that contribute to the volume of this data production are examined in light of the variety and distribution of open data available today.

In seeking to quantify the volume of open data that countries should release, it is first useful to consider how the volume of data should be measured. Davies (2014) notes that assessing the quality of an open data initiative by counting the datasets released by a country is a poor measure but quite common. Counts of datasets are not a wholly accurate measure of the quantity of data as datasets are often duplicated within repositories to provide access to varying file formats or in some cases, to provide different levels of aggregations of the same data<sup>4</sup>. While there is some benefit to be gained from providing the same data in different file formats, providing different aggregates may be questionable in terms of data usability.

---

<sup>3</sup> Subject to the legislative requirements around privacy protection.

<sup>4</sup> Data from a particular organisation within one of the repositories was found to have a very large number of duplicate data files that were the same format, but just different basic summaries of the same data. These files were small, but accounted for over 95 per cent of the total count for that organisation, boosting their count of files with questionable gains in information or usability.

One way to address this issue is to determine the size<sup>5</sup> of the open data files released by a country. With a cumulative file size across a repository, it is much easier to avoid any misrepresentation that a large number of small files contribute significantly to the volume of data. However, using file size alone as a measure of data volume is not sufficient. Some file types are generally larger than others without providing additional information. For example, a PDF file format with tabular data or an image of tabular data is significantly larger than a CSV file format and in this instance, file count may provide a more accurate view of the overall volume of data. A simplistic approach in estimating the amount of government open data that should be released is to consider the trends of data volumes by file type.

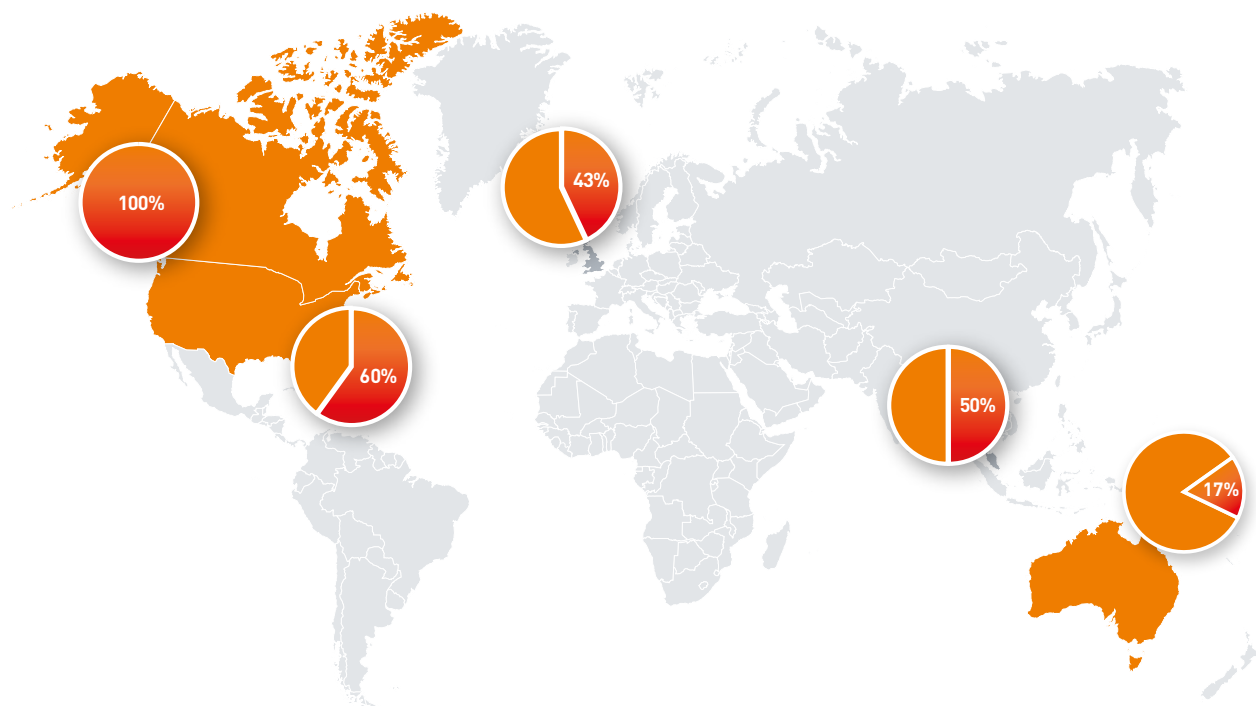
As seen previously, spatial data make up a large proportion of government open data and is generally different in size to other structured file types, particularly when it comes to data size. Given the current standards in technology, data resolution and available geographic information, it is reasonable to expect that the amounts of spatial open data released by a country should be proportional to the landmass of that country. Thus it is possible to estimate the amount of spatial open data by country relative to spatial data released by a country with an active open data culture, such as the UK.

Estimating non-spatial data is much more complex due to the nature of the sort of data collected and how it relates to the economic, industrial and societal activity on a country level. At a first glance, it is reasonable to expect countries with larger populations to collect more government open data due to the increased transactional and infrastructural services provided. Further to that argument, larger populations would also allow a greater level of disaggregation of data, since individual entities might be harder to identify in keeping with the privacy protection requirements that are legislated. However, a large proportion of this data would not be expected to increase with population size with weather data being one example.

Using a similar approach to spatial data, the amount of non-spatial data for each country can be roughly estimated to be a proportion of that country's Gross Domestic Product (GDP) since it would control somewhat for the economic activity within each country. Using trend analysis, we can look to a government with a mature open data policy such as the UK, to estimate their total data file size to GDP ratio in 2020 when it is expected that the early release influx of new previously unpublished government open data has stabilised.

---

<sup>5</sup> File size is provided for a wide range of datasets within CKAN repositories. However these sizes were often inconsistent with the sizes of the files calculated through direct processing. Closer examination suggests that in some circumstances the CKAN repository is inaccurate which may be due to updates to files not modifying the underlying size metadata.



**Figure C: Proportion of released government open data to the estimated potential for Australia, Canada, Singapore, the UK and the USA.**

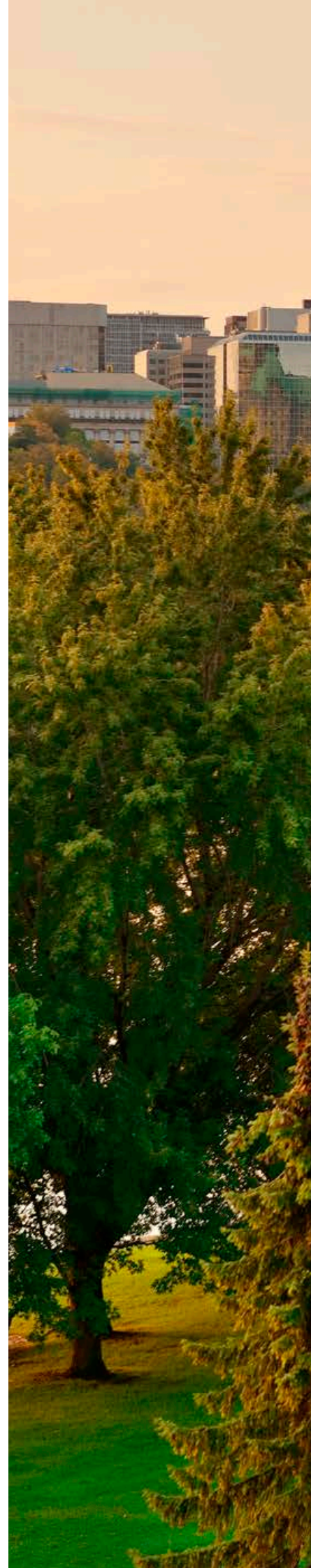
Figure C shows the quantity proportion of released government open data to what could be released by a country if they had open data policies conducive to transparency and openness and had released all possible government open data sets.

Canada is shown at 100 per cent due to its disproportionately high volume of data relative to its GDP and landmass when compared with the UK as a benchmark. Although, both Canada and the USA have gone a large way towards releasing what might be expected from them in terms of data quantity, the file type distribution presented in Figure B earlier suggests that it may be predominantly due to their high proportion of proprietary format files which generally have large file sizes. What this suggests, is that without consistency in the file type distributions, quantity alone may not be a sufficient measure of data usability. It is expected that as these countries move towards more machine readable formats, their aggregate data quantity will be moderated to reflect this.



The UK is tracking well at 43 per cent with its government open data release initiative. Given the UK's file type distribution from Figure B it is expected that they will achieve their optimal open data quantity by 2020. Australia, only having released 17 per cent of the data expected, is clearly an early adopter of government open data policies and has much more scope to realise its data potential. While Singapore is included in Figure C, the volume of data on its beta CKAN repository falls below the threshold for countries for which reliable estimates can be calculated. It appears as 50 per cent as the low number of files results in artificially high estimates of the total spatial data released.


Using a similar approach, the total file size for government open data world-wide may be attempted for the first time. Based on estimates of global landmass and GDP, an estimate of total government open data could potentially reach 20 exabytes by 2020 using file size and number of files across Government open data repositories examined in this research.





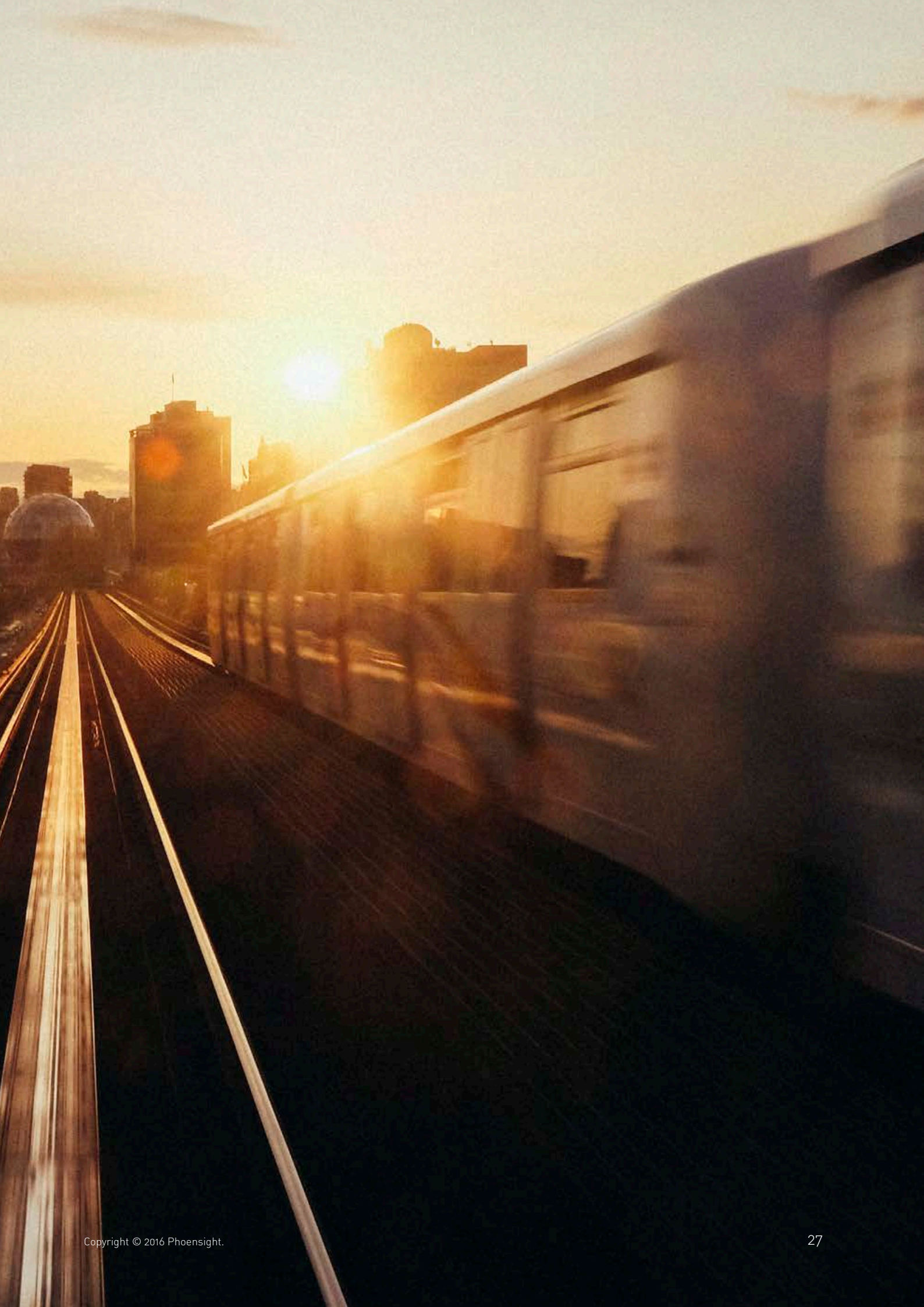




A photograph of a cityscape at sunset. In the foreground, a multi-lane highway with a guardrail runs diagonally from the bottom right towards the center. To the left of the highway is a dark, modern building with a flat roof. In the background, several construction cranes are visible against the orange and yellow sky, along with various city buildings. The overall scene is bathed in the warm light of the setting sun.

# Measures of Data Usability







# The Four Pillars



## **Accessibility**

Accessibility is broadly a measure of how easy it is to access the data both directly and as a result of the implementation of the data repository. Data must be accessible through using a high quality repository with metadata that is complete and useful. Further it must have a well-documented, functional API and the underlying data must be quickly and reliably obtainable.



## **Quantity**

Data quantity is measured both in terms of the number of files or data sets released as well as the size of those files. Quantity is measured at a repository level where the size and count of files must meet a reasonable percentage of the projected volume of government open data expected for that country.



## **Quality**

Better data quality makes data easier to work with and enables more reliable analyses and applications to be built developed. Quality can only be measured by inspecting the data directly. There are many measures that can be applied to ascertain whether a file is of reasonable quality. These include whether the data variable names are meaningful, the data appears complete or has a large ratio of missing data, and if there is any data corruption.



## **Openness**

Openness of data is measured through adopting Tim Berners-Lee's "5-Star deployment scheme". This scheme looks at the licensing of the data, whether it is structured, whether it is released in proprietary or open formats, and whether it conforms to W3C standards with or without links to provide context around the data. The five star rating is simple to implement and allows a quick evaluation of the openness of data across a repository.



# Quantifying Open Data Usability

In order to attempt quantifying the usability of data, it is important to consider each of the 'four pillars' presented earlier to gain an understanding of how they fit together to describe the usability of open data.

Each pillar, accessibility, quantity, quality and openness constitute the different components of information usability and may provide valuable insights into the 'readiness' of the open data supply to be used for analytics. In determining an open data usability score for each country, scores for each pillar based on a number of criteria shown in table A, were calculated using the data, metadata and information about the repositories on which they are hosted.



# The Scoring Criteria

In estimating a measure of open data usability, it is important to examine the performance of government open data for each country against the four pillars. For this reason, the following criteria have been developed against which the data may be assessed and scored.

**Table A: Usability measure scoring criteria**

Accessibility	Quantity	Quality	Openness
			
<b>1 Star: ★</b> Uses a repository that is current with a well-defined API.	<b>1 Star: ★</b> The repository has both spatial and other general files.	<b>1 Star: ★</b> The file contains structured data.	<b>1 Star: ★</b> The data is available under an open license.
<b>1 Star: ★</b> Data can be accessed at a reasonable download speed.	<b>1 Star: ★</b> The repository has a sufficient number of general files relative to GDP.	<b>1 Star: ★</b> The file does not contain unprintable characters.	<b>2 Star: ★ ★</b> The data is structured.
<b>1 Star: ★</b> The repository has a low proportion of broken links.	<b>1 Star: ★</b> The repository has a sufficient number of spatial files relative to landmass.	<b>1 Star: ★</b> The number of meaningless column names is low.	<b>3 Star: ★ ★ ★</b> The data is in a non-proprietary format.
<b>1 Star: ★</b> The repository has a low proportion of missing data in key fields.	<b>1 Star: ★</b> The repository has a sufficient cumulative size of general data.	<b>1 Star: ★</b> The number of meaningless column names is zero.	<b>4 Star: ★ ★ ★ ★</b> The data is published using open standards from the W3C.
<b>1 Star: ★</b> The repository provides a simplified API or advanced visualization of a significant proportion of datasets.	<b>1 Star: ★</b> The repository has a sufficient cumulative size of spatial data.	<b>1 Star: ★</b> The amount of missing data is low.	<b>5 Star: ★ ★ ★ ★ ★</b> The data is published using open standards from the W3C and linked to other data.

# The Open Data Usability Index

Once countries with government open data are assessed against the four measures of open data usability and given a final usability score, it is instructive to compare these scores across countries and examine how they have evolved over time. For this purpose, an *Open Data Usability Index* has been developed for the first time.

The *Open Data Usability Index (ODUI)* is a measure of the usability of a country's open data relative to the score for the UK score in a base year. The UK was chosen as the base country mainly because it is a recognised world leader in open data and has publicly strived to address all four usability measures. For the purposes of this report, the relative base year has been chosen as 2015, so the ODUI in a particular year (t), for a specific country (C) is given by:

$$\text{ODUI}_t(C) = \frac{\text{ODUS}_t(C)}{\text{ODUS}_{2015}(\text{UK})} ;$$

where

$$\text{ODUS}_t(C) = \text{Accessibility}_t(C) + \text{Quantity}_t(C) + \text{Quality}_t(C) + \text{Openness}_t(C).$$

Using the equations above, the ODUI were calculated for each of the five countries and compared in an attempt to understand their data usability performance and identify potential strengths and areas that could be improved.

**Table B: The 2015 usability measures and the Open Data Usability Index (ODUI) by country**

Country	Accessibility	Quantity	Quality	Openness	Usability Score	ODUI
AUSTRALIA	3	2	4.07	1.32	10.39	0.87
CANADA	2	5	3.39	1.26	11.65	0.98
SINGAPORE	3	3	5	0.49	11.49	0.96
UK	4	3	3.14	1.79	11.93	1
USA	3	4	3.41	0.67	11.08	0.93

Table B shows the usability measure scores and the ODUI by country in 2015. The UK takes the lead when it comes to the overall data usability score with consistently good performances across all the measures, topping accessibility and openness. A four point score for accessibility is achieved through a CKAN repository that goes well beyond the default settings to deliver a platform that seeks to make discovery and access of Government open data simple for both developers and non-technical citizens. With a high quantity and reasonable openness score, Canada places second in 2015 with an ODUI of 0.98, followed closely by Singapore at 0.96.

Singapore scores highly on the overall index, however this is primarily due to Singapore providing a relatively small number of high quality files. The Singaporean CKAN repository is currently in beta and had a very limited number of files compared to their previous repository which was not API enabled as CKAN is. Singapore falls below a threshold of data quantity that would be reliably scored by the methodology outlined. It is included for the reason that Singapore is both of regional significance and has delivered thus far, a leading implementation of CKAN. Further, Singapore's previous repository has a substantially large volume of data relative to both its GDP and landmass. It is expected that Singapore will score strongly on the ODUI in coming years as the volume of data on the new repository grows.



Another interesting observation is that the USA has a large number of files that are not released under an open license<sup>6</sup> and as such, although their quantity score is relatively high, their mean openness score is low in comparison.

Australia scores reasonably well on accessibility and has a high data quality measure, resulting in an ODUI of 0.87. Australia’s CKAN repository could be improved through adopting similar features to the UK and Singapore in terms of a simplified API and improved user experience. Australia, being comparatively new to open data and having recently committed to participating

in the international *Open Government Partnership*, has a lot more to gain in terms of releasing a larger volume of data with scores on quantity relative to GDP and landmass being the lowest of the five countries.

Figures D and E show the four usability measures scored for 2015 for each country. They demonstrate that the openness measure is where countries tend to score lowest. Interestingly, although it is expected that the UK would be relatively well balanced, Australia’s usability measure profile is also approximately symmetric.

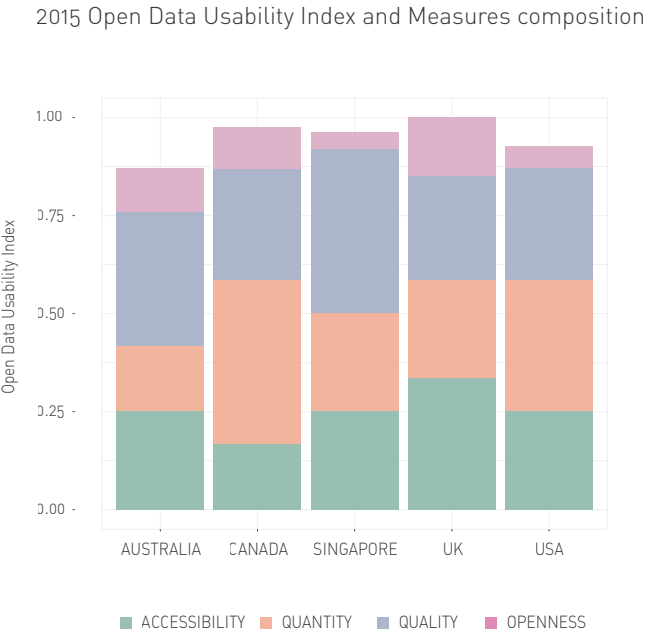


Figure D: Stacked bar chart based on 2015 Open Data Usability Index by country

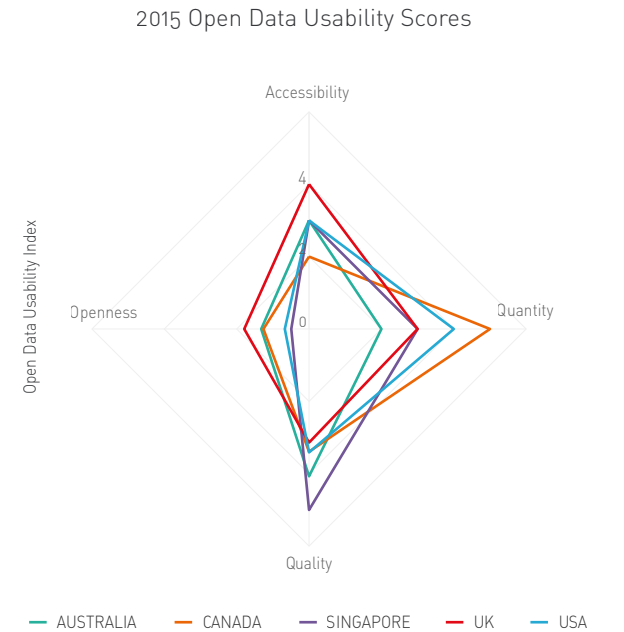
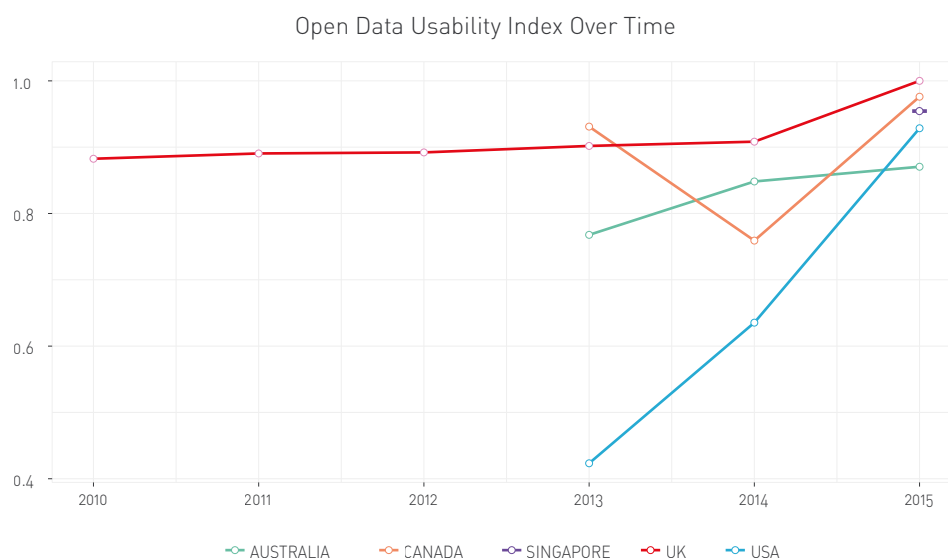


Figure E: Polar chart based on 2015 Open Data Usability Scores by country

<sup>6</sup> A large volume of closed files can suggest a high level of transparency in Government. Noting that openness refers to the open nature of the files themselves, the concept of transparency refers to whether the Government acknowledges the existence of the data. Providing details on these files with a contact point that enables Freedom of Information (FOI) requests would be considered an indicator of good transparency.



**Figure F: The Open Data Usability Index by country 2010 – 2015**

Figure F shows the ODUI for the five countries from 2010 until 2015 with the UK achieving the highest index score for the usability of open data in 2015. Whilst the ODUI for the UK is relatively stable with a larger growth rate between 2014 and 2015, the ODUI for the USA has the greatest growth for the period 2013 until 2015. Australia's ODUI also increases in this period, but it is interesting to note that Canada's ODUI drops in 2014 before increasing again in 2015 mainly due to a drop in the average quality and openness of files in that year.



# Open Data Usability Standards

The Open Data Usability Index and measures demonstrated that there are a number of factors that data suppliers should be concerned with. The steps required to deliver usable open data need not be onerous or overly costly.

At a minimum data must be:

- stored on a central repository that contains or links to the datasets;
- enriched with metadata that includes licensing, the author, time stamps on when the data was created and updated, details on the data itself and where to access it;
- provided in structured non-proprietary formats where possible such as CSV and KML files;
- where possible, not encapsulated in other file formats such as ZIP archives;
- provided in sufficient quantity; and
- supported with an API to query the metadata and the data.

Most of the steps above concord with the principles of the *International Open Data Charter* and are supported by the evidence based on a data-driven approach used in this research for a usable government open data repository.

Of note here is that the metadata itself should be as complete as possible. Care should be taken to ensure that not only are fields filled, but they are filled accurately. Spelling errors or inconsistencies in metadata fields can mean the difference in developers or other users being able to find the data at all. Discoverability is essential for analysis and data driven applications to take place effectively.

In addition to providing data in open structured formats, it is highly desirable to provide data directly<sup>7</sup> via APIs, enabling developers to query data directly for their needs rather than needing to process the entire dataset after requesting it through a repository API. Further, the API should be consistently applied. This research encountered examples of repositories whose API implementations did not match the documentation, failed to operate as expected within some open source languages' supporting packages and had inconsistent syntax across API calls that were similar in function<sup>8</sup>. Consistency and clarity are key to ensuring a functional API that assists developers to innovate, experiment and create with data.

---

<sup>7</sup> Preferably in a serialised format.

<sup>8</sup> Such as filtering search results.









# Industry Impacts

# Industry Impacts

Open data has the potential to significantly impact different industry sectors by creating value in the economy through innovation, enabling improved decision making and providing more efficient service delivery.

The McKinsey Global Institute (2013) estimated that open data enables \$3 trillion USD potential annual value across seven industry domains: education; transportation; consumer products; electricity; oil and gas; health care; and consumer finance.

A recent example of this trend is the banking industry in the UK, with the Open Banking Working Group (OBWG) established in September 2015 to create an *Open Banking Standard*, which would allow for improved competition and efficiency in the sector. The OBWG has recommended that open APIs should be created to enable services to be built using bank and customer data, including open data about products and shared data that entities may choose to share through secure means. The 'Open Banking Project', initiated by TESOBE, is an open source API and App store for banks that allow financial institutions to securely allow third party developers to build applications and services based on account holders' transaction data.

With the FinTech industry leading the front on early adoption of new technology and open data, it is expected that the Open Banking Standard will set a precedent for the development of similar international standards across other sectors. An increasing uptake of open data and data-driven innovation in our global technological landscape will play an important role in boosting productivity and contributing to economic growth across most of our industry sectors in the near future.

Government has an important role to play in encouraging the emergence of open information architectures through open data policy reforms, which may also help facilitate the development of open standards within industries. Khan and Foti (2015) note that there is some evidence that sector-specific approaches to open data might be more conducive to higher rates of implementation than a whole-of-government approach. Effective policy outcomes could potentially draw out valuable private data into the public domain, thereby adding to the data commons and lifting productivity and improving efficiencies for industry and society. The global economy is on the cusp of a digital disruption, which could revolutionise client services by reconfiguring the business model and dynamics of how industries operate.



# Case Study: Data-Driven Innovation in Australia

Digital technology impacts all industry sectors in the Australian economy and according to Deloitte Access Economics' 'Australia's digital pulse' report (2016), is one of the fastest growing sectors forecasted to rise to \$139 billion AUD or 7 per cent of total Australian GDP by 2020.

PwC (2014) estimated that data-driven innovation added approximately \$67 billion AUD in new value to the Australian economy in 2013, which was about 4.4 percent of Australia's gross domestic product in that year. Australian open data policies together with an enriched open data supply that's been optimised for maximum usability are essential components in supporting data-driven innovation.

Lateral Economics (2014) estimates that stronger open data policies in Australia could add around \$16 billion AUD per annum to the Australian Economy. The Australian Government now requires agencies to make data open by default in a machine readable format and has recently released one of its most requested high-value datasets, the Geocoded National Address File (G-NAF) to promote innovation in various sectors. The Open Data 500 Australia surveyed different sectors and organisations in Australia and have identified the most requested data themes as being spatial and land, socio-economic, health, transport and environment.

From an open data market perspective, it is instructive to examine whether the Australian government open data supply is able to meet industry demand. The 'Public Sector data management' report (2015) released by the Commonwealth of Australia, notes that Australia's national data portal Data.gov.au links to approximately 6,700 datasets, of which 75 per cent are from four organisations and less than 25 percent enabled by APIs. Datasets may be used for purposes different from what they were originally intended and opportunities to repurpose data will become increasingly common as information is more usable, better organised and understood.



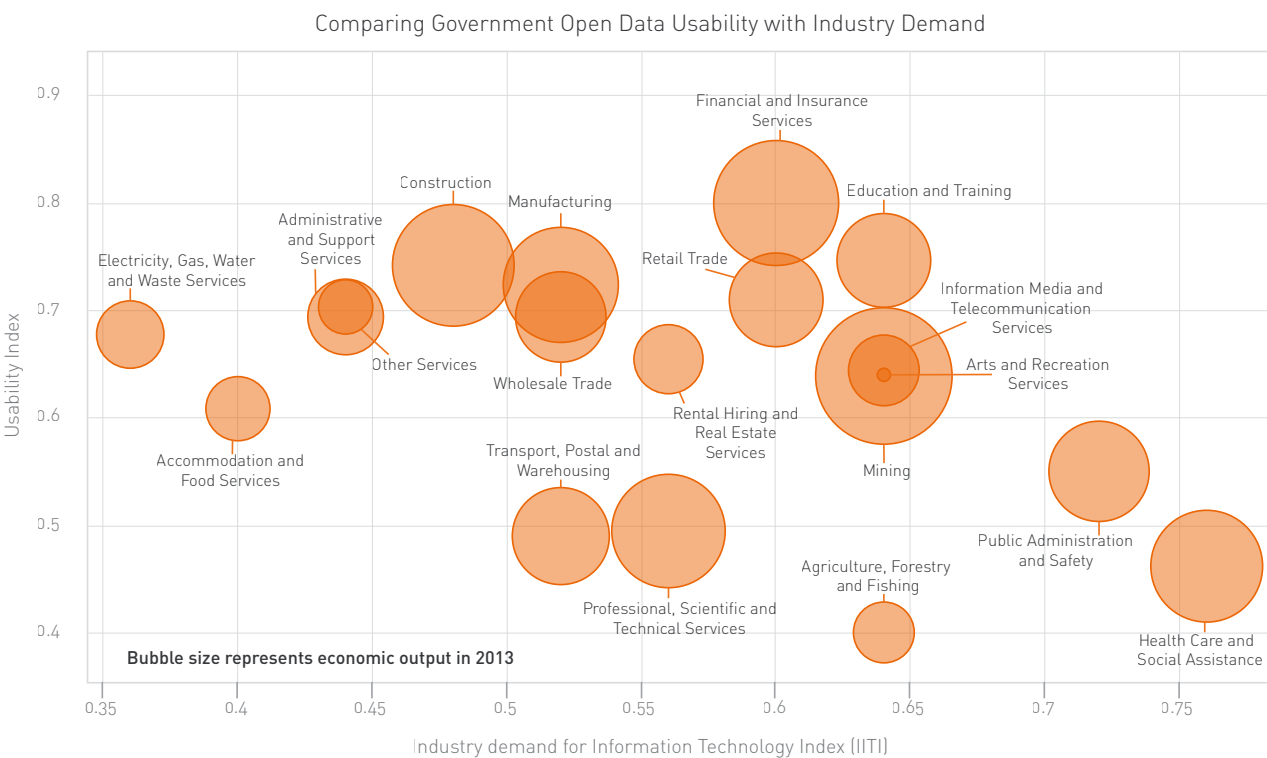
**Figure G: Proportion of total open data in Australia by Industry in 2015**

Figure G presents the proportion of Australian government open data that may be attributed to the different industry sectors across the country. Whilst the proportion of open data relating to each of the industry sectors varies between 3 per cent and 10 per cent, Professional Scientific and technical services followed by Agriculture, Forestry and Fishing, and Public Administration and Safety lead the other Australian industry sectors in terms of being able to access open data that may pertain to their needs.



The Australian Bureau of Statistics (ABS) looks at the business use of information technology by industry as collected annually by the Business Characteristics Survey (BCS)<sup>9</sup>. Using this information, we can gain an indication of how ‘ready’ an industry sector is for digital disruption and this provides a sense of the demand for open data. An Industry demand for Information Technology Index (IITI) was created for the purpose of representing the demand for government open data based on the ABS survey results.

In contrast, the *Open Data Usability Index (ODUI)* presented in this report goes towards indicating the extent of open data supply and its usability in the current data market. In other words, the ODUI may be seen as representing the current state of open data, from which industry will endeavour to extract value from.



**Figure H: Open Data Usability Index against the Industry demand for Information Technology Index (IITI) by Industry Sector for Australia in 2013**

Figure H provides a useful snapshot of how the different industry sectors in Australia are positioned in an open data market. It is interesting to note that although the Health Care and Social Assistance and Mining sectors are well placed to translate data-driven innovation into potential economic value, their low ODUI suggests that open data supplies pertaining to these industries have scope for improvement that could future boost economic activity.

<sup>9</sup> Catalogue no. 8167

Another observation is that the Financial and Insurance Services sector with a relatively high ODUI is well placed to respond to a digital disruption. Indeed, recent events in Australia's FinTech sector suggest that the industry is already reshaping its approach to financial services in a rapidly changing technological environment.

Industries with a low ODUI and a high readiness to consume open data are most likely to gain economically from improvements in the usability of open data relevant to that industry. The implications of these findings support the argument that it may be in the sector's interest to partner with government in ensuring the release of specific open datasets, which contribute to the economic growth of those industries. Furthermore, the nature of open data allows for new local and global entrants into the different industry sectors, which will increase competition, improve services and contribute to global trade.

# Future Possibilities

As traditional Government open data begins to level off in terms of release, new sources of data will be identified.

The bulk of current data is now primarily structured and unstructured spatial data in addition to financial reporting and broad statistical measures. As both openness and transparency grow and an emphasis on the value of releasing data is embedded across government, less traditional forms of data will likely be released.

Reporting for example may move from unstructured, proprietary files to embedded information on the web. Aside from making access to and analysis of the text of the report a simple exercise, other elements such as images may also be included and referenced within open data repositories with metadata for searching. In addition, data sources for embedded visualisations in reports could be sourced via APIs making the entire report a combination of multiple datasets.

Taking this further, government models themselves may be released<sup>10</sup>. The output of which can also be included in reports via API access rather than embedding unstructured text or images of results. Releasing reports in this manner would enable industry and the public to both review the report and provide their own interpretation using underlying data. Further it would dramatically improve the discoverability of Government reporting on areas of interest as search results across repositories could include metadata and content from the report, metadata on included images, metadata on data sources for visualisations and models, and metadata on the models themselves.

As Government recognises the value added to the economy through the release of data, it may choose to price the value of data in collaboration with Industry. This could be done to both assess its current production value and to determine where to focus efforts on collecting, collating and releasing new data. In order to translate the benefits of open data into the economy, data analytics and data science capabilities will play an increasingly critical role in driving innovation.

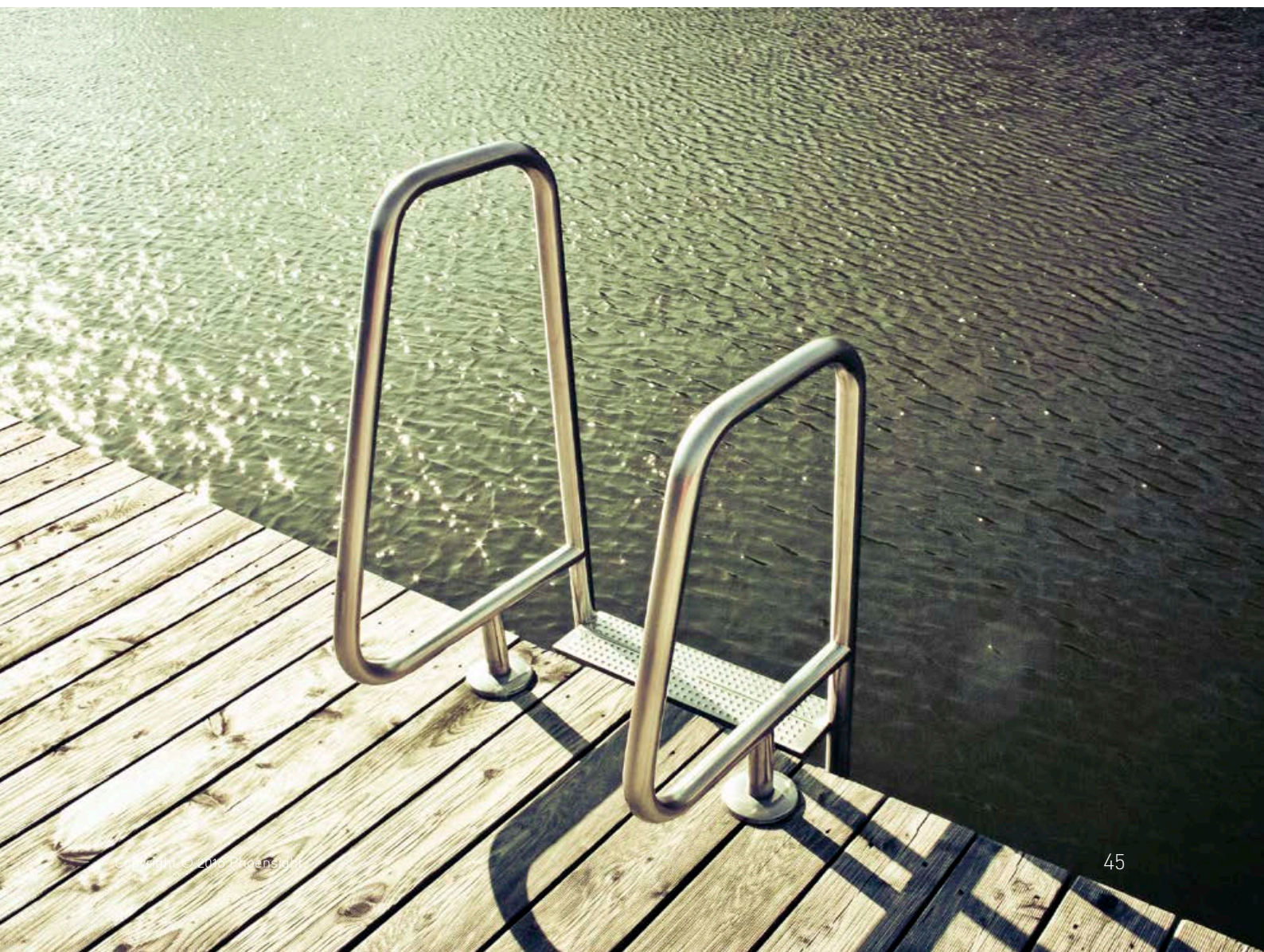
---

<sup>10</sup> (Lobo-Pulo 2015) <http://phoensight.com/evaluating-government-policies-using-open-source-models/>



More work can be done in providing a means to request closed data via Freedom of Information (FOI). Some countries such as the UK have details within their repositories on FOI contacts. This is an excellent start and could be improved upon through working towards more automated approaches to requesting data. While it would be a boon for open data for Governments to adopt more transparent measures and disclose all data sources publicly, it would be of even greater benefit to enable the requests for this data in a simple manner. Modelling could be adopted to learn and determine whether FOI requests should be approved and actioned, with results returned immediately to the requester.

Finally, Government could also seek to release data under collaborative or joint initiatives. These initiatives could leverage existing industry partnerships or be delivered through strategic investment in new data driven partnerships across a range of sectors.





# Conclusion

Government Open Data has a significant role to play in contributing to the global digital evolution today.

For open data to be an enabler for innovation and efficiencies in our economy, not only do countries have to release more data but data usability needs to be optimised for industry to fully realise its potential. Data usability depends on four key components: accessibility, quantity, quality and openness. Together these measures are able to provide valuable insights for data suppliers into how their data could be more useful. Moreover, the *Open Data Usability index (ODUI)* is presented for longitudinal and international comparisons. The ODUI will continue to track the performance of government open data going forward and report on the progress in a growing number of countries. The total amount of potential government open data globally has been estimated for the first time and is expected to be around 20 exabytes. With data-driven approaches becoming increasingly important in our technological landscape, public-private partnerships will feature more prominently in a collaborative approach to drive open data supply.







# Technical Methodology

Metadata was extracted on 2,305,183 files and their broader packages during January 2016 for the five countries of interest.

A collection of 649 fields, including custom fields, was extracted from the repositories. Descriptive statistics such as size, counts and distributions were generated on the metadata to determine features of interest. Analysis suggested that there were 94 unique important fields that would contribute to further analysis on the repositories.



## Accessibility

The accessibility measure used in this report, primarily relates to the state of the data repository itself. Repositories were assessed to determine whether they provided a well-defined API and a suite of features that makes accessing the data straight forward for both technical and non-technical citizens. A small number of other countries were assessed that opted for bespoke solutions that provided neither an API nor a well-documented system in which to access their data.

Access speed of the repository and the websites linked within the metadata that provided the underlying data were assessed to see whether they met reasonable benchmark figures.

A reliable connection was used and repeated benchmarking of connection and download speed were introduced to block for source connection issues impacting on the results. These tests were repeated many times to improve estimates and generate statistically valid results. During this process, results on broken links or links to files that did not match their metadata on CKAN were also collected and incorporated into the accessibility scores.

The ratio of missing data across key default fields in the CKAN repositories including, but not limited to licensing, URLs and authors was calculated to further assess the broad quality of usage of the data repository. Scores were then calculated by comparing the ratio of missing data with a benchmark threshold.

Results across the accessibility criteria were aggregated to generate an overall accessibility score. As this score looks at elements that cannot be measured for past performance, they are required to be calculated annually to track the ongoing performance for each country. For the purposes of this research, current scoring has been rolled back to each prior year.



## Quantity

Assessing the quantity of data required a broad sample of files across the repositories to ensure that all structured and non-structured data files were represented. Using stratified random sampling, tens of thousands of data files were extracted and processed to achieve a statistically valid sample for file types across each country per year with 95 per cent

confidence and less than 5 per cent margin of error. Both Spatial and other file types were segmented with multiple measures taken including size and other metadata. Ratios were calculated to compare to benchmark quantity measures for all repositories. Spatial files for each country were compared to their landmass, while non-spatial files were compared to the country's GDP. The mean file size per year was calculated to ensure an accurate representation of the changes in overall data size and therefore provide a more accurate picture of cumulative sizes. Using the projected 2020 benchmarks for the size and number of both spatial and other file types, scores were calculated based on countries achieving a specified proportion of this estimate for 2015.



### Quality

Quality is scored through assessing files linked on and stored within the CKAN repositories. Like Quantity, stratified random sampling was used to obtain a statistically valid sample. Analysis was focused on assessing open, structured files<sup>11</sup>. Structured data is considered of high usability and its quality is a good indicator of the overall usability of the data across a repository.

The criteria used to assess file quality include the proportion of meaningless column names, the amount of missing data and the proportion of unprintable characters. The scoring methodology for quality included an

assessment of the data files against each of these criteria with reasonable thresholds. These scores were then aggregated and a mean quality score was determined for each country per year.



### Openness

To calculate Openness scores, the Tim Berners-Lee Openness 5-Star deployment scheme was applied based on the types of data stored or referenced in the repositories. To simplify distinguishing between RDF and linked RDF these ratings were combined into a single five star rating. This had a small impact on the overall ratings, as the number of four and five star files was relatively low across the repositories. The results were aggregated across the datasets, generating a mean and trimmed mean star rating for the openness measure. The main issue presented by this approach was that countries that published more closed government data than others tended to receive a lower openness rating since they were assessed based on openness rather than transparency.

### Industries

Industries for Australian data are not included consistently or purposefully across the Australian CKAN repository. As such, multi-label classification models were developed to categorise the data in an automated manner. The models were shown to perform well on training, testing and scoring data.

---

<sup>11</sup> Other file types such as PDFs, images, spatially related and other non-structured files may be assessed in future.

## Full Open Data Usability Index details for 2010 – 2015

Table C shows the Open Data Usability Index results for 2010 to 2015 across the five countries examined in this research.

**Table C: 2010 - 2015 Open Data Usability Index scores**

Country	Year	Quantity	Accessibility	Quality	Openness	Usability Score	ODUI
AUSTRALIA	2013	3	1	3.86	1.30	9.16	0.77
AUSTRALIA	2014	3	1	4.38	1.74	10.12	0.85
AUSTRALIA	2015	3	2	4.07	1.32	10.39	0.87
CANADA	2013	2	2	4.64	2.47	11.11	0.93
CANADA	2014	2	2	3.5	1.56	9.06	0.76
CANADA	2015	2	5	3.39	1.26	11.65	0.98
SINGAPORE	2015	3	3	5	0.49	11.49	0.96
UK	2010	4	1	3.02	2.51	10.53	0.88
UK	2011	4	1	3.25	2.38	10.63	0.89
UK	2012	4	1	3.21	2.43	10.64	0.89
UK	2013	4	2	3.17	1.59	10.76	0.90
UK	2014	4	2	3.2	1.64	10.84	0.91
UK	2015	4	3	3.14	1.79	11.93	1
USA	2013	3	1	-	1.05	5.05	0.42
USA	2014	3	1	3.28	0.30	7.58	0.64
USA	2015	3	4	3.41	0.67	11.08	0.93





30  
31  
32  
33  
34

29

30

31

32

33

34

# Authors



Dr. Audrey Lobo-Pulo

Sydney



Peter Phillips

Sydney



Dr. Graham Williams

Singapore



Luke Singham

Sydney

# Bibliography

- Australian Government, *Australian Government Public Data Policy Statement*, December 7, 2015.
- Bonina, C. M., *New business models and the value of open data definitions, challenges and opportunities*, August 2013.
- CapGemini Consulting, *The Open Data Economy Unlocking Economic Value by Opening Government and Public Data*, 2013.
- Commonwealth of Australia, *Public Sector Data Management*, July 2015.
- Davis, T., *Open Data in Developing Countries - Emerging Insights from Phase I*, The World Wide Web Foundation, July 2014.
- Deloitte, *Open growth Stimulating demand for open data in the UK*, 2012.
- Deloitte Access Economics, *Australia's Digital Pulse*, March 2016.
- G20, *G-20 Anti-corruption Open Data Principles*, November 2015.
- Gigler, S., S. Custer and H. Rahemtulla, *Realizing the vision of open government data: Opportunities, challenges and pitfalls*, The World Bank, 2011.
- HM Government, *Open Data White Paper Unleashing the Potential*, June 2012.
- Khan S. and Foti J., *Aligning Supply and Demand for Better Governance – Open Data in the Open Government Partnership*, Open Government Partnership, May 2015.
- Lateral Economics and Omidyar Network, *Open for business: How open data can help achieve the G20 growth target*, June 2014.
- Lobo-Pulo A., *Evaluating Government Policies using Open Source Models*, September 2015.
- McKinsey & Company, *Open data: Unlocking innovation and performance with liquid information*, October 2013.
- OECD, *Government at a Glance 2015*, July 6, 2015.
- OECD, *Data-Driven innovation Big Data for Growth and Well-Being*, 2015.
- Open Government Working Group, *Eight principles of open government data*, [opengovdata.org](http://opengovdata.org)
- Ordnance Survey, *Assessing the Value of OS OpenData to the Economy of Great Britain – Synopsis*, June 2013.
- Ponte, Diego, *Enabling an Open Data Ecosystem*, Association for Information Systems, 2015.
- PricewaterhouseCoopers, *Deciding with data How data-driven innovation is fuelling Australia's economic growth*, September 2014.
- Scott, A. and Bolotin, L., *Introducing the Open Banking Standard*, Open Data Institute, February 2016.
- The World Bank, *Open data for economic growth*, June 25, 2014.
- The World Wide Web Foundation, *Open Data Barometer – Second Edition*, January 2015.

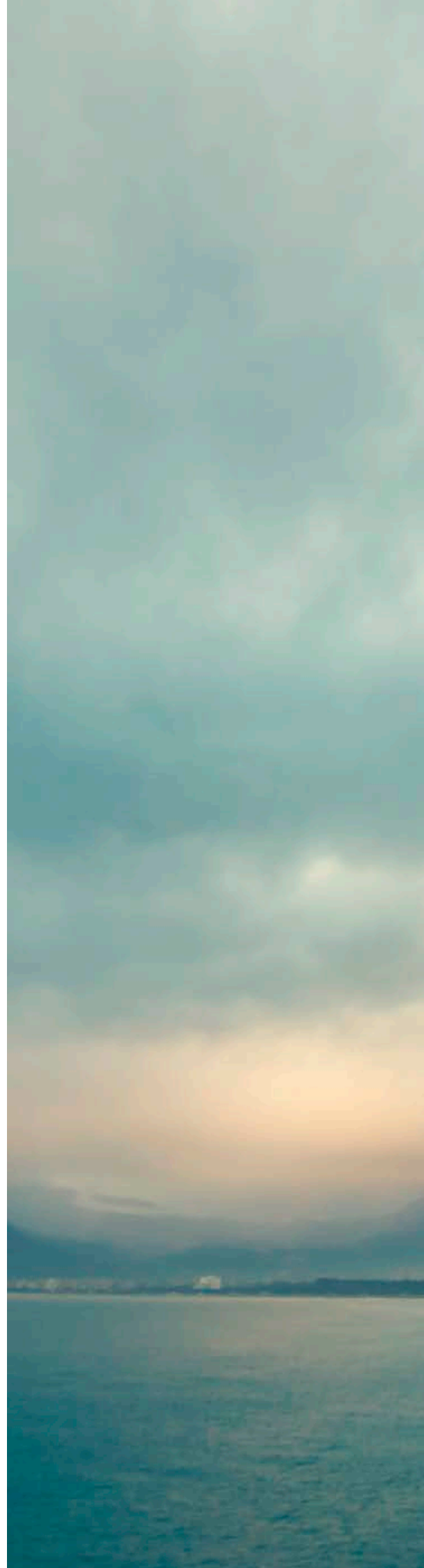


# Term and Conditions

This report has been prepared for general guidance on the subject of interest only, and does not constitute professional advice. Phoensight accepts no duty of care or liability for any loss occasioned to any person acting or refraining from action as a result of any material in this publication.

This report cannot be relied on to cover specific situations; application of the principles set out will depend on the particular circumstances involved and we recommend that professional advice is obtained before acting or refraining from acting on any of the contents of this report. No representation or warranty (express or implied) is given as to the accuracy or completeness of the information contained in this report; and Phoensight reserves the right to alter the information provided in this report at any time.

Phoensight would be pleased to advise readers on how to apply the principles set out in this report to their specific circumstances.





---

## Open Data Supply: Enriching the usability of information



---

Phoensight is a data science consultancy group that specialises in employing contemporary analytics to support economic research and policy analysis. Our diverse combination of skills gives us a unique perspective in gaining valuable insights for business problems. Using sophisticated model development and the latest visualisation frameworks, Phoensight provides a compelling narrative based on innovative solutions.

For more information about Phoensight or to get in touch, visit us at:

[www.phoensight.com](http://www.phoensight.com)

Copyright © 2016 Phoensight

All rights reserved. This report may not be reproduced or redistributed, in whole or in part, without the written permission of Phoensight and Phoensight accepts no liability whatsoever for the actions of third parties in this respect.