Literature Review
—Open Data in New Zealand—

Jimmy Oh

Department of Statistics
University of Auckland

May 29, 2013

Abstract

This Literature Review provides an Overview of Open Data in New Zealand by first defining Open Data, including key desirable properties and terminology relevant to the discussion, followed by an Overview of New Zealand State Sector sources. The State Sector organisations covered are those listed as *Public Service departments* on the State Services Commission's website (`http://www.ssc.govt.nz/state_sector_organisations`).

Contents

1 Introduction

In recent times there has been a movement toward *Open Data*, particularly for government data[1,2] (which also falls under the purview of *Open Government*). The reasons cited for these are typically intangible[3,4,5], but it is easy to appreciate how such benefits can arise. Of course, Open Data includes more than just government data, such as data arising from academic research (which also falls under the purview of *Open Research*), but the vast amount of data openly available from governments around the world make it an ideal low-hanging fruit for initial inquiry.

This Literature Review aims to provide an Overview of Open Data in New Zealand giving particular focus to New Zealand State Sector sources. The intent of this review is not to evaluate the performance of the New Zealand government and its organisations' release of data, but instead to understand the current situation such that we may better grasp what potential barriers exist that may hinder the utilisation of Open Data. In the process we become entangled in some policy issues and at times give thoughts on what would be ideal from our perspective; however such comments are intended primarily to aid in the understanding of the current situation, rather than evaluative remarks on current practice.

To do this we will:

1. Define Open Data (Section 2), including key desirable properties and terminology relevant to the discussion. In doing so we must delve into some policy and procedure issues, as they are relevant to understanding what makes for *Good* Open Data.

2. Provide an Overview of Open Data in New Zealand State Sector sources (Section 3). The State Sector organisations covered are those listed as *Public Service departments* on the State Services Commission's website[6], which includes many of the organisations the general public will be aware of, such as the Ministry of Health or Inland Revenue.

While effort has been made to keep this review up-to-date with recent changes, due to the nature of the subject many details become out-dated at an alarming rate. Many of the points made in the various summaries should remain valid, but the particulars may no longer apply at the time of reading.

Post-Script

This Literature Review is conducted as part of a thesis examining *Presentation Methods for Open Data* with the broad aim of making the data more accessible and usable for a wider audience. The focus on New Zealand State Sector data sources is to narrow the scope of the initial research. It is anticipated that eventually the thesis will move on to international data sources, at which point a supplementary Literature Review may be conducted.

[1] "In many countries across the world, discussions, policies and developments are actively emerging around open access to government data." Davies and Bawa (2012a)

[2] "Over 100 OGD [Open Government Data] initiatives are active across the globe, ranging from community-led OGD projects in urban India, to a World Bank sponsored OGD programme in Kenya, government-led developments in Brazil, civil-society initiated work in Russia, and a World Wide Web Foundation supported programme in Ghana." Davies and Bawa (2012b)

[3] "We promote open knowledge because of its potential to deliver far-reaching societal benefits which include... better governance... culture... research... economy" Open Knowledge Foundation (2012)

[4] "One of the pillars of open and transparent government is open government data and information." New Zealand Government ICT (2012)

[5] "The online publication of structured datasets by governments is seen as playing an important role in driving the transparency and accountability of states, enabling new forms of civic participation and action, and stimulating economic growth and development." Davies and Bawa (2012b)

[6] `http://www.ssc.govt.nz/state_sector_organisations`

2 What is Open Data?

2.1 A Definition

We all vaguely understand what is meant by Open Data, but let us be more formal with our definition. The Open Knowledge Foundation (2012) defines *Open Knowledge* to be "content that people are free to use, re-use and distribute without legal, technological or social restrictions." Davies and Bawa (2012b) observes that:

> "Open data" is just one of a number of high-profile labels with the prefix "open". Open government, open access, open innovation, open education and open knowledge are some of the other initiatives and movements in this area. Many of these draw from the emergence of "open source" as the inspiration for their development.

Cole (2012) notes that there are degrees of openness. For instance, data may only be available on request, or only reveal a subset or an aggregated form of the original data. Additionally, even if the data itself is 'open', the format or content of the data may require additional resources, such as expert knowledge or significant computing power, to process or understand the data. This presents barriers around accessibility and usability, limiting the potential benefits of making data open.

For our purposes the key features of Open Data are:

Free to use and re-use At a minimum this means there are no legal restrictions on using the data for any purpose. Ideally, it also means there are no non-legal restrictions, such as the data format being obfuscated, either intentionally or unintentionally, such that it is unsuitable for use other than for its 'intended' purpose.

Free to distribute Ideally this means we are free to distribute the data itself, as well as any derived output of the data. Sometimes the case may be that the data is freely available and we are free to distribute derived output (so long as it is sufficiently aggregated or processed to no longer be considered raw data), but we are restricted from re-distributing the data itself[7]. In this case, while not ideal, we would still consider this to be (mostly) Open Data.

Freely available Meaning the data is readily accessible, ideally on the internet, with little to no requirements of authentication or approval. Sometimes the case may be that registration and login is required to access the data. If this process is easy and incurs no cost, while not ideal, we would still consider this to be (mostly) Open Data. Other times the data may be 'freely available on request', but with some manner of lengthy approval and response process (often requiring human involvement on the provider side). In such cases we consider this to be a significant barrier for the wider public, and hence we would not consider this to be sufficiently Open to be considered Open Data.

2.2 Users of Open Data

We will classify Users of Open Data into the following categories:

Casual Users The 'ordinary person'. Generally these users will only use data *as is* with little to no independent effort to further process the data. They will appreciate data being presented in a human-friendly fashion, such as hierarchical tables that make it easy to locate a specific value, and will value plots of the data, as they are unlikely to make such themselves.

[7] This may arise for example, when the raw unaggregated data is subject to privacy restrictions that cease to be a problem with sufficient aggregation. Due to New Zealand's privacy laws (see Section 2.5.2) we must remain wary of potential privacy breaches even if we are using data that has been released 'openly' by another party. N.B. When dealing with data from another state, it would be wise to consult the relevant privacy laws for that state.

Semi-Technical Users These are 'ordinary' people with some technical knowledge, or they are in a profession which requires some technical skill with data, such as data journalists or financial professionals. They may carry out some data manipulation, usually manually and often with the aid of a GUI tool such as Excel, to draw conclusions from the data. As manipulation is often manual, they generally prefer human-friendly data formats; machine-readability is a concern to the extent that the data is easily selected for copying or to otherwise refer to.

Technical Users These are users who use script-based approaches to data manipulation to handle very large volumes of data. They typically desire data in as raw and simple a format a possible so that it is machine-readable; they can then use this to produce any other format they may desire, and of course to conduct more sophisticated statistical analyses.

Desirable Open Data will depend on the type of user, ideally Open Data should be released in multiple formats to cater to the different users. Where that is impractical the format should be one that caters to as many users as possible, ideally encompassing *Technical Users* who are capable of taking this data and re-using it in a way the benefits the other users. Conversely, though *Casual Users* form the largest user base they are also not capable of re-using the data to further benefit other users, and thus catering only to *Casual Users* will severely limit the benefits of Open Data.

Our focus will typically be from the perspective of a *Technical User*, desiring machine-readable data as inputs, with the aim of producing tools and output that are suitable for use by both *Casual* and *Semi-Technical Users*.

2.3 Good and Bad Open Data

Within Open Data, it is helpful to have some examples of what might be ideal and desirable qualities to see in Open Data, and conversely, what might be undesirable qualities. These are from the perspective of a *Technical User* (as defined in Section 2.2).

We would consider Open Data to be *Good* if it possessed many desirable qualities and few undesirable qualities. Conversely, Open Data possessing many undesirable qualities (but still classified as Open Data) we will call *Bad* Open Data. Desirable qualities are:

- Any restrictions (or lack thereof) are made clear, e.g. a known copyright licence is clearly attached (see Section 2.5.1).
- Data is accessible in a simple, easy way, e.g. the data can be obtained by simply clicking on a link.
- Data is structured in a simple, easy to use format (see Section 2.4).
- Any supplementary information is clearly connected to the data. These may include:
    1. How the data was collected and processed, including any steps taken to identify and correct errors in the data (such as mistakes during data entry), or any modifications made to ensure confidentiality and privacy protection.
    2. A clear definition of the headings, groupings or other notation used in the data.
    3. Any relevant information that may affect analysis and interpretation, e.g. if there was a change in how data was collected, processed or classified over time.

The lack of desirable qualities (such as not having a copyright licence) is undesirable. However in some cases, there may be specific features that are undesirable, such as:

- Data is only accessible via an inconvenient or slow user interface, such as a restrictive online form.
- Data is structured for presentation and human reading. This could be through excessive aggregation of the data before release, or a layout that is chosen for aesthetic appeal rather than ease of data processing. In particular, this makes it very difficult to use the data in any other way than originally envisaged by the data provider (which may be intended in some cases), stifling innovation and potential re-use of the data.

2.4 Data Formats

The data format can significantly influence how the data can be used. This section discusses some common data formats we have seen used in State Sector Open Data Sources and how these formats can enable or prevent usage of the data.

2.4.1 Executive Summary

- Releases in CSV are the simplest, and hence enable the greatest data re-use but requires some experience with data manipulation making it less accessible to non-technical users.
- Releases in XLS can have significantly different implications for data use and re-use depending on the specific internal format. Of these internal formats the Pivot Table is most recommended as it caters to the widest range of users, enabling some data manipulation by users within Excel, while also making it possible to convert and export the data to a simpler, machine-readable format. However, ideally the data should also be released in a simple multi-platform format, such as CSV.
- Releases in HTML Tables are easier to share via the web while still being structured to some extent, making it possible to salvage and convert the data to a more convenient format. However, where possible the data should also released in a different format (such as CSV) so such conversions are not necessary.
- Releases in PDF Tables should never be done as the PDF is purely a presentation format and not suitable for data transfer. Where data is included in a PDF document it should also be available separately in a better format, and this should be clearly indicated and referred to in the PDF document.

2.4.2 CSV

Perhaps the most basic data format is the *CSV*, or *Comma Separated Values*. Despite the name, CSV-like files that use other delimiters, such as a blank space or a tab, are sometimes called CSV. In CSV data is stored in *long-form*, that is each row represents an individual observation and contains all the information relating to that individual in that single row, split several 'columns' as indicated by the delimiter. An analogy would be if we stored information on people, then each row would be a person and the columns would be attributes like height, weight, gender and age. We do not require information from any other person (row) in order to obtain information about any given person. This simplicity of structure makes CSV files very easy to read and manipulate via a computer as there is no need to decipher relationships in the data, but this simplicity can make the format cumbersome to work with for users unfamiliar with data manipulation of such a format.

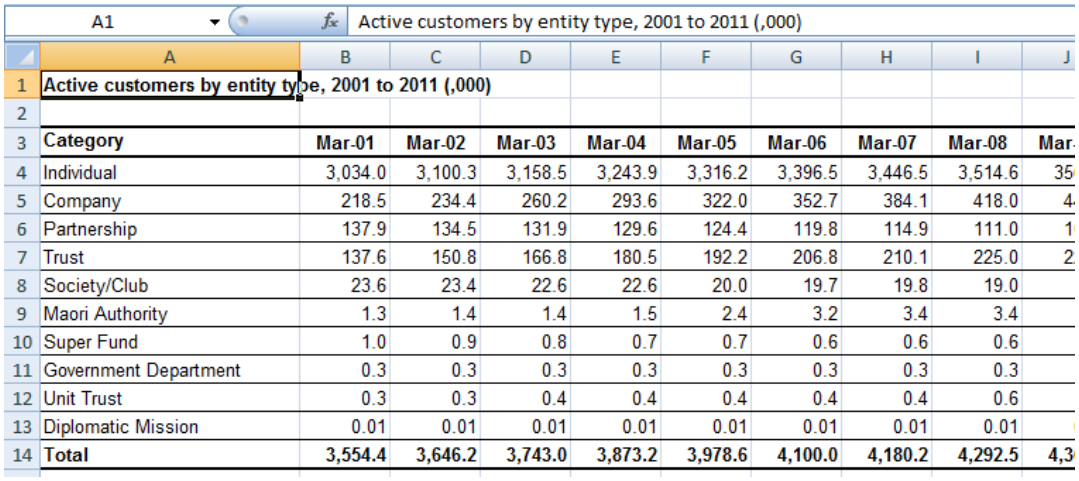
`Col1,Col2,Col3`
`"A",1,19`
`"B",2,28`
`"C",3,37`
`"Longer text",4,46`
	`Col1 Col2 Col3`
	`"A" 1 16`
	`"B" 2 25`
	`"C" 3 34`
	`"Longer text" 4 46`
		`Col1  Col2  Col3`
		`"A"  1  16`
		`"B"  2  25`
		`"C"  3  34`
		`"Longer text"  4  46`

Table 1: Example CSV and CSV-like data using (from left-to-right) a comma, a single space and a tab as the delimitters. The simplicity of data structure makes it very easy for this data format to be read and manipulated via a computer, but the lack of nice alignment of cells makes it unsuited for the naked eye.

2.4.3 XLS

Another common format is the *XLS* (*Excel Spreadsheet*), and this appears to be the preferred data format for the majority of State Sector Open Data Sources. We will further classify XLS files as follows:

XLS-Longform - When the data in the spreadsheet are presented in long-form. These are very similar to a CSV, but can add hurdles to users who may wish to work in a software other than Excel as they must convert the data first. There is little benefit in releasing data in XLS-Longform rather than CSV, as Excel is perfectly capable of reading both, while XLS-Longform can only be read by Excel or similar XLS-compatible software. See Figure 1.



A1 | Active customers by entity type, 2001 to 2011 (,000)

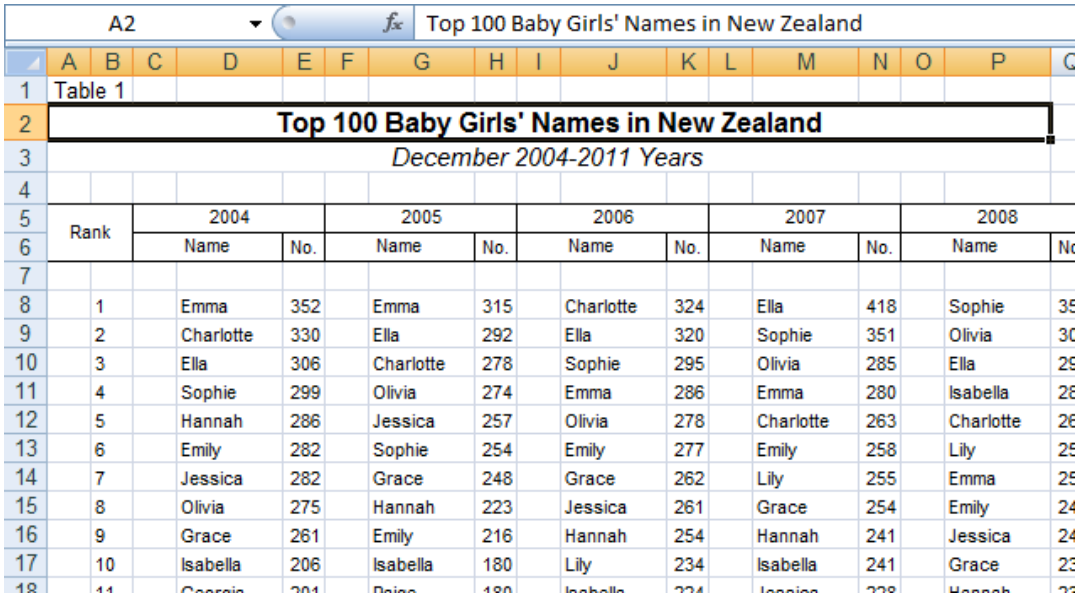
	A	B	C	D	E	F	G	H	I	J
1	Active customers by entity type, 2001 to 2011 (,000)									
2										
3	Category	Mar-01	Mar-02	Mar-03	Mar-04	Mar-05	Mar-06	Mar-07	Mar-08	Mar
4	Individual	3,034.0	3,100.3	3,158.5	3,243.9	3,316.2	3,396.5	3,446.5	3,514.6	35
5	Company	218.5	234.4	260.2	293.6	322.0	352.7	384.1	418.0	4
6	Partnership	137.9	134.5	131.9	129.6	124.4	119.8	114.9	111.0	1
7	Trust	137.6	150.8	166.8	180.5	192.2	206.8	210.1	225.0	2
8	Society/Club	23.6	23.4	22.6	22.6	20.0	19.7	19.8	19.0	
9	Maori Authority	1.3	1.4	1.4	1.5	2.4	3.2	3.4	3.4	
10	Super Fund	1.0	0.9	0.8	0.7	0.7	0.6	0.6	0.6	
11	Government Department	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	
12	Unit Trust	0.3	0.3	0.4	0.4	0.4	0.4	0.4	0.6	
13	Diplomatic Mission	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	
14	**Total**	**3,554.4**	**3,646.2**	**3,743.0**	**3,873.2**	**3,978.6**	**4,100.0**	**4,180.2**	**4,292.5**	**4,3**

Figure 1: An example of a XLS-Longform. The data is *active-customers-by-entity-type-2001-to-2011.xls* from *Inland Revenue*.

XLS-Table - When the data is presented in some form of table that relies on non-trivial relationships. Unlike a long-form where all the information for a piece of data is essentially contained in a single row, a XLS-Table requires collecting information across multiple rows to form the full picture. Such tables are relatively easy for a human brain to decipher and may even be easier to read for a human compared to a long-form, however the non-trivial relationships makes it extremely difficult to machine-read the data. See Figures 2 and 3.

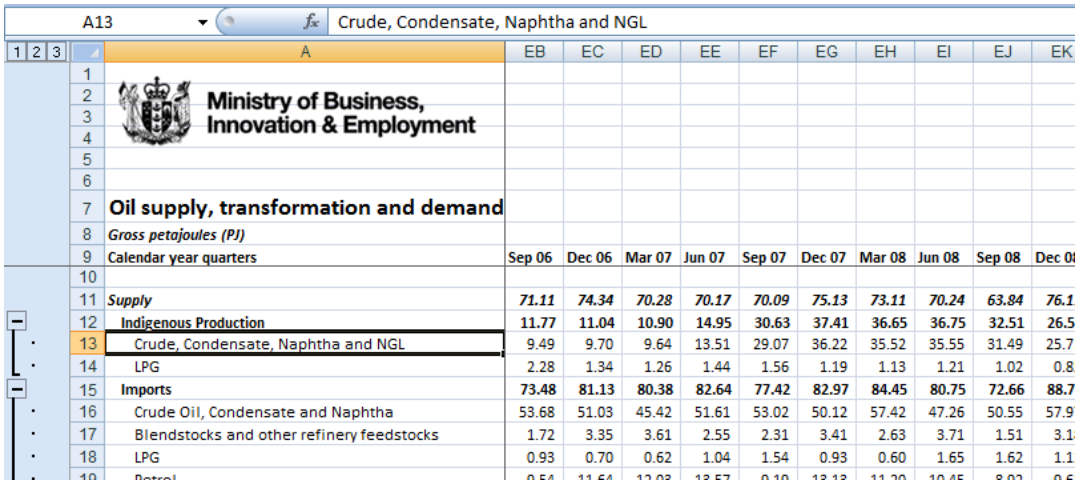
XLS-PivotTable - When the data in the spreadsheet is presented using the Pivot Table feature of Excel. The Pivot Table is an interactive feature of Excel that allows dynamic creation and modification of tables via a graphical user interface. It greatly enhances data exploration that can be done within Excel, though at first glance appears to be a significant barrier to analysis outside Excel. However, this is not the case, as it is possible to extract the underlying data using *Show Details* from the right-click menu. This makes XLS-PivotTables much better than XLS-Tables as it caters to many different types of users. See Figures 4 and 5.

XLS-ReportTable - When the data is presented as some form of report or table that is highly tuned to human-reading. We classify a XLS file as ReportTable if there are substantial barriers to machine reading. Generally any data released as a XLS-ReportTable can only be used manually and presents significant barriers to any re-use, even by manual means. This form of release is only appropriate if the underlying data is also released in a simpler form. See Figure 8.



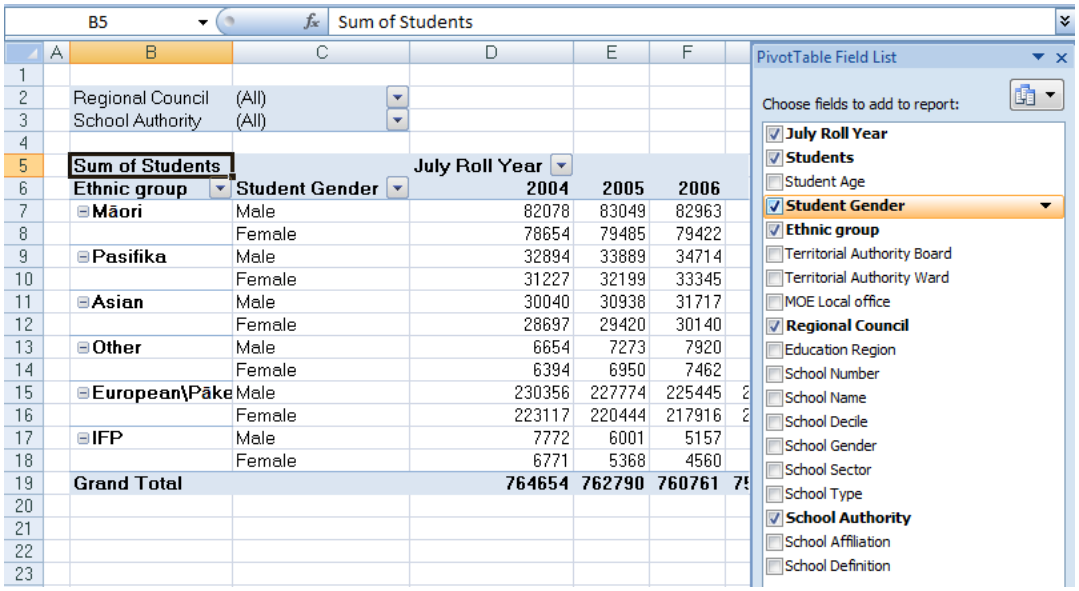
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Table 1															
2	**Top 100 Baby Girls' Names in New Zealand**															
3	*December 2004-2011 Years*															
4																
5	Rank		2004			2005			2006			2007			2008	
6			Name	No.		Name	No.		Name	No.		Name	No.		Name	No
7																
8		1	Emma	352		Emma	315		Charlotte	324		Ella	418		Sophie	35
9		2	Charlotte	330		Ella	292		Ella	320		Sophie	351		Olivia	30
10		3	Ella	306		Charlotte	278		Sophie	295		Olivia	285		Ella	29
11		4	Sophie	299		Olivia	274		Emma	286		Emma	280		Isabella	28
12		5	Hannah	286		Jessica	257		Olivia	278		Charlotte	263		Charlotte	26
13		6	Emily	282		Sophie	254		Emily	277		Emily	258		Lily	25
14		7	Jessica	282		Grace	248		Grace	262		Lily	255		Emma	25
15		8	Olivia	275		Hannah	223		Jessica	261		Grace	254		Emily	24
16		9	Grace	261		Emily	216		Hannah	254		Hannah	241		Jessica	24
17		10	Isabella	206		Isabella	180		Lily	234		Isabella	241		Grace	23

Figure 2: An example of a XLS-Table. While this may be quite easy for the human brain to decipher, it is a nightmare for a computer. Consider for instance trying to obtain all the information for the first 'piece' of data. First we need to get the Rank from B8, understand this number is a 'Rank' by getting the column label from A5 (noting this is a different column to where the actual Rank value is stored). Second we need the name from D8, and the label for this from C6. Third we need to get the count of babies from E8 with label from E6. Finally we need the year from C5. The data is *Top100BabyNamesNewZealand2011.xls* from the *Department of Internal Affairs*.



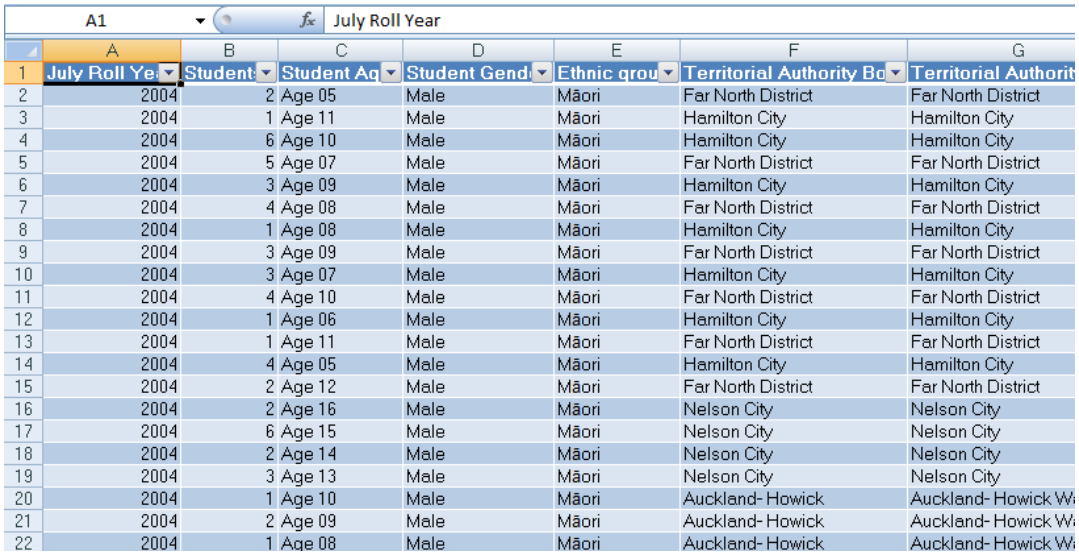
	A	EB	EC	ED	EE	EF	EG	EH	EI	EJ	EK
1											
2	Ministry of Business, Innovation & Employment										
3											
4											
5											
6											
7	**Oil supply, transformation and demand**										
8	***Gross petajoules (PJ)***										
9	**Calendar year quarters**	**Sep 06**	**Dec 06**	**Mar 07**	**Jun 07**	**Sep 07**	**Dec 07**	**Mar 08**	**Jun 08**	**Sep 08**	**Dec 0**
10											
11	***Supply***	***71.11***	***74.34***	***70.28***	***70.17***	***70.09***	***75.13***	***73.11***	***70.24***	***63.84***	***76.1***
12	**Indigenous Production**	**11.77**	**11.04**	**10.90**	**14.95**	**30.63**	**37.41**	**36.65**	**36.75**	**32.51**	**26.5**
13	Crude, Condensate, Naphtha and NGL	9.49	9.70	9.64	13.51	29.07	36.22	35.52	35.55	31.49	25.7
14	LPG	2.28	1.34	1.26	1.44	1.56	1.19	1.13	1.21	1.02	0.8
15	**Imports**	**73.48**	**81.13**	**80.38**	**82.64**	**77.42**	**82.97**	**84.45**	**80.75**	**72.66**	**88.7**
16	Crude Oil, Condensate and Naphtha	53.68	51.03	45.42	51.61	53.02	50.12	57.42	47.26	50.55	57.9
17	Blendstocks and other refinery feedstocks	1.72	3.35	3.61	2.55	2.31	3.41	2.63	3.71	1.51	3.1
18	LPG	0.93	0.70	0.62	1.04	1.54	0.93	0.60	1.65	1.62	1.1

Figure 3: An example of a XLS-Table. Once again, this may be relatively easy for a human brain to decipher, but a nightmare for a computer. Note that in this instance in-cell indentation and the Excel Grouping feature are used to communicate the hierarchical relationships in the data rows. Such information will typically be lost when converting the XLS file to another format, effectively forcing the user to a) Use Excel and b) Conduct any analysis manually. As Excel itself understands some of this information, it may be possible to write code in VBA (*Visual Basic for Applications*, a slightly modified version of *Visual Basic* integrated with Microsoft Office applications) to extract this information, but a cursory look at the code suggests this might be quite a challenge. The data is *Oil.xls* from the *Ministry of Economic Development* (now a part of the *Ministry of Business, Innovation & Employment*).



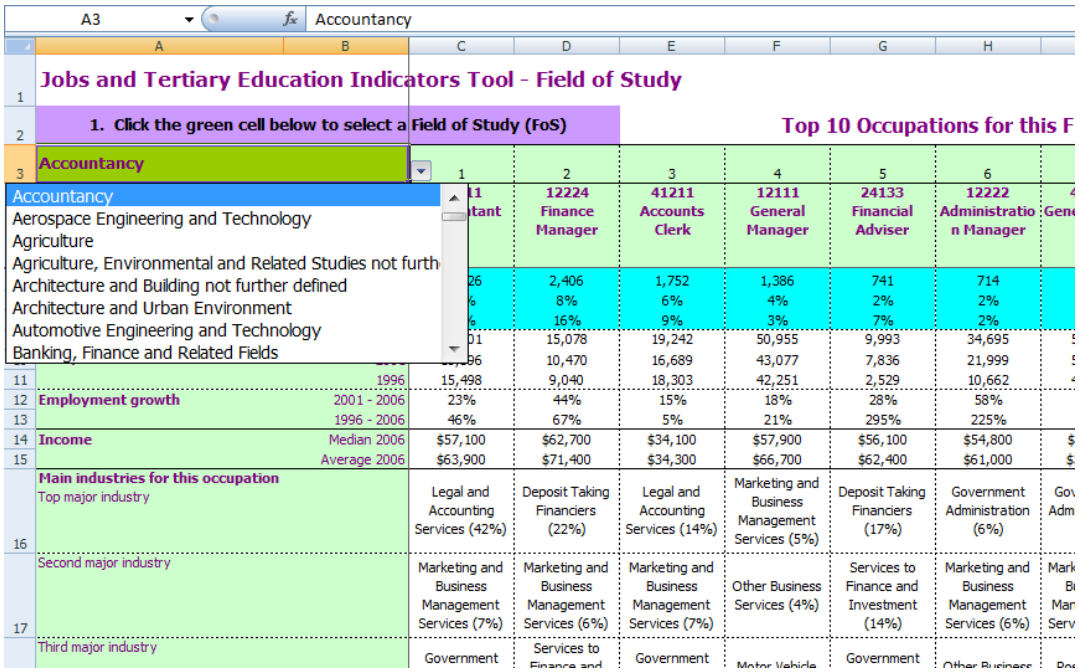
	Regional Council	(All)				
	School Authority	(All)				
	Sum of Students		July Roll Year			
	Ethnic group	Student Gender	2004	2005	2006	
	Māori	Male	82078	83049	82963	
		Female	78654	79485	79422	
	Pasifika	Male	32894	33889	34714	
		Female	31227	32199	33345	
	Asian	Male	30040	30938	31717	
		Female	28697	29420	30140	
	Other	Male	6654	7273	7920	
		Female	6394	6950	7462	
	European\Pāke	Male	230356	227774	225445	2
		Female	223117	220444	217916	2
	IFP	Male	7772	6001	5157	
		Female	6771	5368	4560	
	Grand Total		764654	762790	760761	75

Figure 4: An example of a XLS-PivotTable. The Pivot Table is an interactive feature of Excel that allows dynamic creation and modification of tables via a graphical user interface. It greatly enhances data exploration that can be done within Excel, though at first glance appears to be a significant barrier to analysis outside Excel. However, this is not the case, as it is possible to extract the underlying data using *Show Details* from the right-click menu. The data is *Pivot-Table-Student-Numbers-by-Age.xls* from the *Ministry of Education*.



July Roll Ye	Student	Student Ag	Student Gend	Ethnic grou	Territorial Authority Bo	Territorial Authorit
2004	2	Age 05	Male	Māori	Far North District	Far North District
2004	1	Age 11	Male	Māori	Hamilton City	Hamilton City
2004	6	Age 10	Male	Māori	Hamilton City	Hamilton City
2004	5	Age 07	Male	Māori	Far North District	Far North District
2004	3	Age 09	Male	Māori	Hamilton City	Hamilton City
2004	4	Age 08	Male	Māori	Far North District	Far North District
2004	1	Age 08	Male	Māori	Hamilton City	Hamilton City
2004	3	Age 09	Male	Māori	Far North District	Far North District
2004	3	Age 07	Male	Māori	Hamilton City	Hamilton City
2004	4	Age 10	Male	Māori	Far North District	Far North District
2004	1	Age 06	Male	Māori	Hamilton City	Hamilton City
2004	1	Age 11	Male	Māori	Far North District	Far North District
2004	4	Age 05	Male	Māori	Hamilton City	Hamilton City
2004	2	Age 12	Male	Māori	Far North District	Far North District
2004	2	Age 16	Male	Māori	Nelson City	Nelson City
2004	6	Age 15	Male	Māori	Nelson City	Nelson City
2004	2	Age 14	Male	Māori	Nelson City	Nelson City
2004	3	Age 13	Male	Māori	Nelson City	Nelson City
2004	1	Age 10	Male	Māori	Auckland- Howick	Auckland- Howick W
2004	2	Age 09	Male	Māori	Auckland- Howick	Auckland- Howick W
2004	1	Age 08	Male	Māori	Auckland- Howick	Auckland- Howick W

Figure 5: An example of the XLS-PivotTable from Figure 4 after clicking *Show Details* from the right-click menu. Note that the data is presented in long-form and could now be converted to a CSV for analysis in different software.



A3 | Accountancy

Jobs and Tertiary Education Indicators Tool - Field of Study

1. Click the green cell below to select a Field of Study (FoS)

Top 10 Occupations for this F

Accountancy

Accountancy
Aerospace Engineering and Technology
Agriculture
Agriculture, Environmental and Related Studies not furth
Architecture and Building not further defined
Architecture and Urban Environment
Automotive Engineering and Technology
Banking, Finance and Related Fields

			1	2	3	4	5	6	
			11 ...tant	**12224 Finance Manager**	**41211 Accounts Clerk**	**12111 General Manager**	**24133 Financial Adviser**	**12222 Administratio n Manager**	**Gen**
			26	2,406	1,752	1,386	741	714	
			%	8%	6%	4%	2%	2%	
			%	16%	9%	3%	7%	2%	
			01	15,078	19,242	50,955	9,993	34,695	
			96	10,470	16,689	43,077	7,836	21,999	
11		1996	15,498	9,040	18,303	42,251	2,529	10,662	
12	**Employment growth**	2001 - 2006	23%	44%	15%	18%	28%	58%	
13		1996 - 2006	46%	67%	5%	21%	295%	225%	
14	**Income**	Median 2006	$57,100	$62,700	$34,100	$57,900	$56,100	$54,800	$
15		Average 2006	$63,900	$71,400	$34,300	$66,700	$62,400	$61,000	$
16	**Main industries for this occupation** Top major industry		Legal and Accounting Services (42%)	Deposit Taking Financiers (22%)	Legal and Accounting Services (14%)	Marketing and Business Management Services (5%)	Deposit Taking Financiers (17%)	Government Administration (6%)	Gov Adm
17	Second major industry		Marketing and Business Management Services (7%)	Marketing and Business Management Services (6%)	Marketing and Business Management Services (7%)	Other Business Services (4%)	Services to Finance and Investment (14%)	Marketing and Business Management Services (6%)	Mark Bu Man Serv
	Third major industry		Government	Services to Finance and	Government	Motor Vehicle	Government	Other Business	Ro

Figure 6: An example of a XLS-ReportTable. Much like a XLS-Table, these are relatively easy for a human brain to decipher. However, we classify a XLS file as ReportTable if there are substantial barriers to machine reading. Generally any data released as a XLS-ReportTable can only be used manually and presents significant barriers to any re-use, even by manual means. The data is *jtei-july-08.xls* from the *Department of Labour*.

2.4.4 HTML Table

Another format is to release the data as a *HTML Table* element. Like XLS files, HTML Tables could also be released in simple long-form or in more complex structures, though practically when HTML Tables are used, they tend toward the former as it is not very suited for more complex structures. As the data is still stored in a structured format, there are several tools to convert HTML Tables to a different format, such as a CSV, for analysis in a variety of software tools. Though easy to share on a website, as HTML Tables do incur a conversion cost when re-used, the underlying data should also be released in a different downloadable format such as CSV.

Income bands for salaries and wages, 2002 to 2011

2002		
Wage/Salary Income Band	Number of People	Wage/Salary Income ($000,000)
$1 - $1,000	153,900	65.1
$1,001 - $2,000	90,320	133.7
$2,001 - $3,000	73,640	183.9

Figure 7: An example of a HTML-Table. The data is *Income bands for salaries and wages, 2002 to 2011* found under the *Wage/salary distributions for individual customers* category from *Inland Revenue*.

Results for Subject and Standard (broken down by Ethnicity and Gender)

All Subjects — National

Standard Type	Ethnicity	Gender	# of Results	# Not Achieved	# Achieved	# Merit	# Excellence
Unit Standard			1,432,835	274,170	1,157,636	921	108
	NZ Maori		337,953	73,256	264,628	68	1
		Male	171,927	41,375	130,516	35	1
		Female	166,008	31,881	134,094	33	
		Unknown	18		18		
	NZ European		745,140	123,433	620,947	663	97
		Male	400,991	75,797	324,698	420	76
		Female	344,138	47,625	296,249	243	21
		Unknown	11	11			
	Pasifika Peoples		182,627	46,485	136,055	81	6
		Male	92,961	25,122	67,814	19	6
		Female	89,650	21,352	68,236	62	
		Unknown	16	11	5		
	Asian		131,717	24,084	107,540	89	4
		Male	70,216	14,961	55,201	51	3
		Female	61,501	9,123	52,339	38	1
	Other/Unspecified Ethnicity		35,398	6,912	28,466	20	
		Male	19,266	4,099	15,148	19	
		Female	16,132	2,813	13,318	1	
Internally Assessed Achievement Standard			1,846,630	362,734	693,185	457,735	332,976
	NZ Maori		306,350	88,250	124,996	60,859	32,245
					60,779	25,751	12,473

Click here to show Standards by Subjects for Internally Assessed Achievement Standard

Figure 8: An example of a HTML-Table that does not use HTML Table code (instead utilising `div` and `span` for visual formatting). This means the data is not structured making it extremely difficult to extract the data. Luckily in this case, a link is provided (not shown) to download the data as an XLS-Table. The data is a *Results by Subject and Standard* report from the *New Zealand Qualifications Authority*.

2.4.5 PDF Table

Finally, the last format we cover is a table embedded inside a *PDF* document. These can barely be considered data as PDF documents do not retain any data structure, being purely a presentation format. Generally, any data released as a PDF can only be used manually and any re-use will be costly in time and effort. Data should never be released exclusively in a PDF format, any data included in a PDF document should be released separately in a proper data format.

2.5 Supplementary Glossary

We provide a brief glossary of certain relevant topics.

2.5.1 Licensing

Without going into technical details, a standardised *Copyright-Licence*, such as one of the *Creative Commons* licences or the *GNU General Public License*, can specify a certain set of copyright conditions or waivers. By attaching such a licence to Open Data, the legalities of copyright become clearer, allowing a user to know exactly where they stand with regards to what they may do with this data.

The *Creative Commons* licences appears to be the most common licence used by New Zealand State Sector Open Data Sources, so we provide a very brief overview here. For more information visit the *Creative Commons* website[8]. The default CC licence is the 'baseline' licence. It in effect allows the user to freely copy, distribute or make use of the work, but still assigns ownership of the copyright to the original holder. A CC licence may attach additional conditions, such as:

by Attribution - must reference the original author in the manner specified.

nc Noncommercial - can only be used for noncommercial purposes.

nd No Derivative Works - a derivative work that includes major elements of the original cannot be made. Thus you can still copy and pass on the original to others, but you cannot make any changes, alterations or additions.

sa Share-alike - derivative works may be made, but they must have a licence identical to the original.

Thus, a **CC by** licence would allow free use, modification, distribution, etc, as long as the correct attribution is given.

An alternative to using a Licence is to simply make it Public Domain, which waives any and all copyright.

2.5.2 Privacy

The relevant legislation in New Zealand is the Privacy Act 1993, which restricts anyone's collection, storage and use of data about an identifiable private individual. In essence, data on an identifiable individual can only be collected for a specific purpose, requires the person's permission, should (where possible) be taken directly from the person, should be stored securely, and should not be released to anyone else. This data should only be used for the purpose it was originally collected for (which was communicated to the person) and once it serves its purpose, it should be securely disposed (Penk and Tobin 2010).

These restrictions apply not only to the data collector, but extend to anyone who subsequently gains possession of the private data. This creates a problem: if Open Data is released that is in fact breaching privacy, any user of this 'Open' data becomes liable for any privacy breaches themselves - blame cannot be shifted to the original data provider. Fortunately, the New Zealand Privacy Commission takes a corrective role rather than a punitive one, thus in most cases the 'liability' will be minor (e.g. required to apologise to affected individuals), especially if the breach was unintentional.

[8] `http://creativecommons.org/`

Obtaining permission directly from any affected individuals, to effectively waive their privacy rights relating to the particular data collected, is allowed and can be used to avoid privacy issues. However in the case of Open Data, often the users are separate and distinct from the data collector and provider, thus obtaining permission is rarely possible.

N.B. It may be of interest to know that private information collected purely for 'domestic' use (for personal affairs) is not bound to the restrictions of the Privacy Act 1993.

3 Overview of Open Data in New Zealand

3.1 Overview of the Overview

3.1.1 Motivation

We give an overview of Open Data in New Zealand for two reasons:

1. It is useful to understand what Open Data is actually out there in practice, by who and in what form. It is also helpful to know how this data might be acquired.

2. Restricting to New Zealand limits the scope to something that is manageable.

The focus of the overview will be on the State Sector, as this is the most abundant source of Open Data. In particular, the New Zealand Government is making a push for an Open and Transparent Government and we can expect to see even more data coming from the State Sector. Related initiatives include:

Declaration on Open and Transparent Government [9] - "Building on New Zealand's democratic tradition, the government commits to actively releasing high value public data [non-personal and unclassified data.]... To support this declaration, the government asserts that the data and information it holds on behalf of the public must be open, trusted and authoritative, well managed, readily available, without charge where possible, and reusable, both legally and technically."

NZGOAL (New Zealand Government Open Access and Licensing)[10] - "The Government wants to encourage... individuals and organisations to be able to leverage State Services agencies' data stores for their own, agencies' and others' benefit... In essence, NZGOAL... sets out a series of open licensing and open access principles"

The government has recently released the **2012 Report On Agency Adoption of the New Zealand Declaration on Open and Transparent Government**[11] (New Zealand Government ICT 2012). Of particular interest is paragraph 23 which highlights what the various departments and agencies consider to be barriers to releasing 'high value public data'.

- restrictive licensing terms imposed by third parties
- issues with data quality and inconsistent data
- a lack of data standards, which causes confusion with data format
- a lack of resources to address the above barriers
- a shift in culture is required
- considerable time is required to analyse the risks of releasing data for re-use
- a perceived lack of data to release
- a lack of analysis of what information is of interest to consumers

[9] `http://ict.govt.nz/programme/opening-government-data-and-information/declaration-open-and-transparent-government`

[10] `http://ict.govt.nz/guidance-and-resources/information-and-data/nzgoal`
Note that `http://nzgoal.info/` is not the official website for NZGOAL.

[11] `http://ict.govt.nz/programme/opening-government-data-and-information/declaration-open-and-transparent-government/2012-report-adoption-declaration`

As supplement to the report, the government also released information providing details of Current and Future Data Releases. We have however encountered data releases not found in either of these lists, so they are not a comprehensive source of all data releases.

3.1.2 The Method

To gain the overview we desire we examined the organisations listed as *Public Service departments* on the State Services Commission's website[12]. For each organisation the website was examined to gain some understanding of how they release data. A small sample of the discovered data was then examined in greater depth to gain an understanding of any difficulties that might be encountered in obtaining and trying to make active use the data. The organisations were then contacted to confirm our findings and to better understand the rationale behind how they release data.

We classified the organisations based on the amount of data they release, as it is reasonable to expect that organisations that release more data will have better processes in place. Additionally some were classified as being Outside the Scope of this Literature Review.

Major Sources Those identified as releasing significant amounts of data regularly and often have an extensive history of data releases.

Minor Sources Those identified as releasing data regularly, or have otherwise released a non-trivial amount of data, but not enough to be considered *Major*.

Not Sources No significant datasets found.

Outside Scope Some organisations only release geospatial data, and these are classified as Outside the Scope of this Literature Review. The complexity and specialised nature of geospatial data means it is a topic deserving of its own Thesis and is considered beyond the scope of the Thesis for which this Literature Review is conducted.

3.1.3 Situation Report

As we might expect, those sources that release large amounts of data tended to have better data releases than those that release smaller amounts. In particular, those classified as *Major* generally have a mandate to release data, often being *Tier 1 Statistics* (see Section 3.2.2), and thus have good reason to focus on releasing data. On the other hand, other organisations must focus first on their primary objectives; release of data is a secondary, even tertiary objective, one with no additional funding attached.

However, we have seen a few organisations taking the recent government initiatives as an opportunity to focus on releasing data, and some activity has been seen in attempting to overcome the lack of resources with a cross-agency solution. LINZ led a working group to develop a business case for a shared Open Data Service (ODS). The ODS Agency Demand Survey notes:

> At present agencies provide their released information through a variety of mechanisms ranging from website content though to custom designed database applications. This variability means that the data can be difficult for users to find, access and use, thereby limiting its potential reuse and subsequent benefit to the New Zealand economy.
>
> The goal of the ODS is to reduce duplication and get the best return for investment in the production of systems that enable government agencies to release high value, non-personal data, and to provide a portal or portals where users can seamlessly access information from a number of different agencies.

As noted in the above quote, such a shared effort will standardise how the data is released across many state organisations, making it much easier to find and obtain data.

Unfortunately, the initiative never made it past the consultation stage. LINZ notes:

[12] `http://www.ssc.govt.nz/state_sector_organisations`

> A survey of local, regional and central government agencies found that there was a strong level of interest in using an ODS if it were available. The same survey also revealed that agencies were unwilling to prioritise investment in the development of the service in a constrained fiscal environment.
>
> The working group concluded that while there was strong interest in the concept of an ODS, the inability to obtain funding to carry out the development meant that it was not possible to produce a viable business case for an ODS.

Unfortunately this means the current situation, where there is a lack of standardisation across the different agencies, continues. Each agency has its own website layout that can make it difficult to find the data, with some agencies even lacking a listing of all data releases. A data aggregator, like *data.govt.nz*, alleviates this problem and is currently one of the best methods for finding data (see Section 3.3 for details). Once found, it is often the case that the data is not in a format ideal for re-use (see Section 2.4 for more on data formats), requiring extra work to make use of the data.

Though change and improvements are on the horizon, the current situation poses many challenges in making use of Open Data released by state organisations.

3.2 Statistics New Zealand

`http://www.stats.govt.nz/`
Statistics New Zealand Tatauranga Aotearoa is a government department and New Zealand's national statistical office. We're New Zealand's major source of official statistics and leader of the Official Statistics System.

3.2.1 Official Statistics in New Zealand

Statisphere[13], "New Zealand's official statistics portal" managed by Statistics New Zealand (StatsNZ) "on behalf of all government departments" provides a good explanation of Official Statistics in New Zealand.

> Official statistics are defined in section 2 of the Statistics Act 1975. They are statistics derived by government departments from:
>
> - statistical surveys
> - administrative and registration records and other forms and papers that are published regularly, or planned to be published regularly, or could be published regularly.
>
> Statistical survey is defined in the Statistics Act as: "a survey of undertakings, or of the public of New Zealand, whereby information is collected from all persons in a field of inquiry or from a sample, by a Government Department with the authority of this Act or any other Act, for the purpose of processing and summarising by appropriate statistical procedures and publishing the results of the survey in some statistical form".
>
> Statistics New Zealand is New Zealand's national statistical office. Statistics New Zealand is the leader of the Official Statistics System and is the major producer of official statistics in New Zealand.
>
> The Government Statistician, who is also the Chief Executive of Statistics New Zealand... [provides] direction to the Official Statistics System and engaging other government departments to build shared ownership, minimise duplication, and maximise reuse of data.

Thus Statistics New Zealand is very important on matters relating to policy, as well as being an important source of data.

[13] `http://www.statisphere.govt.nz/`

3.2.2 Tier 1 Statistics

In 2003 a review conducted by StatsNZ identified a set of key official statistics; these are known as Tier 1 statistics[14] which...

- are essential to critical decision-making
- are of high public interest
- meet expectations of impartiality and statistical quality, in accordance with the Tier 1 principles and protocols
- require long-term data continuity
- allow international comparability
- meet international statistical obligations.

A principle and protocol of particular interest is **Protocol 5: Release Practices**. The following are excerpts from the *Principles and Protocols for Producers of Tier 1 Statistics*[15]:

- Tier 1 statistics producers will ensure equality of access. Statistics are presented in an understandable manner and are widely disseminated.
- Release of Tier 1 statistics is by the Chief Executive of the producing agency, according to a calendar of release dates published at least six months in advance.
- Information dissemination practices are responsive to the needs of users.
- Statistics are released in a variety of formats that meet the needs of users.
- Tier 1 statistics producers endeavour to integrate and harmonise their publications and products with users' needs and give them easy access to related statistics through common gateways or interlinked websites.
- Tier 1 statistics producers respond to changing expectations about access to outputs. Formats, media, content and support materials are regularly reviewed and are modified to meet users' current and future needs.
- The cost of accessing Tier 1 statistics is minimal.
- Catalogues and directories are readily available so that potential users know where and what statistics are available. A list of official statistics is available at `www.statisphere.govt.nz`
- Tier 1 statistics producers provide facilities (by electronic and/or print media) to ensure easy, user-friendly access to statistics for everyone, including regular, professional users as well as casual users and the interested public.
- Regularly recurring statistical releases are delivered in consistent formats. The format is sufficiently flexible to allow explanations of the data as they vary between periods.

Unfortunately, but understandably, practical application of these protocols appears to be very difficult, most likely due to limited resources, and these protocols seem more a goal rather than a reality.

[14] `http://www.statisphere.govt.nz/tier1-statistics.aspx`
[15] `http://www.statisphere.govt.nz/tier1-statistics/principles-protocols.aspx`

3.2.3 Statistics New Zealand as a Data Source

The majority of data is released through one of the online tools, of which there are currently three: Infoshare, Table Builder and NZ.Stat.

NZ.Stat is a new tool released near the end of 2012 that replaces Table Builder. Table Builder is scheduled to be decommissioned at the end of February 2013. NZ.Stat is also intended to replace Infoshare in 2014.

The general idea behind all the online tools are the same. They are graphical user interfaces that are used to:

1. Narrow down user query to a specific category.
2. Provide a breakdown of variables related to the chosen category.
3. Generate a table based on the selection.

The output is a table tailored more for human consumption and is not presented in Longform.

At this stage, it would appear there is no API support for any of the online tools, though NZ.Stat may have this capability in the future "We will also be providing more functionality, such as the ability to create your own custom queries that can be shared with colleagues, and machine-to-machine data transfer."[16]

Some data is released via direct download, usually in an XLS format. The internal format of these files are variable, but they appear to be intended for direct consumption by human audiences (i.e. Table, not Longform). In some cases, the files mention that more data can be found via one of the tools, though output generated from Infoshare appeared to be considerably more aggregated than the data available via direct download.

3.3 data.govt.nz

data.govt.nz is a catalogue of Open Data released by New Zealand government agencies. It is administered by the Department of Internal Affairs and has official support through the New Zealand Declaration on Open and Transparent Government initiative.

Advantages:

- good coverage of data released by government agencies
- good search engine which includes some metadata (mainly in the form of tags and categories)
- generally much easier to search for a particular agency in *data.govt.nz* than it is to look through the agency's official website
- Has official backing from the New Zealand Declaration on Open and Transparent Government initiative.

Disadvantages:

- Not all open government data can be found on *data.govt.nz*
- A lot of metadata is incomplete or left unfilled, e.g. the RBNZ datasets are missing information for many of the categories, such as Contact Person, Email, Phone, Date of creation and Frequency of update.
- *data.govt.nz* does not hold the data themselves and often does not link directly to the data, leading instead to a webpage on the agency's official website, where it may or may not be obvious how the sought data should be obtained.
- Lacks a documented API to facilitate computerised calls to search and/or access metadata (though DIA plans for a blog post related to this at `http://webtoolkit.govt.nz/`).

[16] `http://www.stats.govt.nz/tools_and_services/tools/nzdotstat.aspx`

It is worth noting that *data.govt.nz* is not the only data catalogue. Some others are:

Statisphere `http://www.statisphere.govt.nz/` Serves as a listing of Tier 1 Statistics, though this feature is currently unavailable "pending a review and update of the content".

DigitalNZ `http://www.digitalnz.org/` A catalogue of any *digital content*.

Open Data Catalogue `http://cat.open.org.nz/` "an open, independent catalogue of Government and Local Body datasets.".

3.4 The Template

The data sources are presented using a standard format. This information is recorded in *NZDataSources.xml* and is later parsed by script to this Literature Review. As it is a structured document there are other possible ways of parsing or using this information and for such purposes the XML document is intended to be released to the public in the future, but is unavailable at this time due to unresolved confidentiality issues. The template is as follows:

Name of Organisation

`URL to Organisation's homepage`
Organisation's Purpose, quoted from their homepage.

Comments An overview of the organisation.

Data Grab Pattern Step-by-step instructions for finding the data from main page URL, generally referring to the names of the links to follow.

Data Format The type of data formats found (see Section 2.4).

Related Information Where related information (see Section 2.3) can be found. The two types found were In-file (where the information is found in the same files as the data) and webpage (where the information is found on a webpage located near where the data is found).

Found on data.govt.nz Whether the data is also found on *data.govt.nz*. Can be TRUE, FALSE or Partial.

Copyright Licence The copyright licence attached to the data.

3.5 Major Sources

3.5.1 Ministry of Economic Development

`http://www.med.govt.nz/`
The Ministry of Economic Development's purpose is to create the conditions for businesses to succeed and New Zealanders to prosper.

Comments The Ministry of Economic Development (MED) is one of the organisations that have been merged under the Ministry of Business, Innovation and Employment (MBIE). Internally this merge is more or less complete though the merging of website content has been a low priority. Thus MBIE still makes MED-related releases through the former MED website.

The MED does not yet have a centralised process for data releases and instead provides data along two primary groups: Tourism and Energy. These two groups have different teams involved and hence have different processes.

The Tourism group classes users into three categories based on their data needs: 1) Requiring high level aggregate stats, 2) Requiring flexible cross tabulation (similar to Statistics NZ's Table Builder or Infoshare), 3) Microdata (eg respondent-level survey data complete

with weights and replicate weights from standard error estimation). They see the greatest demand from category 2. Data under category 3 are provided to "bona fide researchers who asks and we are sure is not going to hurt themselves with it."

Some data is released via a third-party software called Infoview, which allows for some interactive visualisations and data exploration.

They face minimal privacy or confidentiality issues, thus the biggest constraints on releasing data are time and resources to communicate required information to users, including background necessary to understand the data.

The Energy group generally release all data unless there is a good reason not to. They currently have a number of informal procedures in place to identify potential risks of data release, and to check data quality.

All Energy data is released via Excel, in some cases using some non-trivial Excel features to present the data, such as in-cell indentation and Excel's grouping feature.

Data Grab Pattern Sectors & industries tab ->Energy data and modelling (or) Tourism research and data.

Some data not found this way, best found via data.govt.nz.

Data Format CSV, XLS-Report-Table, XLS-Table, HTML-Table

Found on data.govt.nz TRUE

Copyright Licence Crown Copyright, restrictions may apply, refer to `http://www.med.govt.nz/help-support/about-this-site/legal-notices/copyright`

3.5.2 Ministry of Education

`http://www.minedu.govt.nz/`
Building a world-leading education system that equips all New Zealanders with the knowledge, skills, and values to be successful citizens in the 21st century.

Comments The Ministry of Education releases its data through a separate website, Education Counts. Note that the New Zealand Qualifcations Authority also releases some data relating to Secondary Schools, but this data is collected as part of NZQA's operations and is often different from the data collected by the Ministry. The Ministry assesses secondary schooling achievement using NZQA data and non-NQF (National Qualifications Framework) data collected directly from schools.

The data is typically collected for administrative reasons, rather than for research purposes, though the latter does happen. Generally, the Ministry will publish all such data, using Education Counts for routine releases.

They carry out several data quality checks, including looking for results and trends out of the ordinary, checking with schools if necessary to confirm correctness of their data.

National Standards which have been introduced recently with a lot of surrounding controversy (in particular relating to potential misuse of the data by unfairly ranking schools) are currently catering to the needs of each individual school, and thus standardised national data collection is not possible. However, it is anticipated that this will eventually be possible.

The Ministry prefers to release data in tables with accompanying explanatory notes, rather than simple text. They also make use of XLS-PivotTables to allow end-user exploration of the data. Where applicable they use standardised terms and definitions, usually defined by Statistics New Zealand. They regularly receive data requests for specific breakdowns of routine data releases, which are answered where possible, subject to privacy and data quality issues.

Data Grab Pattern NZ Education ->Researchers ->Education Counts

Links to http://www.educationcounts.govt.nz/

Data Format XLS-Table, XLS-PivotTable, HTML-Table

Found on data.govt.nz TRUE

Copyright Licence Crown Copyright, but also "CC By", refer to `http://www.educationcounts.govt.nz/help/privacy`

3.5.3 Inland Revenue

`http://www.ird.govt.nz/`
We collect most of the revenue that government needs to fund its programmes. We also administer a number of social support programmes.

Comments The Inland Revenue Department (IRD), more commonly just "Inland Revenue", makes regular releases of tax statistics via an easy to use listing of data on their website. The data release pages themselves provide simple plots along with links to the data.

They also release some other data that are not on the tax statistics listing (as they are not tax statistics), such as the quarterly Customer Satisfaction and Perceptions Survey results or the Industry Benchmarks. These are however not found via data.govt.nz.

The Industry Benchmarks are created by Statistics New Zealand from data provided by the IRD, the results of which are published on the IRD website.

IRD has a formal 15 point process for releasing data. Broad steps involve quality control, peer-review, production of tables and graphs and an assessment of confidentiality and other disclosure procedures (including accompanying notes).

To encourage re-use of the data the IRD clearly releases all data under a Creative Commons licence (CC BY), and releases all data in both HTML and Excel formats.

Data Grab Pattern About us ->Research and tax statistics
(The sources below are not linked to in data.govt.nz)
Business & employers ->Industry benchmarks ->Find Your Industry Benchmarks

Data Format XLS-Table, HTML-Table

Found on data.govt.nz Partial (Only lists Tax Statistics)

Copyright Licence CC By

3.5.4 Ministry of Transport

`http://www.transport.govt.nz/`
The Ministry of Transport is the government's principal transport adviser, and the bulk of our work is in providing policy advice and support to Ministers.

Comments The Ministry of Transport releases a large volume of data in a reasonably easy to use format, primarily XLS or CSV. Though some data is only available as part of a PDF report, in personal correspondence these have been identified as mostly a legacy feature, and the Ministry intends to release such data in XLS or CSV in the future.

The Ministry considers the main cost to be in collecting data. Once collected, they do not consider the release process to be a major burden. They also possess more detailed data (unit-record level, with potential privacy issues) that may be available to researchers on request.

Data Grab Pattern Research

Data Format XLS-Table, HTML-Table, PDF-Report-Table

Found on data.govt.nz TRUE

Copyright Licence Crown Copyright, restrictions may apply, refer to `http://www.transport.govt.nz/copyright-and-disclaimer/`

3.5.5 Reserve Bank of New Zealand

`http://www.rbnz.govt.nz/`
The Reserve Bank of New Zealand is New Zealand's central bank. We promote a sound and dynamic monetary and financial system.

Comments The Reserve Bank of New Zealand (RBNZ) releases a significant amount of data often going back decades. They release data in two formats, HTML-Tables for recent data and XLS-Tables for historical data (which includes recent data).

They also release Key Graphs on topics of high public interest, such as Inflation, GDP or Unemployment. These graphs also have accompanying commentary and a link to the data (XLS).

The RBNZ maintains an Advance Release Calendar which provides details on when data is updated, or the update frequency (e.g. Daily).

Data Grab Pattern Statistics

Data Format XLS-Table, HTML-Table

Found on data.govt.nz TRUE

Copyright Licence Slightly conflicting. The Statistics page states: "Material published in our website Statistical tables may be used without restriction, but acknowledgement would be appreciated." However, the copyright page states some restrictions: `http://www.rbnz.govt.nz/0161308.html` Most likely, the statement on the Statistics page is given precendent.

3.6 Minor Sources

3.6.1 Department of Building and Housing

`http://www.dbh.govt.nz/`
Our vision is for a building and housing market that delivers good quality homes and buildings for New Zealanders that contribute to strong communities and a prosperous economy.

Comments The Department of Building and Housing (DBH) is one of the organisations that have been merged under the Ministry of Business, Innovation and Employment (MBIE). Internally this merge is more or less complete though the merging of website content has been a low priority. Thus MBIE still makes Building and Housing related releases through the former DBH website.

As things stand, the website lacks a comprehensive listing of all data releases, and not all data releases are listed on data.govt.nz, presenting a significant barrier in finding the data.

The Department releases reports with accompanying commentary, graphs and data, but this data is not available separately making re-use difficult. They have however indicated that more data may be available upon request, e.g. they regularly provide raw rental bond data to those who request it.

They have indicated a desire to release their Key Indicator Reports (currently released as PDF reports) as 'live' spreadsheets, "so that interested people can access the raw data behind the graphs", though the integration into MBIE is likely to delay this development.

They have recently made open some data behind their Key Indicator Reports. Currently this only consists of some market rent data, but the data is in a very simple, machine-friendly format.

Quality assurance procedures exist for any data releases, including data quality checks and peer review of analysis.

Data Grab Pattern No Standard Pattern.

Data Format HTML-Table, PDF-Report-Table, XLS-Report-Table

Found on data.govt.nz Partial

Copyright Licence Crown Copyright, restrictions may apply, refer to `http://www.dbh.govt.nz/disclaimer-and-copyright`

3.6.2 Department of Conservation

`http://www.doc.govt.nz/`
This Department of Conservation (DOC) site has information about the protection of New Zealand's natural and historic heritage, how and where you can enjoy public conservation places and how to get involved in conservation.

Comments The Department of Conservation (DOC) has a history of releasing data even prior to official government policy on data releases, however there were no formal procedures or processes in place. While care was given when releasing a dataset, including considerations for using common formats for releases rather than what may be most convenient to use internally, these have been ad hoc. Their website management model has also meant that it was easy for data releases to be duplicated, as evidenced by the website currently having two independent but very similar listings of data (see DataGrabPattern).

Despite this, the data releases are in fairly good condition. Much of the data released are geospatial released through a third-party provider, Koordinates Limited. Other releases are provided as HTML-Tables, though a few appear to only be available as part of PDF reports.

DOC had actively participated in the Open Data Service initiative (see Section 3.1.3), which they had intended to use for future data releases.

They are now investing in a federated infrastructure for a biodiversity data exchange. This is intended to enable sharing across agencies (both nationally and internationally) and also enable access by the public.

Data Grab Pattern About DOC ->Role ->Visitor statistics & research

Publications ->About DOC ->Role ->Maps & statistics

The two pages above have some duplicates between them, but one may have something the other doesn't.

Data Format HTML-Table, PDF-Report-Table

Found on data.govt.nz Partial

Copyright Licence Crown Copyright, restrictions may apply, refer to `http://www.doc.govt.nz/footer-links/copyright/`

3.6.3 Department of Corrections

`http://www.corrections.govt.nz/`
The Department of Corrections enforces the sentences and orders of the criminal courts and parole board. Corrections improves public safety by ensuring sentence compliance and works to reduce re-offending by providing offenders with rehabilitation programmes, education and job training.

Comments The Department of Corrections does not release a lot of data, but they do make regular quarterly releases for "community sentences and orders statistics" and "prison statistics". These releases are as HTML-Tables and come with some simple graphs.

They are currently in discussions with the Department of Internal Affairs on how best to support the Open Data initiative, and are likely to adopt what DIA prescribe.

Data Grab Pattern About Us ->Facts and statistics

Data Format HTML-Table

Found on data.govt.nz FALSE

Copyright Licence Crown Copyright, restrictions may apply, refer to `http://www.corrections.govt.nz/utility-navigation/disclaimer-and-copyright.html`

3.6.4 Ministry of Health

`http://www.health.govt.nz/`
The Government's principal advisor on health and disability: improving, promoting and protecting the health of all New Zealanders

Comments The Ministry of Health releases data primarily to satisfy various requirements placed upon the Ministry. This data may be classified under Tier 1 Statistics, or are released for other accountability purposes, and hence have various requirements on how they are released. Typically these data releases are for direct human consumption and are often not ideal for machine-reading.

Internal discussions are underway to adopt the Government's Open Data Initiative, including working towards more machine readable formats, but as no additional resources were provided it may take a while (potentially several years) to incorporate the Open Data processes into existing projects.

Data Grab Pattern Health statistics

Data Format XLS-Report-Table

Found on data.govt.nz TRUE

Copyright Licence Crown Copyright, restrictions may apply, refer to `http://www.health.govt.nz/about-site/copyright`

3.6.5 Department of Internal Affairs

`http://www.dia.govt.nz/`
The Department of Internal Affairs serves and connects people, communities and government to build a safe, prosperous and respected nation.

Comments The Department of Internal Affairs (DIA) does not release significant amounts of data but is notable for being the department that administers data.govt.nz, webtoolkit.govt.nz and ict.govt.nz, which provides the platform, tools and policies that will no doubt affect the open data releases of other government organisations.

The DIA does not maintain its own independent listing of data, instead opting to utilise data.govt.nz's ATOM feed to generate a list of their data releases, which is then presented to the user. This is good to see as it reduces duplication of effort (DIA does not have to spend additional resources to update its own independent listing) and makes it more likely that their listing on data.govt.nz is more comprehensive and up-to-date.

Unfortunately this method of using data.govt.nz to generate a listing automatically has not been suggested to the other agencies, though our contact with DIA considered it to be a good idea. It would be good to see the method adopted by other agencies in the future.

Being the agency that administers ict.govt.nz, DIA also has extensive documentation on data release processes which are available publically at ict.govt.nz:

`http://ict.govt.nz/library/3191296DA%20-%20Process%20for%20Prioritisation%20and%20Release%20of%20High%20Value%20Public%20Data%20for%20Reuse_0.pdf`

`http://ict.govt.nz/library/3191296DA-Open-Data-Identification-Prioritisation-and-Planning-Template-Worksheet_0.xls`

Data Grab Pattern Data and Statistics

Data Format XLS-Report-Table, XLS-Table, CSV

Found on data.govt.nz TRUE

Copyright Licence Licence generally denoted in the data.govt.nz listing (usually "CC By"). Where it's not specified, it is Crown Copyright, restrictions may apply, refer to `http://www.dia.govt.nz/Legal-Copyright-Index`

3.6.6 Department of Labour

`http://www.dol.govt.nz/`
To grow New Zealand's economy and improve the quality of lives through a high-performing labour market and immigration system.

Comments The Department of Labour is one of the organisations that have been merged under the Ministry of Business, Innovation and Employment (MBIE). Internally this merge is more or less complete though the merging of website content has been a low priority. Thus MBIE still makes Labour related releases through the former Department of Labour's website.

The website links to a Statistics page which leads to a Health and Safety Statistics page that contains Workplace fatalities, serious harm and prosecutions data. These datasets cannot be found via data.govt.nz, though data.govt.nz links to several hidden pieces of data (often contained in reports) that were not found by following the Statistics link from the website.

No formal policies existed for the release of data. Such work was underway when the Department was merged into MBIE and has now ceased, presumably to be replaced by a unified policy from MBIE.

Data Grab Pattern Statistics

Data Format XLS-Report-Table, HTML-Table

Found on data.govt.nz Partial

Copyright Licence Crown Copyright, restrictions may apply, refer to `http://www.dol.govt.nz/common/copyright.asp`

3.6.7 Ministry for Primary Industries

`http://www.mpi.govt.nz/`
Our vision is to grow and protect New Zealand. We do this by: maximising export opportunities for the primary industries; improving sector productivity; increasing sustainable resource use; and protecting New Zealand from biological risk.

Comments The Ministry for Primary Industries (MPI) releases data collected from their own activities, and also data sourced from Statistics New Zealand. Some of these are available only as a part of a report making re-use difficult. For some reports (e.g. Farm Monitoring Reports) the data is available separately, but only on request.

The Ministry used to release data in both Excel and CSV format, but have since stopped as they found the process to be too cumbersome. CSV format is still available, but only on request.

Most of MPI's data releases are Tier 1 Statistics and follow the guidelines specified for Tier 1 Statistics (see Section 3.2.2). The Ministry intends to incorporate the NZGOAL framework (see Section 3.1.1) for future releases.

Data Grab Pattern News & Resources ->Statistics & Forecasting

Data Format XLS-Table

Found on data.govt.nz Partial

Copyright Licence Unusual copyright, restrictions may apply, refer to `http://www.mpi.govt.nz/Copyright.aspx`

3.6.8 Ministry of Social Development

`http://www.msd.govt.nz/`
Our Ministry is all about helping to build successful individuals, and in turn building strong, healthy families and communities.

Comments While the Ministry collects large amounts of data as part of their core purpose (of Social Development), much of the data collected is sensitive information and collected to fulfil the Ministry's primary objectives. Any release of such data to the wider public is incidental to their core purpose, which contributes to the manner in which they release data. The Ministry of Social Development mostly releases data as reports rather than in formats more conducive to re-use.

The Ministry is aware of some of the limitations in how they currently release the data but their focus is on fulfilling their core purpose, not on the public release of collected data. They receive hundreds of Official Information Act requests every year, in addition to requests from parliament, media and researchers, which provides some idea of what sort of data is in demand, but the potential for misuse of data limits what they can release as Open Data.

The Ministry has very detailed data quality procedures which was audited by PricewaterhouseCoopers in 2006 and is periodically audited by Audit NZ.

Their listing on data.govt.nz reveals statistics released by StudyLink (which is one of the services provided by the Ministry), and links to their Annual Statistical Reports, but does not link to their primary listing of Statistics found via the *Data Grab Pattern* steps below.

Data Grab Pattern Publications & Resources ->Statistics

Data Format HTML-Table, PDF-Report-Table, DOC-Report-Table

Found on data.govt.nz Partial

Copyright Licence Crown Copyright, restrictions may apply, refer to `http://www.msd.govt.nz/about-msd-and-our-work/tools/copyright-statement.html`

3.7 Not Sources

3.7.1 Ministry of Business, Innovation and Employment

`http://www.mbie.govt.nz/`
A new 'parent' Ministry for: the Department of Building and Housing, the Ministry of Economic Development, the Department of Labour and the Ministry of Science and Innovation.

Formed on 1 July 2012, internally the merge is more or less complete but the merging of website content has been a low priority. MBIE still communicates mainly via the old websites and because of this, we cover the websites of the above agencies separately.

3.7.2 Crown Law Office

`http://www.crownlaw.govt.nz/`
No Comments.

3.7.3 Ministry for Culture and Heritage

`http://www.mch.govt.nz/`
No Comments.

3.7.4 Ministry of Defence

`http://www.defence.govt.nz/`
No Comments.

3.7.5 Education Review Office

`http://www.ero.govt.nz/`
ERO is not a true data provider (instead see Ministry of Education for education data), but the School Reports `http://www.ero.govt.nz/Early-Childhood-School-Reports` contain some broad "About the School" information presented as an HTML-Table. Contains information like: Decile, School roll, Gender composition, Ethnic composition.

3.7.6 Ministry of Foreign Affairs and Trade

`http://mfat.govt.nz/`
The Ministry appears to be policy focused and do not release trade-related data.

3.7.7 Government Communications Security Bureau

`http://www.gcsb.govt.nz/`
No Comments.

3.7.8 Ministry of Justice

`http://www.justice.govt.nz/`
Some very old data might be found via data.govt.nz, recent data appears to be held by StatsNZ.

3.7.9 Ministry for Maori Development

`http://www.tpk.govt.nz/en/`
Only 'data' is the Iwi Directory. This is grouped by region but these region names are in Maori with no English translation provided, reducing accessibility.

However, a simple but effective interactive map is also provided, which can also be used to navigate to the region sub-pages (still only has Maori names, but can at least go off the geography).

Beyond the directory, there is also information on each Iwi's 'population' (found in the individual Iwi's sub-page).

3.7.10 New Zealand Customs Service

`http://www.customs.govt.nz/`
Somewhat disappointing that no data is apparently available to the public. Might have been interesting to see data such as a table of goods with annual tariff revenue. Note StatsNZ does have some data on imports and exports, such as goods by value and volume by country, but does not appear to have customs specific data (e.g. tariffs, goods seized, etc.).

3.7.11 Ministry of Pacific Island Affairs

`http://www.mpia.govt.nz/`
Some simple statistics available on their website, but no real data. Statistics are likely derived from StatsNZ data.

3.7.12 Department of the Prime Minister and Cabinet

`http://www.dpmc.govt.nz/`
No Comments.

3.7.13 Serious Fraud Office

`http://www.sfo.govt.nz/about`
No Comments.

3.7.14 State Services Commission

`http://www.ssc.govt.nz/`
List of state organisations `http://www.ssc.govt.nz/state_sector_organisations`

3.7.15 Ministry of Women's Affairs

`http://www.mwa.govt.nz/`
Only 'data' found is Students' Occupational Choice Study, Dunedin, Auckland, which is mostly a report but includes some HTML-Tables of data.

It also has a Statistics page `http://www.mwa.govt.nz/women-in-nz/stats` that links to some other agencies (Ministry of Social Development, StatsNZ, Department of Labour).

3.8 Outside Scope

3.8.1 Ministry for the Environment

`http://www.mfe.govt.nz/index.html`
Only release geospatial data and all data appears to be released via a third-party provider, Koordinates Limited (also used by the Department of Conservation).

3.8.2 Land Information New Zealand

`http://www.linz.govt.nz/`
Geospatial data only, see `http://www.geodata.govt.nz/`.

4 Conclusion

The value of Open Data can be significant but realising this value is dependent on how easy the Open Data is to find, access and use. This Literature Review has examined features necessary for *Good* Open Data - clear copyright terms, easy to find, easy to access, good data structure and accompanying documentation. It is also clear from Protocol 5 of *Principles and Protocols for Producers of Tier 1 Statistics* that the government agrees these, and other considerations, are important. Unfortunately, resource constraints mean few government agencies are able to fully meet these requirements.

While a lot of Open Data is out there, and while improvements are on the horizon, there are currently many challenges in making use of this Open Data. This also means there are many opportunities to create tools that will make this data more useful, and hence help unlock the benefits of this Open Data.

References

Cole, R., 2012. Some observations on the practice of open data as opposed to its promise. The Journal of Community Informatics 8 (2).
URL `http://ci-journal.net/index.php/ciej/article/view/920`

Davies, T., Bawa, Z., 2012a. The Journal of Community Informatics 8 (2).
URL `http://ci-journal.net/index.php/ciej/issue/view/41`

Davies, T., Bawa, Z., 2012b. The promises and perils of open government data. The Journal of Community Informatics 8 (2).
URL `http://ci-journal.net/index.php/ciej/article/view/929`

New Zealand Government ICT, 2012.
URL http://ict.govt.nz/

Open Knowledge Foundation, 2012.
URL http://okfn.org/

Penk, S., Tobin, R. (Eds.), March 2010. Privacy Law in New Zealand. Brookers Ltd.

We would also like to thank all the State Sector Organisations who took the time to answer our questions. Without their help this Literature Review would be far less informative.