Schiff Consulting.

sapere
research group

REVISED VERSION

Getting more value from data sharing: definitions, participants and barriers

31 January 2017

Hayden Glass (hglass@srgexpert.com, 021 689 176)

Aaron Schiff (aaron@schiff.co.nz, 021 288 0665)

Summary

1. This is the first report on our project for the Data Futures Partnership Working Group (DFP) to explore barriers to the generation of value from data sharing.

What we have done

2. As part of this project:

- We have developed a simple economic framework for thinking about the definition of value and the motivations and barriers to sharing data.
- We have reviewed the available literature, and considered some other material on how New Zealand is going with open data, including public feedback solicited by the DFP.
- We have interviewed 14 people involved in New Zealand's data ecosystem to develop our thinking and test our developing conclusions. We have also reflected in this paper feedback from the DFP on our work including feedback derived from DFP's engagements with others.
- We have put these pieces together into a long-list of barriers to value generation from data sharing, assessed those barriers against some simple criteria, and discussed them with the DFP as a way to come to a shorter list.
- We have identified three specific barriers for which we will look to develop some ideas on solutions in the next phase of this project.

Some initial conclusions as a result of this work

3. More data sharing could be very valuable (with "value" meaning a wide range of different things). The available literature does not identify exactly where the most value might lie from advancing data use, in terms of sectors or industries.

- On the public sector side, the annual reporting on achievements of the government's open data programme reveals that progress is limited outside of agencies whose core job is producing data, and that 90 per cent of user requests for datasets to be released are not met. The Open Data

Barometer provides guidance on what datasets to release and how: New Zealand has some clear gaps in the provision of machine readable data in general and in datasets on companies, government spending, public contracts and public transport timetables in particular.

- On the private sector side, there is a functioning market for data, and people we interviewed gave us examples of private arrangements for multi-lateral data sharing. There are few examples of private to public sharing as yet, e.g., open sharing of data by corporates for public use.

4. We expect that those who benefit from sharing data will engage in it under their own steam and private arrangements will emerge to get data from those who collect it to those who can use it for something valuable. We see substantial data sharing between government agencies in the social sector, for example, and we were told of data sharing initiatives organised at the industry level in the private sector as well.

5. The barriers that we focus on are situations where these economic incentives seem likely not to lead to the optimal level of data sharing. In particular:

   - Government agencies collect a lot of data but are not designed to respond to market incentives nor to be especially innovative. Public servants are worried that something will go wrong. There is no process for correcting (or indeed expecting) errors, and a feeling that data has to be perfect to be published. Examples of sharing that have worked out well are lacking. Data sharing is something unusual rather than a standard part of all data collection for those agencies whose main job is not data-related.

   - There can be strategic incentives to withhold data. One thing that stops private-sector data sharing is the sense that someone else might get a commercial benefit from the data that I have collected. There are some government agencies whose business models rest on selling data.

   - The uses of data are uncertain and the costs of sharing are very low. This implies, subject to controls for personal and confidential information, that the widest possible sharing of data would be socially optimal. That seems unlikely to emerge from uncoordinated private action because there are private gains from restricting data release and incentives not to share. We found a useful distinction between a passive approach to data publication, in which data is released in response to requests, as in the OIA, versus an active approach, where agencies have processes in place to publish data routinely, as does Statistics NZ, for example.

   - There are many issues that would benefit from coordination between those that collect and share data, but coordination does not easily emerge from individual organisation decision-making. For example, establishing standard formats for address data, or agreeing standard legal templates for data sharing would save a lot of effort across the system.

   - Sharing data can also create new costs and risks. For data that identifies people, data subjects retain control of the uses of data, and there are additional obligations on those collecting and sharing data.

- Willingness to share data depends on trust. Because of the relative levels of trust, data sharing is easier to arrange between two public sector agencies than between two private sector competitors.

6. Put very simply, one major challenge in creating value is getting data from the hands of those that collect it to someone who can use it for something valuable. This is tricky because valuable uses and users that can create value are hard to discover except by trial and error.

7. A second major challenge is making sure that data collectors take some extra effort to make it easier for others to use data that they have collected. There are many coordination issues, e.g., shared data definitions, and data quality issues, e.g., assembling reliable metadata. Organisations lack the incentives, information and in some cases capability to share well-prepared data. Some organisations that are trying are developing their own ad hoc systems to do so, leading to a lack of consistency and duplication of costs.

8. More broadly, and in line with some of the thinking associated with James Mansell, core difficulties around data collection and sharing are related to the way that our institutions are designed. There could be a lot of value from a fundamental rethink, making institutions for the information age. We note that the Australian Productivity Commission has recently released a draft report on its inquiry into Data Availability and Use that calls for substantial institutional change.

Barriers to focus on

9. We isolated three barriers to focus on in the next stage of our work. The reasons that we chose these three are explained in more detail in body of this paper.

   - Each individual agency must create its own bespoke process for sharing data with others. There is no detailed guidance on how, when, and with whom to share data, how to address ethics, confidentiality, and privacy issues, and how to revise or correct errors in data that has been shared. This creates costs and risks that weaken the incentives of agencies to share data.
   - Data that is shared is coded in many different ways, stored in many different formats and systems, definitions are not standardised, and the quality and coverage of metadata is variable. This limits the ability to combine data from different sources for analysis, and imposes costs that weaken the incentives to use data that has been shared.
   - Some government agencies (e.g. NIWA, QV, local councils) have business models based around selling data that severely restrict sharing and use of their data, limit experimentation to discover new uses, and make it difficult to combine their datasets with others.

What comes next

10. We are on the hook to provide a final paper by the end of January that will present some proposals for solutions to overcome these barriers.

Table of contents

Purpose

11. This paper defines what we mean by "value" when we talk about value being generated from data sharing, it broadly defines how such value is created, who is involved in this process, and it explores some barriers to generating more value from data sharing for New Zealand.

Background

12. The DFP asked us to help to identify some "barriers to generation of value for New Zealand from the trusted use of data" and possible ways to reduce or remove them.

13. We have focused our attention on data sharing, i.e., the use by one organisation of data collected by another. We do not consider issues relating to the creation of datasets in the first place, i.e. we are not concerned with barriers to data collection, aggregation or use by the organisation that collected the data. The limits of the social licence as a barrier to data sharing are also outside the scope of our work.

14. Our focus in this paper is mostly on data collected by central government agencies and use of that data by the private sector. We briefly discuss data sharing within government, across departments, and also some incentives for private sector operators to share data. We leave a lot of terrain uncovered for want of time, including sharing of data from local government, academia, CRIs and SOEs.

15. This is the first report of our work, explaining our progress so far. A second report is due by the end of January 2017 that presents some ideas for solutions to overcome the three barriers choose to focus on from those that we identify.

16. This paper is in four stand alone parts:

- A longlist of barriers to greater data sharing with some suggested criteria for considering them and an initial assessment of the barriers against the criteria. This is the main result of our work so far.
- A definition of what we mean by value and a conceptual overview of how it is created. We also define some other terms, including "open data" and "data sharing".
- A brief summary of what we learned from the 14 interviews we have done so far in the course of this project.
- A review of the existing literature and work to date on this topic plus a scan of the available information on how well New Zealand is doing at this.

Barriers to data sharing

17. In this section we consider barriers to data sharing identified from our work, from the literature we reviewed, from our interviews and from discussions with the DFP in the course of this project.

18. We first list the barriers that we have identified, and then we assess them against some criteria that we propose in order to derive a shortlist. We discussed this list with the DFP and took on board some other feedback to isolate the list of three barriers to focus on.

Volume of data available

19. The core challenge is that a lot of data that is collected is not published for reuse.

- Unless it is their core job (Statistics NZ, LINZ) government agencies are not very interested in data sharing (there are not enough carrots or sticks). Compliance with the Declaration on Open and Transparent Government is effectively voluntary.
- Each individual agency must create its own bespoke process for sharing data with others. There is no detailed guidance on how, when, and with whom to share data, how to address ethics, confidentiality, and privacy issues, and how to revise or correct errors in data that has been shared. This creates costs and risks that weaken the incentives of agencies to share data.
- Agencies fear that data release will cause some kind of trouble for them if it is interpreted by others badly or if there are any errors. Agencies think that data has to be perfect to be released at all.
- Some government agencies (e.g. NIWA, QV, local councils) have business models based around selling data that severely restrict sharing and use of their data, limit experimentation to discover new uses, and make it difficult to combine their datasets with others.
- Unjustified fears about privacy rules limit sharing of personal data, and the perceived risk of re-identification limits sharing of even anonymised personal data.
- In the private sector, fears of a loss of commercial advantage or upsetting customers limit data sharing.
- For not-for-profits, the state of their own data and their lack of resources are both barriers to more data sharing.

Usability of data

20. The core challenges are that it is too hard to find data that has been released, and data is not released in the best formats for reuse.

- Data that is shared is coded in many different ways, stored in many different formats and systems, definitions are not standardised, and the quality and coverage of metadata is variable. This limits the ability to combine data from different sources for analysis, and imposes costs that weaken the incentives to use data that has been shared.

- Organisations do not know what data they have or know what condition it is in, in the way they know about their other assets.
- There is a lack of aggregation of data, which means users have to gather it from scattered sources. The Register (data.govt.nz) is not comprehensive.
- Data that is released is mostly suitable for consumption only by experts.

Institutional factors

21. The core challenge is that our institutions are designed for a world where data was hard to share

- Our funding model for government agencies does not support data sharing. Agencies are not funded for it. They struggle to work across organisational boundaries.
- Data sharing requires bespoke legal agreements and complex negotiations between organisations to divide up the benefits. Agencies that want to share data are having to solve the same legal, ethical, commercial, trust and technical problems independently.
- There is growing but still limited evidence of positive stories from data sharing (see case studies of the IDI, Open Data NZ, LINZ, Creative Commons, Digital NZ). In particular, there are few examples of successful private sector/public sector data sharing, and few examples of where things went wrong but everything turned out okay.

Other

- There is still a lot of work to do to get policy to start with data. People are more comfortable talking about assumptions in narrative. Ministers are the same. Data is a second or third thought.
- We treat data analysis as a specialist expertise, which means that decision-making and data analysis are barely connected, let alone integrated.
- General understanding of data is poor and the distinction between data and insight is lost on most.
- Privacy rules (especially the requirement for upfront consent to all uses) limit valuable data sharing.
- People providing data fear that it might be used against their interests, or conversely are too willing to provide data in return for services without thinking through the consequences.

Criteria used in short-listing barriers

22. We roughly scored options as a way to shorten the list on four criteria:

- The expected increase in **value** from removing this barrier (scored 1-5 from least to most value)

- The expected total **cost** of removing this barrier (scored 5-1 from least costly to most costly: in reverse order so that higher total scores are better options)
- DFP ability to **influence** on this issue (score 1-5 from little to a lot)
- Our **confidence** in our estimates of these things (scored 1-5 from uncertain to confident)

23. In response to DFP feedback, we considered also including as specific elements of value the three other Data Futures Forum principles: inclusion, trust and control. We concluded that we did not know enough about the definitions or measurement of “inclusion”, “trust” or “control” to use these terms to score options systematically, and developing a clearer understanding of what they mean was outside the scope of what we could do in this work.

24. Without any detail on the definition of these terms, generally we thought that:

- Releasing more data openly (subject to privacy and confidentiality limits) would boost trust
- Transparency on when and what data is being shared would boost trust
- Enabling more control by data subjects of data sharing would boost both trust and control
- Releasing more usable and detailed data would boost inclusion, since it expands the audience of people that can use data, and it lets people understand and solve their own problems.

25. We do not think that the barriers identified differ greatly in terms of their impact on inclusion, trust or control. An approach focused on inclusion might weight more highly initiatives that boost usability of data. An approach more focused on trust and control might view a lack of transparency in data sharing for personal data as a barrier in itself.

Assessment of barriers against criteria

26. The table below shows the result of a quick assessment of each barrier from 1 to 5 on each of the four criteria, with 1 being relatively bad news and 5 relatively good news. Scores are totalled in the last column. This is an aid to discussion, not a scientific assessment process. The numbers below are an average of independent assessment by each of the two authors.

Figure 1: Assessment of barriers against criteria

	Value (1-5)	Cost (5-1)	DFP influence (1-5)	Confidence (1-5)	Sum (4-20)
Data availability					
Agency interest	5	4	2	3	13
Difficulties with process	5	3	4	4	16
Agency fear	4	4	2	2	11
Agency business models	5	2	2	4	14
Fear of privacy breaches or reidentification	3	3	4	3	12
Private sector fears of loss	4	2	1	2	9
NGO capability	3	2	2	2	9
Data usability					
Data standardisation	5	3	4	5	17
Data governance	5	2	3	3	12
Fragmentation of release	4	4	4	4	15
Aimed only at experts	4	3	3	5	15
Institutional issues					
Agency funding model	4	1	2	3	10
Complex bespoke negotiations	4	3	3	4	12
Not enough positive stories	3	5	5	4	16
Other					
Data driven policy	5	2	3	5	14
Siloed data analysts	4	3	3	3	13
Lack of data understanding	4	3	4	4	14
Privacy rules	3	2	3	3	11
Data subjects' behaviour	3	3	4	3	11

27. Some basic statistics on these assessments follow:

- Maximum 17 (one barriers)
- Minimum 9 (two barriers)
- Average 13

- Median 13

Top six barriers

28. The top six barriers on these assessments are (not in order of priority):

Data availability

- Each individual agency must create its own bespoke process for sharing data with others. There is no detailed guidance on how, when, and with whom to share data, how to address ethics, confidentiality, and privacy issues, and how to revise or correct errors in data that has been shared. This creates costs and risks that weaken the incentives of agencies to share data.
- Some government agencies (e.g. NIWA, QV, local councils) have business models based around selling data that severely restrict sharing and use of their data, limit experimentation to discover new uses, and make it difficult to combine their datasets with others.

Data usability

- Data that is shared is coded in many different ways, stored in many different formats and systems, definitions are not standardised, and the quality and coverage of metadata is variable. This limits the ability to combine data from different sources for analysis, and imposes costs that weaken the incentives to use data that has been shared.
- There is a lack of aggregation of data, which means users have to gather it from scattered sources. The Register (data.govt.nz) is not comprehensive.
- Data that is released is mostly suitable for consumption only by experts.

Institutional issues

- There is growing but still limited evidence of positive stories from data sharing (see case studies of the IDI, Open Data NZ, LINZ, Creative Commons, Digital NZ). In particular, there are few examples of successful private sector/public sector data sharing, and few examples of where things went wrong but everything turned out okay.

29. The scoring system means that these are barriers that are expected to make a big difference to value and where removal is relatively cheap and easy, they are things on which we think the DFP has influence, and we are relatively confident in the assessment itself.

30. The scoring system also means that barriers that are very difficult to change (even if they would have strongly positive impacts) tend to fare worse in the assessment.

31. From discussions with the DFP, we shortened this list of six to three that we will focus on in the next stage of our work:

- Each individual agency must create its own bespoke process for sharing data with others. There is no detailed guidance on how, when, and with whom to share data, how to address ethics, confidentiality, and privacy issues, and how to revise or correct errors in data that has been shared. This creates costs and risks that weaken the incentives of agencies to share data.

- Some government agencies (e.g. NIWA, QV, local councils) have business models based around selling data that severely restrict sharing and use of their data, limit experimentation to discover new uses, and make it difficult to combine their datasets with others.
- Data that is shared is coded in many different ways, stored in many different formats and systems, definitions are not standardised, and the quality and coverage of metadata is variable. This limits the ability to combine data from different sources for analysis, and imposes costs that weaken the incentives to use data that has been shared.

The value of data

32. In this section we explain what we mean by the “value” of data, and discuss how this value is created and distributed. We think this framework is useful to explain some features of the data landscape, and we used it in the previous sections to help identify and analyse barriers to creating value for New Zealand from data.

33. In this analysis we are looking for features of the data sharing ecosystem that might limit the value that New Zealand can secure from data sharing. It has many similarities to the “market failure” analysis that is a standard part of public policy.

What data we are talking about

34. This paper is about data sharing, i.e., the use of data by someone other the organisation that collected it. We do not talk about issues of data collection, except to the extent that what the collector does with the data it collects makes a difference to the usefulness of the data once shared.

35. We are using the word “sharing” in the non-technical sense of the word, i.e., data collected by one organisation being used by another. This is, for example, wider than the meaning of “shared data” on the Open Data Institute’s Data Spectrum, where it is an intermediate step between “closed” and “open” data.[1]

36. We use the term “open data” and “open government data” interchangeably to refer to the publication of data by any government agency. This includes when the data is available to only a small audience, like the IDI, or when it includes everyone. We also talk a little in what follows about the open government data programme, but our conception of “open government data” is much wider than just the efforts under that programme.

37. Data sharing can be one-off or repeated. It can be unilateral (which we also call “data release”), bilateral, or multilateral. We found it useful to distinguish data sharing for a particular purpose, e.g., IRD sharing student loan data with border agencies to enable better enforcement of student loan rules, with multi-party data stores that can be used for many purposes, e.g., the IDI.

Value comes in many forms

38. Value is a broad concept that could include a wide range of benefits for individuals, businesses, government organisations, and society as a whole.

39. The DFP's terms of reference are to "increase the value being generated by New Zealand's data", to "broker and stimulate more data driven innovation", with a principle that "New Zealand should use data to drive economic and social value and create a competitive advantage".

[1] See http://theodi.org/data-spectrum

40. Focusing just on specific types of value makes it easier to measure the scale of benefits and how they are generated, and it simplifies judgements about what counts as "valuable".

41. One set of value is that associated with economic activity, which we call "economic value" and measure through market transactions. Economic value can be generated by business activity, and by the actions of government agencies in markets. Economic value accrues to businesses and their owners in the form of profits and to individuals and consumers in the form of greater wellbeing (or "welfare") from market transactions (which we take to be the difference between what a consumer pays for something and its value to that person).

42. But economic activity is not the only factor that affects people's wellbeing. Other aspects of value derived from data might be improvements in social values (e.g., feelings of community safety from better information on crime), environmental values (e.g., increases in the efficiency of irrigation where water is provided without charge).

43. There might also be value derived through data from improvements in government services. This could be because of better understanding of public policy problems, better targeting of interventions, easier evaluation of existing initiatives, or more transparency in government that leads to greater citizen engagement in the processes of governance.

44. We expect individuals, businesses and government agencies to try to generate and secure value where they can. We therefore focus on barriers to the incentives or ability to generate and capture value from data.

45. Such barriers may arise from unwillingness of data subjects to have their data shared. We treat these issues as part of the conversation around “social licence”, which has been excluded from our remit.

46. Note that the value of data sharing and attitudes or barriers to sharing may differ in different situations. For example, databases of street addresses have many valuable uses and are not considered to be very sensitive for most people (e.g. the electoral roll is published). But if addresses are connected to other data, such as health or criminal records, the value of shared data and the acceptability of sharing this data may be quite different.

47. Also, as we discuss further below, our concept of value in this paper is instrumental, i.e., value comes from use. But one can also conceive of more intrinsic value in data, e.g., value that comes about because the data is somehow associated with a person or an area and is valuable in itself regardless of whether anything is done with it.

48. That said, in this paper we are working under the general assumption that sharing more data between organisations than is done today is valuable (within some limits) and that more sharing could create more value.

How sharing data can increase economic activity

49. "Data-driven innovation" (DDI) is innovation and resulting economic and social value created from data analysis by public and private sector organisations to make better decision and create new products and services (Sapere and Covec, 2015). DDI creates value by:

    - Reducing costs to provide goods and services, e.g. a retailer analysing sales data to avoid holding too much stock of less popular items.
    - Increasing consumers' willingness to pay for existing goods and services through quality improvements, e.g. a media company uses data on viewer preferences to optimise the mix of content it provides.
    - Allowing organisations to make better investment decisions, leading to higher returns on investment, e.g. an electricity network uses forecasts of customer connections to determine when and where to invest in expanding its network.
    - Creating new goods and services, e.g. Uber uses data to create a ride sharing service.

50. We use DDI as an overarching term for the above uses of data. In addition, economic value is generated when any of the following happen:

    - Impediments to market transactions or competition are reduced or removed, leading to a higher volume of market transactions, e.g. an airline analyses travel and trade data to identify new destinations it could serve.
    - People have better information that enables them to make better decisions when buying and selling goods and services and making investments, e.g. buyers on online auction sites are able to check the history and reputation of sellers.
    - Government agencies are able to operate more efficiently, enabling more government goods and services to be provided for the same amount of tax, e.g. social interventions can be targeted to where they are more beneficial.

Who benefits from this increase in value

51. Economic value from the use of a dataset accrues to the user of the dataset, and to the provider of the data if there is a payment for its use. Consumers also benefit from uses of data that create or improve goods and services.

52. Where data is used to improve business decision-making and efficiency, some of these gains will be retained by the businesses and some will be passed to consumers and society as a whole via the process of competition. The extent to which this occurs depends on the intensity of competition in relevant markets.

53. Where data is used to improve government services, the effects are slightly less straightforward: benefits can still accrue to specific individuals, e.g. those who receive better health care as a result of data analysis, but government agencies have more complex incentives than businesses and do not benefit in as direct a fashion from improved performance (i.e., there is no profit incentive).

54. Where multiple parties are involved in putting a dataset to use, each will need sufficient incentives to participate. Some of the value may need to be redistributed away from the immediate user of the data and towards other parties in the "value chain". We expect that "brokers" (e.g. data sharing platforms) will emerge to facilitate these trades. However, brokers will need to overcome coordination problems and find a sensible way to redistribute value, which might be hard in multi-party collaboration.

There are three central ideas for our model of data use

55. If a use of data is valuable, people and organisations will work to realise that value. However, barriers may limit the amount of value that can be generated. These barriers can be understood by analysing the way that value is generated from data.

56. Our model of how the use of data generates value depends on three important ideas:

- Data generates value when it is used, but using data does not use it up and it is hard to know in advance what the ultimate economic value of any given dataset will be. This implies that the greatest economic gains will come if data is shared widely in ways that encourage experimentation.
- Making data usable by others requires extra effort from those collecting data. Without good incentives to make those efforts, insufficient data will be published, and the data that is available will be in formats that are not easy to use.
- Using data also requires complementary inputs, such as the skills and tools to analyse datasets and to understand and act on the results of analysis.

57. The study by Sapere and Covec (2015) estimated that in 2014 DDI was used by only 10-15 per cent of organisations in New Zealand. This was low in comparison with other studies of DDI in Australia, Japan, and Singapore. This suggests that significant barriers exist to the analysis and use of data for decision-making and innovation in New Zealand.

Data only generates economic value when it is used

58. For the purposes of this paper, we assume that a dataset does not have intrinsic value. The value of a dataset, and consequently its economic value, arises only from its use.

59. Using data does not use it up. In fact, if a dataset is used widely and understood by more people, its value may increase as more ways to combine it with other datasets are discovered and more people understand how to use the data. It is also very cheap to store, analyse, and share data.

60. The way we have set up our system, those who collect data control who else can use it and what for (economically this is a compensation for the effort involved in collecting and storing it).

61. No one knows in advance all of the valuable use of a dataset. No one knows who the most valuable user is.

- In many cases, the sources and amount of value that can be created from any given dataset are difficult to predict. Innovation by nature is uncertain and often involves some trial and error.
- The value from a dataset may not come from a specific product or service but by allowing people to make better decisions or by improving business operational efficiency. Such sources of value may not be obvious to outsiders, including the organisation that has collected the data.
- When combining datasets, it is not always clear in advance which combinations will be valuable. The collector of a dataset may not be able to anticipate all of its potentially valuable uses, particularly where this requires combination with datasets from other sources.

62. The core problem for generating value is getting data from the point of collection to someone who can use it for something valuable, and finding as many people as possible who have the skills and tools to make use of it.

63. This implies that sharing data widely is likely to be best for society (the costs of collection are incurred once, but the benefits of use can be generated many times), and that permitting a wide range of uses of data and allowing data users to explore and experiment will be important. In contrast, business models or processes that seek to predict potential uses of a dataset and extract value from those may not realise the full potential value of the data.

64. Given that the additional (marginal) cost of sharing an existing dataset is zero or close to zero, the optimal price for accessing a dataset should also be zero or close to it. However, there can be significant fixed costs in the collection and maintenance of datasets, so the question in practice is how close we can get to this optimum while still providing sufficient compensation to collect and publish data in the first place.

65. The economic characteristics of data also create problems for sharing. Once data has been published in electronic format and shared, it is hard to un-share it, and hard to completely control its use. This may deter organisations that collect data from sharing it, especially if they are nervous about the uses to which the data they have collected might be put. It is also hard to properly value data without first sharing it and discovering how it can be used and the value of those uses.

Datasets are more valuable if they are usable

66. Creating value from data sharing between organisations depends on it being possible, permissible and worth the effort involved. As explained above, the worth of data is discovered through use, but there are pre-conditions to this use.

67. For data use to be possible:

- The data must be available and able to be found by those who can create value from it – Data users need to know what datasets already exist, or have an efficient way of finding relevant data.

- The data must be in a usable format – Most of the effort in any data project is data cleaning and manipulation prior to analysis. This is a fixed cost borne by every user of dataset. If every user has to do this work, then the benefits from use of the dataset will be reduced and some analysts may not bother, since they judge the effort required does not exceed the benefit from being able to use the data.
- The data must be accurate, reliable, and trusted by those who use it – This requires robust systems to be used for recording and storing data, and that data providers maintain a reputation for the quality of their data.
- Sufficient metadata is recorded to enable the data to be understood and interpreted correctly – Metadata provides the context for any data analysis.

68. For data use to be permissible:

- Use or re-use of the data must be allowed by its licence – Without appropriate or clear licencing, the possible uses of a dataset may be restricted.
- Data subjects (if any) must permit the data use – Where a dataset contains personal information, for legal or ethical reasons consent of data subjects to the sharing and use of the data will usually be required.
- Any relevant legislation must permit data use – Legislation can create additional restrictions or obligations on the use of data, depending on the type of data and the use.

69. To best combine datasets together:

- Datasets should contain fields that enable them to be combined. Linking records is easiest if there is a single identifier that is used across all records. Otherwise, concordances or matching algorithms must be developed to combine datasets, which is potentially costly and reduces the quality of linking.
- Metadata must allow people to understand the meaningful ways in which datasets can be combined.
- The relevant licenses, legislation, and regulation must permit combination, and the cost of using the combined data must not exceed its value to a user.

70. Failure to do any of these will impose costs on users. In the worst case, these costs could prevent any use of data at all. For example, licensing that does not enable republishing of data after use will prevent derivative work from being shared. And in practice, pricing of commercial datasets rarely considers the total cost that users face to obtain data from different sources.

71. Put another way, making sure these conditions are met will increase the value of data in use. But putting data into a form suitable for others to use imposes costs on data collectors.

Use of data also requires other efforts

72. A dataset cannot be used on its own. Use also requires some or all of:

- Analytical skills and tools – Data in its raw form usually needs to be processed and analysed to be put to use. Such analysis requires that the analyst have the relevant skills and knowledge, as well as access to software tools.
- Algorithms – For certain kinds of analysis with complex datasets, algorithms may need to be developed or adapted. This can be a costly process.
- Communication and interpretation – The results of data analysis must be communicated to those who will use it to make decisions. This requires people skilled in data visualisation, communication, and interpretation.
- Product and service design and production – If data is being used to create a new product or service, many other parts of the service must be designed, implemented, and produced.
- People, businesses, and government agencies are willing and able to act on the results of data analysis – There are a range of institutional factors that could limit enthusiasm for making data-informed decisions.

73. Providers of these complementary inputs form an important part of the data "ecosystem". Weaknesses or constraints in the provision of any complementary inputs will be barriers to the creation of value from data.

74. Where data is combined from multiple sources, there may be several parties who have to provide data or complementary inputs, each with their own set of costs. This makes data sharing agreements complex, as each party needs to cover its costs and have an incentive to contribute, and the value generated will need to be re-distributed among them. There is a risk that the costs of negotiating such agreements will outweigh the benefits of data use.

Barriers in general

75. The analysis above leads us to some conclusions.

76. We expect that the benefits available from using data will generally motivate those who want to use data to come to arrangements with those who collect it.

77. There are some reasons to think that these economic incentives to share data will not lead to the socially optimal level of data sharing or to the collection of high quality data for all uses:

- Government agencies are not designed to respond to market incentives nor to be particularly innovative, meaning they will under-provide data they collect relative to the hypothetical ideal. Agencies whose purpose is to share data (like Statistics NZ or LINZ) will do a better job.
- Related: provision of high-quality data is a specialist job. Agencies that generate data as a by-product of their work could see it as a sideline or (worse) a distraction from their core business, or an opportunity to earn revenue in excess of their costs.
- There might be strategic incentives to withhold data. One thing that stops private-sector data sharing is the sense that someone else might get a

commercial benefit from the data that I have collected. There are also some government agencies whose business models rest on selling data.

- Because the optimal marginal price for data is zero or near zero, data collectors may not benefit from making it available or may have to charge more to cover their fixed costs. Any additional efforts required to make data usable by others may also go uncompensated.
- There are many issues that would benefit from coordination between those that collect and share data, but coordination does not easily emerge from individual organisation decision-making. For example, establishing standard formats for address data, or agreeing standard legal templates for data sharing would save a lot of effort across the system.
- Sharing data can also create new costs and risks. For data that identifies people, data subjects retain control of the uses of data, and there are additional obligations on those collecting and sharing data.
- Willingness to share data depends on trust. Because of the relative levels of trust, we would expect it to be easier to arrange to share data between two public sector agencies than between two private sector competitors.

78. In general, if we want to get more value from data sharing, then we could:

- On the supply side: reduce the costs and risks of making data available to others, more effectively direct government agencies to share data openly, relook at strategic incentives for sharing data that is being sold by government agencies, and compensate those who collect and share data for their efforts in some way other than having them sell data directly.
- On the demand side: make it easier for people to find relevant data, or boost the value of data that is found by making data easier to use.

79. There is no obvious institutional problem with a lack of relevant technology or skills or commitment to data-driven decision-making. If data-informed decisions are better, then we would expect more data-informed decisions will be made over time. This is not to say that efforts to boost the number of people with relevant skills or to encourage data-informed decisions would not help, but just that our analysis says they are unlikely to be the most pressing barrier to resolve.

Literature review

80. In this section we summarise the literature we collated and reviewed as part of this project. A list of references is attached. We also briefly review the available information on how New Zealand is going with progress with its open government data programme and mention some public feedback to DFP on its Diagnose and Fix workstream, of which this project forms part.

81. In general, the literature is quite voluminous but rather underdeveloped. In particular, despite the popularity of government open data programmes, there is little literature assessing them comparatively, in a policy sense or in a quantitative way, except for the question of whether government agencies should charge for data (answer: no).

The value of data in general

82. OECD (2013) links the increased value of data to changes in the ease with which data is collected, analysed and shared (page 4):

> "Economic and social activities have long relied on data. Today, however, the increased volume, velocity and variety of data used across the economy, and more importantly their greater social and economic value, signal a shift towards a data-driven socioeconomic model. In this model, data are a core asset that can create a significant competitive advantage and drive innovation, sustainable growth and development."

83. Taking better advantage of data links at the highest level with improved national prosperity. OECD (2013) sees capability in data use as a type of Knowledge-Based Capital (KBC) and smart use of data as a way for countries to create significant national competitive advantage. The Productivity Commission (de Serres et at, 2014) says that New Zealand's low levels of investment in KBC are a major reason for New Zealand's relatively poor national economic performance over the last forty years.

84. There are many different ways to think about the value of data. We found material on:

- The economic value of data
- The value of data for improving government services
- The value of data for redistributing power within society
- The value available from data in different sectors, for different types of data or for different data-use environments
- The value of specific datasets.

Economic value

85. Economic value is generated from data in diverse ways. Referring just to data collected by the public sector, the OECD (2015, page 5) says:

> “Public sector information (PSI) can be used directly to generate products and services, and it contributes in a wide variety of ways to improving efficiency and productivity across the economy.”

86. Data enables businesses to be better informed, and therefore to make choices with larger or more certain economic payoffs. Benefits for existing businesses come about from both cost savings and revenue increases. Glass et al (2014) cite a long list of possible routes to value including enhanced marketing and sales, improved customer interactions, more efficient advertising, easier recruitment, better management information, leaner processes and more efficient production, and new organisational forms that become possible when data is more easily gathered and shared. OECD (2013) differentiates between data-driven research and development, data-driven products, data-driven processes, data-driven marketing, and data-driven organisations.

87. Data also enables businesses to create new goods and services where the data itself is the product or a major component of the product (OECD 2013). There are also thought to be economic benefits from the creation of new businesses based on wider availability of especially government data. OECD (2015) describes these as “clearly seen, although [the] evidence is mainly anecdotal”. Zuiderwijk (2014) says that very little research has been done on how exactly open data can result in innovation, although there is some evidence describing various business models that have emerged.

88. The economic value available from the use of data is thought to be high.

    - McKinsey Global Institute (2013) estimates that greater access to open data could add $3 trillion annually to global economic activity, equivalent to boosting world trade by 16 per cent.
    - Dekkers et al (2006) estimate the size of the market for public sector information in the EU27 plus Norway at between €10 billion and €48 billion, around one quarter of one per cent of GDP.
    - PIRA (2000) estimated the economic benefits of commercial use of public sector information in Europe were €68 billion (around one per cent of EU GDP). Vickery (undated, but likely 2011) updates that estimate to €140 billion a year or 1.2 per cent of EU27 GDP.
    - OECD (2015) estimates that the economic benefits of public sector data use were around USD 500 billion in 2012/13 or 1.5 per cent of OECD GDP, with growth in the market for public sector information of between 6 and 18 per cent a year.

89. There are also valuations of more granular aspects of data use:

    - The value of Data Driven Innovation (DDI) has been estimated in Singapore, Japan, Australia and New Zealand. DDI was found to contribute between 1.4 per cent and 4.4 per cent of GDP or GVA (AnalysysMason 2014a, AnalysysMason 2014b, PWC 2014, Sapere and Covec 2015). New Zealand is at the lower bound. Although there are some methodological differences, the main reason for the impacts of DDI in New Zealand being so much lower is the much lower estimated takeup of DDI.

- Fioretti (2010) cites a £6 billion boost to UK GDP from that government's open data plans, €1.6 billion in market revenues for geo-spatial information in Germany in 2006, and around €15 million in direct financial benefits to Denmark from open publication of that country's official address database.
- Missingham (2009) quotes an Australian National Audit Office study that concludes that the federal government's information assets could be worth as much as $3.9 billion.
- Closer to home, Bakker (2013) estimates the value of the use of Census and associated population statistics to New Zealand at $1 billion over 25 years. And ACIL Tasman (2009) put a figure of $1.2 billion on the value of geospatial information for New Zealand in 2008.
- At the firm level, Barua et al (2011), Brynjolffson et al (2011) and La Valle et al (2011) all show improved financial performance for firms that use data effectively.

90. More broadly, there are studies of the impact on productivity of investment in or use of Internet services. Generally these show that the Internet is a strong force for productivity growth and business success (Varian et al 2002, Micus Consulting 2008, McKinsey Global Institute 2011b, MYOB 2011, Deloitte Access Economics 2013, Grimes et al 2012, Statistics NZ 2013, Glass et al 2014, Cap Gemini undated). Koski (2011) shows that architectural, engineering and technical consulting firms grow about 15 per cent faster in countries where public sector agencies provide basic geographic information for free or at marginal cost. The growth effect is visible a year after a switch to lower pricing but stronger after two years. SMEs benefit most strongly, suggesting that marginal cost pricing lowers their barriers to enter new market areas and compete with larger firms.

Government services

91. There are thought to be significant improvements in the efficiency of government services available from better use of data. OECD (2013) reports some European governments saying cost savings could reduce administrative expenses by 15 to 20 per cent. Venkatraman (2010), writing at a time when fiscal austerity was a major political focus in Europe, says the main attraction of open access to government data is cost savings. The Canterbury District Health Board reports substantial improvements in services as well as better value for money from use of integrated health data (Sapere and Covec, 2015). Bill English, then Minister of Finance, said in September 2016:[2]

> "it's hard to overstate the value of what's being revealed to us for the first time through the use of the IDI and other data stores in government"
>
> "For instance, it means we can understand how changes in policy or funding in one part of the social sector affects spending and outcomes in other parts of the social sector months or years downstream based on real events."

[2] https://www.beehive.govt.nz/speech/social-investment-analytics-layer-launch

Wider social value

92. Susha et al (2015, page 182) refers to wider social benefits from open government data, including improved citizen participation, public engagement and self-empowerment of users. It can:

> "enable individuals to make better decisions in their lives and increase participation in public affairs. Besides, by innovating with the data, citizens become active contributors and designers of content and services, thereby entering into a more participative and empowering relationship with the government."

93. But she goes on to point out that "evidence of any such effects ensuing from [open government data] initiatives is scattered, not well-understood and at times even contradictory", citing a paper by Peled that concludes that the US open data initiative mostly empowered those who already had the money and skills to use data.

94. Jetzek (2015, page 2) focuses on the benefits of Open Government Data (OGD) for transparency in government and redesign of institutions:

> "OGD implies that the public sector relinquishes its role as information gatekeeper in lieu of a new role as information publisher. Thereby, OGD involves a realignment of the power dynamics between the public and private sectors. Proponents of OGD hope that such shifts will readjust the power balance between government and citizenry and subsequently strengthen democracy and improve government work through increased participation, collaboration and transparency. OGD advocates are also motivated by the potential of open data for promoting innovative entrepreneurs, who can use the open data to propel economic growth as well as to address social challenges. Moreover, advocates of OGD argue that it enables greater government efficiency through an information infrastructure that allows for better data re-use within the public sectors and inter-agency coordination."

95. In a similar vein, Mantelero (2014) draws a connection between access to information and societal power held by public and private entities. He distinguishes between cultural barriers "such as [a] low level of education, technical language, linguistic barriers", and restrictions on access, including "access to libraries, copyright, trade secrecy, documents, and databases owned by governmental agencies not available to the citizens or to any citizen."

96. He says (page 1):

> "These limits create an asymmetric distribution of information and knowledge in society, which implies an unequal distribution of the opportunity to use them to understand, predict, and manage the different aspects of social interaction, from business to government policies, from technological innovation to science."

97. He sees that despite having more information collected and available, there is neither more knowledge available nor more diffuse access to information. Indeed he sees "increasing concentration of the control over the information in the hands of a limited number of private and public entities".

Value in different sectors

98. Some sectors are thought to be more important than others:

- McKinsey Global Institute (2013) focuses on the benefits of greater access to open data in education, transportation, consumer products, electricity, oil and gas, healthcare, and consumer finance.
- In the Japan DDI study AnalysysMason (2014b) sees relatively small impacts in the health, education and social services sector (although non-market impacts might be larger there), and much stronger impacts in transport and logistics, and manufacturing. It focuses in particular on the value of "services for here and now", like mapping services or fleet management systems, and "intelligent planning", like analysis of historic traffic congestion data to aid in urban planning.
- In a similar vein, the largest impacts of DDI in New Zealand (Sapere and Covec, 2015) are in the transport and logistics, retail, and other services sectors, reflecting higher takeup of DDI in those sectors.
- OECD (2013) focuses on benefits in online advertising, healthcare, electricity generation and distribution, transport and the public sector, sectors that accounted for around a quarter of total economic activity in OECD countries in 2010.
- De Saulles (2007, page 10) cites estimates that the public sectors of Europe and North America are the largest generators of commercially valuable information "derived from a variety of sources including regulatory bodies, statistical offices, mapping agencies and treasury departments".

Value from different types of data

99. Some types of data are also thought to be more important. Drago (2009) says that geographic data is the most important, both because it is the largest category and because it is responsible for half of all the value available from reuse of public section information.

100. Zakaria and McBride (2000) say that more valuable data is:

- Data that is comprehensive – Public authorities have the advantage of drawing data from large populations, either from compulsory information gathering or by providing very popular services.
- Data that has long time series – Many public data sets have been collected on comparable bases for decades.
- Data that is accurate – Public authorities expend significant effort in data collection and analysis.

101. We can use the ten criteria of the Open Data Barometer to show what data characteristics might matter most (World Wide Web Foundation, 2016):

- Does the data exist?
- Is it available online from government in any form?
- Is the dataset provided in machine-readable formats?

- Is the machine-readable data available in bulk?
- Is the dataset available free of charge?
- Is the data openly licensed?
- Is the dataset up to date?
- Is the publication of the dataset sustainable?
- Was it easy to find information about this dataset?
- Are (linked) data URIs provided for key elements of the data?

Value from the data use environment

102. Susha et al (2015) note that citizens can be involved in three different ways:

- more passively, as the end-users of newly developed services based on open data, or the objects of development
- in an in-between way, when citizens are represented in the open data process by intermediary organisations, and
- more actively, when citizens act as the subjects of development, or collaborators in the production of new services.

103. Jetzek (2015, page 4) distinguishes between the economic value of different datasets and the amount of external (to the public sector) participation in value creation. The table below reproduces the thinking: datasets that are high in economic value and involve high levels of external participation are candidates for the creation of new services and businesses; those that are low on economic value and external participation are relevant to public sector transparency. This is a conceptual rather than an empirical model.

Table 1: Types of value as a function of dataset use

	Economic value	
External participation in value creation	Low	High
Low	Transparency	Participation and collaboration
High	Efficiency of public services	Creation of new services and businesses

Source: Jetzek (2015)

The value of specific datasets

104. The 15 categories that the World Wide Web Foundation (2016) uses in its Open Data Barometer provide a ready reckoner of which are the most important datasets to make available:

- Mapping data
- Land ownership data
- Key national statistics
- Detailed budget data
- Government spending data
- Company registration data
- Legislation data
- Public transport timetable data
- International trade data
- Health sector performance data
- Primary and secondary education performance data
- Crime statistics
- National environmental statistics
- National election results
- Public contracting data

105. The ODB clusters these datasets into three groups:

- innovation (maps, public transport, trade, crime statistics, public contracts) and is expected to be of business value for building open data applications
- social policy (health, education, environment, Census data), useful for social policy development and inclusion
- accountability (land ownership, legislation, elections, government budget and spending data, company registration), and is used for public accountability.

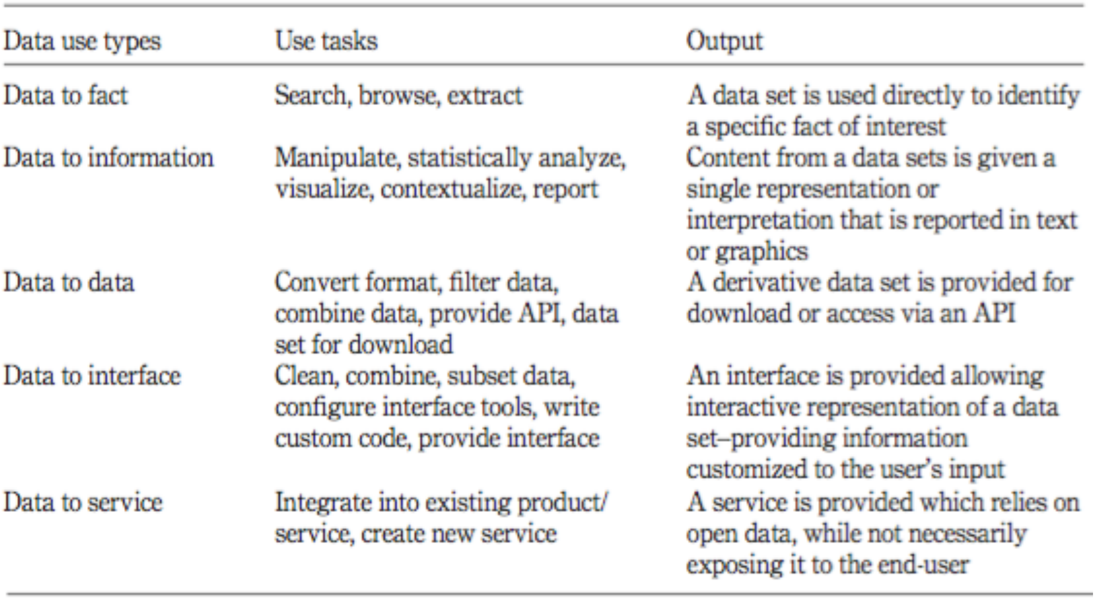
106. Dekkers et al (2006) cite 28 areas with the most relevant data to make available based on directives of the European Union, divided into six domains: company information, geographic information, legal information, meteorological information, social data, and transport information.

107. Weiss (2002) explores the value of wide dissemination of meteorological data. He cites examples from Switzerland, Germany, the Netherlands, the United Kingdom, Sweden, Finland of public meteorological bodies restricting access to data they collected in order to hobble private sector entities that were competing with some of their commercial efforts.

The value of different types of use

108. Davies (2010), as cited in Susha et al (2015), categorises five different types of uses of open data.

Figure 2: Five data use types



Data use types	Use tasks	Output
Data to fact	Search, browse, extract	A data set is used directly to identify a specific fact of interest
Data to information	Manipulate, statistically analyze, visualize, contextualize, report	Content from a data sets is given a single representation or interpretation that is reported in text or graphics
Data to data	Convert format, filter data, combine data, provide API, data set for download	A derivative data set is provided for download or access via an API
Data to interface	Clean, combine, subset data, configure interface tools, write custom code, provide interface	An interface is provided allowing interactive representation of a data set–providing information customized to the user's input
Data to service	Integrate into existing product/ service, create new service	A service is provided which relies on open data, while not necessarily exposing it to the end-user

Source: Davies (2010)

Methods for valuation

109. From our review, there is no straightforward or accepted way to value the use of data in practice. Most of the studies we came across that included quantitative estimates were built from top-down views based on a sample of impacts or assumed impacts.

110. Bakker (2013) estimated a low-end value for the Census based on the uses to which Census information is primarily put (e.g. funding of education services). Use by local and central government organisations, as well as the private sector were examined, but specific types of applications of data innovation by consumers and businesses got relatively limited attention.

111. The DDI studies were based on assumed values for cost reductions and revenue increases and takeup in different sectors, with the assumptions derived from relevant statistics and interviews or surveys.

112. Other methodological contributions focus on assessing the sorts of policy settings that are thought to be correlated with value. Good examples are Nugroho et al (2015) and Nugroho (2013), papers that compare the national open data policies of the USA, UK, the Netherlands, Kenya, and Indonesia. The authors found that the main emphasis has been on data release, with less attention paid to the use and re-use of data, and no measures of the value of open data were found in any country. However, the USA and UK, which are more advanced in their open data programmes, were said to have now reached a stage where generating value from the use and re-use of data that has already been published is receiving more attention from policymakers.

113. As a starting point for addressing the measurement issues, Ubaldi (2013) provides an analytical framework for the evaluation and empirical analysis of open government data (OGD) initiatives. This framework includes collecting data on OGD initiatives in terms of:

- The overall vision, governance, and institutional framework;
- The legal framework and policy environment;
- Technical barriers to the publication of open data;
- Economic and financial benefits and costs, and the health of the "ecosystem" in which open data is used;
- The extent of organisational and cultural change that has occurred in the public sector to support the release of open data;
- The extent of communication between users of open data and the public sector on the quality and quantity of available data; and
- The impact of the OGD initiative in terms of:
    - Concrete measures of economic and social impacts
    - Measured effects on collaboration and innovation
    - Statistics on open data users and dataset usage
    - Steps taken to encourage more advanced use of open data (beyond data delivery).

114. In terms of the commercial value of data, Taylor (2012) summarises an interview with Gartner research vice president Doug Laney about the treatment of data and information as a corporate asset. While there is no standard for calculating the asset value of data held by a business, Laney notes that value could be based on the difference in performance resulting from the use of data and information versus not using it. Such values could be measured (in a large organisation) through the use of controlled experiments.

115. Viscusi et al (2014) focus on the social value created from open data and provide a framework for classifying datasets in terms of their social value. Rather than providing a single measure of value, this framework suggests a number of metrics that can be used to understand the contribution that publishing a dataset makes to social welfare or wellbeing. The metrics are specific to each dataset, and are intended to reflect the potential social value created by that dataset. For example, indicators relevant to a safety-related dataset could include crime and accident rates. These do not measure the contribution of the dataset to reducing crimes and accidents, but indicate its potential significance in terms of specific measures of wellbeing. Viscusi et al do not attempt to aggregate these indicators, but discuss how they can be collected and presented in a systematic way.

116. There are other papers along similar lines:

- Vaish et al (2011) suggest that the value of a published dataset is proportional to the number of times that it is used, the duration of time over which it is available, and its accuracy (determined through a subjective survey).

- Prime (2015) suggests that the commercial value of a dataset can be determined by the number of times it is used, the frequency of its use, the variety of its uses, and its accuracy.
- Englesman (2007) suggests a number of other factors in commercial valuation of datasets, including whether or not they can be combined with other datasets, the ease of sharing data, the flexibility of use and re-use, the formatting, and timeliness.

117. Morris et al (2014) suggest that methods for assessing damages in IP litigation might be useful for valuing data:

> "the courts consider the extent to which information is disclosed internally, the extent to which the information is shared with outsiders, and the extent of efforts to guard the secrecy of the information in determining damages due to misappropriation of trade-secret information. Accordingly, trade-secret data that an organization shares with 600 people are assumed to be less valuable than data that are shared with only six people."

Barriers

118. We found quite a large number of papers on barriers to data sharing, although nothing that weighs up barriers or compares them. We classify them into:

- General approaches to defining barriers
- Privacy related barriers
- Some specific barriers to sharing in the public and private sectors respectively
- Questions of funding
- Some risks created by sharing that operate as barriers.

In general

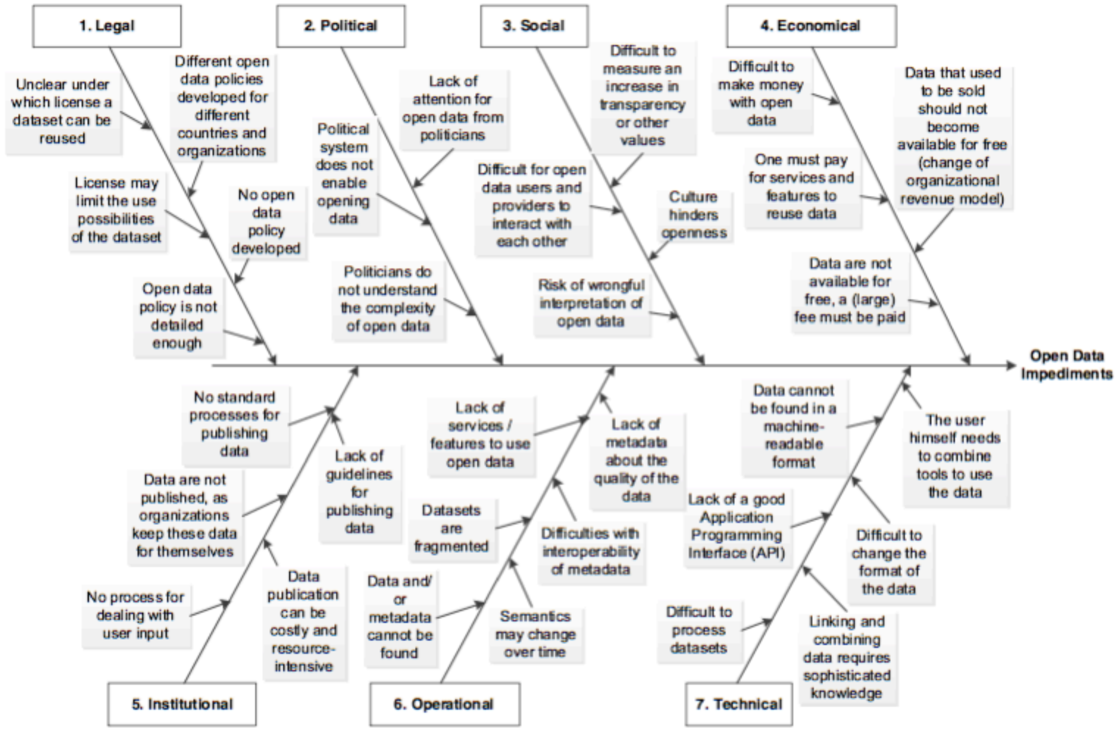
119. Zuiderwijk (2014) classifies barriers to open data into political, social, economic, institutional, operational, technical and legal matters. There are many ways to think about which of these matters most. A diagram from the paper is reproduced below. She cites previous work documenting 118 barriers to open data:

> "The barriers were divided into ten categories, namely 1) availability and access, 2) findability, 3) usability, 4) understand-ability, 5) quality, 6) linking and combining data, 7) comparability and compatibility, 8) metadata, 9) interaction with the data provider, and 10) opening and uploading. "

120. Unsurprisingly, Zuiderwijk said that getting value from open data is complex (page II):

> "The complexities emerge from several factors, including the large number of actors involved in the process, the variety of social and technical contexts, uncertainty surrounding how open data will be used, and the difficulty valuing intangible impacts generated through open data innovation. "

Figure 3: Framework for thinking about barriers to open data


1. Legal
2. Political
3. Social
4. Economical
Unclear under which license a dataset can be reused
Different open data policies developed for different countries and organizations
License may limit the use possibilities of the dataset
No open data policy developed
Open data policy is not detailed enough
Political system does not enable opening data
Lack of attention for open data from politicians
Politicians do not understand the complexity of open data
Difficult to measure an increase in transparency or other values
Difficult for open data users and providers to interact with each other
Culture hinders openness
Risk of wrongful interpretation of open data
Difficult to make money with open data
One must pay for services and features to reuse data
Data that used to be sold should not become available for free (change of organizational revenue model)
Data are not available for free, a (large) fee must be paid
Open Data Impediments
No standard processes for publishing data
Data are not published, as organizations keep these data for themselves
No process for dealing with user input
Lack of guidelines for publishing data
Data publication can be costly and resource-intensive
Lack of services / features to use open data
Datasets are fragmented
Data and/ or metadata cannot be found
Semantics may change over time
Lack of metadata about the quality of the data
Difficulties with interoperability of metadata
Data cannot be found in a machine-readable format
Lack of a good Application Programming Interface (API)
Difficult to process datasets
The user himself needs to combine tools to use the data
Difficult to change the format of the data
Linking and combining data requires sophisticated knowledge
5. Institutional
6. Operational
7. Technical

Source: Zuiderwijk (2013)

121. Janssen (2011) is more specific: the main barrier in Europe is the “general lack of awareness of the public sector about the benefits and risks of opening up their data”. Public agencies “feel that the data they have collected for their internal purposes do not have much value for the outside world, or they are worried that the data are of insufficient quality to be disseminated or that the data might be misused or misrepresented harming their reputation”. There are also challenges with public agencies that are required to operate commercially, who have incentives not to share the data they collect with firms that might compete with them in downstream markets.

122. Janssen also points to practical challenges: the difficulties of finding re-usable data (because portals are under-developed or data is hidden away), unclear licensing or pricing regimes, a lack of metadata, or data only being available in the wrong formats. This is often because agencies lack the time, skills or experience to deal with these issues.

Privacy issues

123. Scassa (2014) identifies three privacy challenges that need to be resolved before open government programmes can succeed:

- Balancing the goal of open and transparent government with privacy, given the volume of personal data that governments have, and the ease with which such data can be accessed and combined if it is published online.

- A blurring of the formerly well-defined boundary between the public sector and the private sector, leading to less clarity about who is responsible for protection of personal information and what acceptable uses are.
- The increasing volume of data that is publicly available, increasing the scope for re-identification of individuals within anonymised datasets when datasets are combined with others. (We talk about this issue more below).

Sharing in the public sector

124. Yang and Maxwell (2011) look at factors that affect information sharing within and between public-sector organisations. Within organisations, culture, beliefs, and attitudes are the key determinants of the extent and quality of sharing. Between organisations, sharing may be promoted or hindered by technological, political, and organisational/managerial issues.

    - Technological barriers may arise due to the use of different types of hardware and software in different organisations, as well as incompatible information security practices, and issues created by outsourcing of IT functions.
    - Political or organisational barriers may arise from legislation and policy differences between organisations, as well as a lack of desire to collaborate, differences in cultures, values, and experience, (lack of) resourcing, competing interests and resistance to change, concerns about loss of control and autonomy, lack of trust, lack of support from leadership, and the difficulties associated with negotiating agreements to share information.

125. de Rosnay and Janssen (2014) also look at challenges for open data publication across the public sector. Key obstacles identified include overlapping and conflicting laws and regulations regarding open data, unresolved issues relating to privacy and data protection, barriers created by intellectual property rights and licensing, resistance from private sector businesses that earn money from data publication, and fears about the exposure and liability created by data publication. The authors suggest that a common legal and institutional framework and detailed implementation guidelines for open data schemes are required to address some of these barriers.

126. Charbonneau and Nayer (2012) focus just on the use of benchmarking information by managers within local government, based on responses to a survey in Quebec. The most significant barriers identified were the lack of usefulness (or perceived usefulness) of benchmarks by managers, the difficulty for small municipalities with limited resources to analyse benchmarking data, lack of support or understanding by elected officials, and lack of resource including time and people with appropriate expertise.

Sharing in the private sector

127. Wiewiora et al (2014) consider factors affecting the sharing of knowledge between projects within a private-sector organisation. Via a series of case studies, they find that organisational culture affects the extent and types of knowledge that are shared, and whether sharing is tacit or explicit. They also find that where there is a higher level of trust between people within an organisation,

knowledge sharing is deeper and more fruitful. This suggests that organisations need to foster a culture where knowledge sharing is supported and promoted, and where internal trust is high, in order to make effective use of internal data sharing.

Funding issues

128. Kitchin et al (2015) consider fourteen different funding models for open-access data repositories, including simple core funding from public funds, public-private partnerships, philanthropy, and various commercial or semi-commercial models including selling data, research, and other services. The authors note that funding is a critical issue for the success of open data programmes, not only to cover costs but to provide sufficient incentives for contributing data, compensating owners of intellectual property, and building up sufficient expertise within public sector organisations. Lack of a robust funding model raises the risk of financial failure as well as reputational damage, which may be a barrier to future open data efforts. The authors argue that core public funding is the most robust model, as all of the other models lead to risks and trade-offs that could undermine the creation of data platforms that are truly open and are free to use.

Risks created by sharing

129. Sharing data might lead to unauthorised disclosure. Morris et al (2014) quantify the scale of data breaches (in US dollars), and highlights the scale of the risks of sharing with third-parties:

> “In 2011, the Ponemon Institute conducted a study of data breach cases and found the average cost per compromised record was $194. The total cost per breach incident in their sample ranged from $750,000 to $31 million. In addition to direct costs related to repairing and mitigating the damage, such as legal expenses, organizations also experienced indirect costs resulting from loss of customer confidence and reputation. In the incidents studied, 42 percent were caused by third-party organizations with which data were shared”

130. There is particular concern about sharing personal data, and whether data that is thought to be anonymous turns out not to be in practice, particularly if combined with other datasets.

131. Sweeney (2000) shows that 87 per cent of respondents to the 1990 US Census are unique in the population based just on their 5-digit ZIP code, gender and date of birth, data that was readily available publicly. Nearly one fifth of the population is unique based on their county of residence, gender and date of birth. Golle (2006) is a similar study for the 2000 US Census. Zhang and Bolot (2011) show similar results for cellphone call data. De Montjoye et al (2013) study fifteen months of cell phone data for one and a half million individuals and find that, in a dataset where the location of an individual is specified hourly at the cellsite level, four data points are enough to uniquely identify 95 per cent of the individuals.

132. Being identifiable in theory is not the same as being identified in practice. But the risks of de-identification are demonstrated by Sweeney (2002), who reports being able to identify some medical records as relating to the then-governor of Massachusetts by combining together an anonymised public employee health

insurance database with public voter registration data and using the governor's (publicly known) ZIP code, gender and date of birth.

133. The Australian Productivity Commission (2016, page 8) noted that risks related to the potential to inappropriately identify people or businesses from datasets included:

- "discrimination
- loss of control over the boundaries around the 'you' that the world sees
- reputational damage or embarrassment
- identity fraud
- other criminal misuse of the data
- commercial harm."

134. The Commission also noted the need not to overstate the risks. Most personal information used in identity theft is obtained online through theft or hacking or phishing, not through data sharing. It goes on (page 11):

> "In reality, most risks of data misuse arise not through the public release of robustly de-identified data, but rather from poor or outdated data collection, storage and management practices, often coupled with malicious intent to gain access and use data that would otherwise not have been available."

How to boost value

135. There is thought to be a lot of value from removing barriers to data sharing. OECD (2015) reckons the economic impacts of the use of public sector information would be 40 per cent higher if barriers to use were removed, skills enhanced and the data infrastructure improved. This would add USD 200 billion in additional gains to the USD 500 billion existing economic impacts. This assessment builds on that of OECD (2013), which says that a new approach to privacy protection, more open access to data, increasing availability of needed skills, better technical infrastructure, and improving the evidence base on what works would be helpful.

136. Vickery (nd) has similar numbers but just for Europe. He suggests that moving to a more open data stance, with easy access to data for free or at marginal cost could add 40 per cent to the value of open data, making open data worth around €200 billion in 2010, around 1.7 per cent of EU27 GDP. He cites substantial cost savings through efficiency gains for public bodies from open data, but notes that revenue from data users is likely to be only modest because the greatest benefits come from the widest possible distribution of data.

137. We found a number of papers on how to improve the environment for data sharing, but we agree with Nugroho et al (2015), who note that "there is hardly any research systematically comparing open data policies". We group the ideas we found into:

- Improving the open government data programme

- Reducing pricing of public data
- Standardising data collection and release
- Making changes to privacy rules
- Some proposals for technical innovations, and
- Fundamental institutional reform.

138. There is also a set of papers on making sure that options for change reflect the issues that are actually need to be addressed and the objectives being pursued.

Improving open government data programmes

139. At a high level, Dragos and Neamtu (2009) suggest three important institutional features to generate value from open data:

- a central government body that is focused on encouraging reuse,
- an asset list maintained by all public sector bodies of all data they hold for reuse, and
- a prohibition on exclusive arrangements for data provision by any public agency unless that arrangement is somehow necessary to provide a service to the public.

140. Nugroho et al (2015) suggest:

- making sure the law requires “continuous release” of data from government bodies, including regulating the minimum number of data sets each government organization should release regularly.
- Requiring that data be machine-readable and accessible through an open data portal, published in open formats and licensed for reuse as well as having appropriate metadata.
- A way to disseminate the data that are published to all citizens is to present it in an understandable format or in usable applications that more people can relate to.
- Enabling ways to enable public interaction on data and feedback to improve quality.
- Requiring checks on quality before publication
- Assigning the implementation of open data policies and an ICT infrastructure to a designated agency or taskforce

141. Susha et al (2015) provides a wide range of best practices for those involved in publishing open data drawn from a review of the literature and case studies, focusing primarily on ensuring strong engagement with users of open data as part of the process of design and operation of open data systems. The authors suggest that the choice of what data to publish and how should be driven from real problems that users face, and done in a way that enables real-life problem-solving by users of the data.

142. She notes a gap between the theory of open data (where it was enough just to publish something) and the reality (where in fact a lot more effort is required to turn open data into something of value.

> "In practice and research, the emphasis has so far been on how to open data and much less on how these translate into services of public value when used. The insights into user perspectives are largely lacking, which contributes to a gap between the promises of open data and the real-life use."

143. Similarly, Kaschesky and Selmi (2014) review progress on open government data and develops a framework

> "To yield measurable impact, it is not sufficient to just make data "available"; it must become "actionable"…. government agencies should publish the data proactively with a mission to attract uptake. That requires data stewardship and a data strategy."

144. Parycek et al (2014) report on a survey of those involved in open government efforts in the city of Vienna. Operational solutions to generate more value include an easy way for organisations to share data externally, common data standards and definitions, appropriate solutions for organisations that lose revenue from not being able to charge for open data, showcasing potential usage to users, and publishing more data.

145. Petychakis et al (2014) present the results of a survey of (more than 150!) open data portals across the EU in terms of their licensing, languages, how they acquire data, how data is catalogued and able to be searched, how data is provided to users, what formats are available and what categories of data are available.

Pricing of public data

146. There is quite a substantial literature on the appropriate pricing of public sector information in Europe.

147. PIRA (2000) and Dragos and Neamtu (2009) contrast the EU and USA approaches to open data, arguing that the EU approach of cost recovery is holding it back. PIRA (2000) says that the USA market for public sector information is between two and five times the size of the EU market and that this is because of the charging approach in EU, where information is often seen as an asset to be exploited by public sector agencies. The US has "a strong freedom of information law, no government copyright, fees limited to recouping the cost of dissemination, and no restrictions on reuse". PIRA's figures show that Europe's less liberal approach generated a return to the national economies of seven times initial investment in public sector information, whereas in the USA investment returns a multiple of 39 times.

148. Deloitte Access Economics (2011) focuses on the value of improving access to information held by public sector organisations. It is based on a review of practices in 21 public sector bodies across Europe, endeavouring to determine which charging model for publicly held information works best. The authors conclude that lower charges are generally better. There are very substantial

increases in re-use of public data from reduced charges and more open availability. Barriers to lower prices include dealing with the loss of revenue from lower prices (although in some cases lowering prices increased revenue because of the extra demand that it created), and how to manage the transition from existing arrangements that involve a commercial partnership with a single company that then has incentives to delay or prevent a move to more open data.

149. De Saulles (2007) points out that a conflict of interest exists for public agencies that charge for data between their commercial activities and public responsibilities. He argues that this has a wider negative impact on the economy, "with short-term financial interests of the public sector restricting innovation within the private sector".

150. OECD (2015) calculates the welfare gains to member countries from moving from an average cost or cost recovery pricing model to marginal cost pricing for public sector information at USD 145 billion. These positive gains stem from removing widespread disincentives to use data, including high prices, restrictive licensing practices, differences in licensing systems across national institutions, lack of information, and poor interoperability. They say (page 6):

> "it is clearly demonstrated that government revenues foregone by moving to open PSI are very largely outweighed by government benefits including public sector productivity gains, more effective service delivery, improved policy development, cost savings through common data access, etc., and outweighed even more by wider economic and social benefits.

Create data sharing templates and standards for data collection

151. Kaschesky and Selmi (2014) note:

> "Data models for different systems are arbitrarily different. The result of this is that complex interfaces are required between systems that share data. These interfaces can account for between 25% and 70% of the cost of current systems. Data cannot be shared electronically with customers and suppliers, because the structure and meaning of data has not been standardized".

152. Standardisation comes with advantages, such as reuse and compatibility, and disadvantages, such as additional costs and constraints for those that collect data.

153. Higgins et al (2014) point to the lack of commonly available frameworks for data sharing between public sector organisations (in the UK). It reports on a two-year project to share information between the fire and rescue service, the local council, the NHS primary care trust, and the police force in one (undisclosed) area of the UK. The project aimed to ensure that data gathered concerning 'at risk' individuals, and social groups within the region studied was more effectively processed and disseminated, and to reduce duplication of effort between the fire and rescue service, the local council, the NHS primary care trust, and the police force.

154. Zuiderwijk et al (2014) looks at how to design a process for the publication of open data. The authors note that publication is often cumbersome and there are

no standard procedures and processes for opening data. They propose that instead open data publication run on five principles, built from detailed engagement with a Dutch government agency in the justice sector:

- start thinking about the opening of data at the beginning of the process
- develop guidelines, especially about privacy and policy sensitivity of data
- provide decision support by integrating insight in the activities of other actors involved in the publishing process
- make data publication an integral, well-defined and standardized part of daily procedures and routines
- monitor how the published data are reused.

Improve privacy rules

155. Custers and Ursic (2016) distinguish between various types of re-use by those processing data and by data subjects. This is a legal analysis, calling for privacy frameworks that encourage rather than hinder data sharing and re-use, while still being based on consent.

156. Ohm (2010) suggests regulation of the sharing of datasets that present high risks of damage if they were to be inappropriately disclosed.

157. Rohunen (2014) explores what makes data subjects more willing to disclose data. The value of the service they get in return matters, so does whether they get some additional benefit for disclosing more sensitive data (for example, personalised service). Other factors include trust in the organisation involved, informing users of how their information will be handled, and providing control over data sharing. More sensitive is less likely to be shared.

158. OECD (2013) suggests that governments ensure more effective privacy protection to support "open, secure, reliable and efficient" data flows, leading by example by providing open access to government data, working to increase the supply of relevant skills in STEM subjects, ensuring infrastructure is ready to support the demands of the Internet of Things, and investing in better measurement of the value of data to build the evidence base.

159. Forgeron (2015) provides some guidance on the legal risks for institutions considering big data projects. Understanding the legal basis of the data collection, the impact of particular analysis platform choices, the legal controls on data processing, and ensuring appropriate control of project delivery are important.

Technical issues

160. Pang et al (2014) say that there has been very little research on the value of IT investments in the public sector (as opposed to the private sector) and existing models for how IT helps firms do not take into account the characteristics of public sector organisations. They argue, following earlier work, that public entities should actively work to create greater public value (rather than just do what they are told by Ministers), just as managers in the private sector work to create greater shareholder value.

161. Morris et al (2014) proposes a secure information market as a way for agencies to benefit from pooling of data without having to share their information with competitors or take a risk of wider disclosure of their information. In operation it sounds like a data commons, or some aspects of the IDI, with each agency contributing data to a neutral third party that provides back a consolidated view.

162. Hofman and Rajagopal (2014) consider some technical issues and software choices in data sharing systems, noting that system design should be built from engagement with the users of data and an understanding of their needs (not from the perspective of the data collector).

163. Rehman et al (2013) suggest that organisations should collect less data in order to lower the costs of data manipulation as well as enhance trust and preserve privacy.

164. Hammond (2013) says that the real issue is getting from data to useful insight automatically without needing analysts (so as to avoid the bottleneck between data provision and information availability).

Fundamental institutional reform

165. The Australian Productivity Commission released the draft report on its inquiry into Data Availability and Use in October 2016. It reviews the benefits and costs of greater access and use of all public and private sector data, and looks at individual access to data about themselves, standardisation of data collection, and boosting social trust in data sharing.

166. The draft report (Australian Productivity Commission, 2016) recommends very significant changes based on new Federal legislation (the Data Sharing and Release Act) to encourage data sharing. It divides the treatment of data up by the risks that it presents. Non-personal, non-confidential data should be readily available to all. De-identified and identified data would be available to trusted users. For a subset of data that presents unusual security or confidentiality risks, release would be up to the custodians of that data.

167. A very wide range of changes is proposed: All government agencies would be required to produce comprehensive easy to access registers of all data they hold within a year and publish them on http://data.gov.au, to develop data management standards, and to release the data they collect in standards-compliant ways with the provision of quality metadata. Datasets of national interest would be able to be identified, maintained and specially curated. Consumers would have new rights to access information about themselves, to change it, and to share it. New institutions would be established, including a National Data Custodian to with oversight of the data system, and Accredited Release Authorities that would decide whether a dataset should be publicly released or released to a more restricted audience.

168. The Commission’s final report is due by 21 March 2017.

Ensuring approaches take account of the context and objectives

169. Koerten and Veenswijk (2013) suggest that the best approach to encouraging open data must take account of the characteristics of the government and the country concerned.

170. Dragos and Neamtu (2009) say that to develop an effective market for the reuse of public sector information, governments have to clarify the type or categories of information that can be disclosed and reused for commercial purposes, the type of mechanisms used (licensing, exclusive contracts, free access), and the prices to be charged.

171. Woods quoted in Dragos and Neamtu (2009) offers the following distinction with regard to public sector information:

    - A core of public sector activities which can be clearly categorized as directed towards the production of information by mandate (e.g., national statistics services),
    - A second set of activities which are not primarily directed towards producing information, but which nonetheless produce valuable information as a by-product on a significant scale (e.g., education, welfare, healthcare, environmental monitoring, weather systems);
    - Activities such as scientific research or the financing of libraries and museums, that are not always included in the narrow definitions of public sector information and might require a customised approach.

172. Lassinantti et al (2014) isolate two different approaches to open data from survey work with two Swedish municipalities (Stockhol and Skelleftea). Stockholm saw its open data efforts as a platform for economic growth through technology, so it was focused on what data would be most useful for commercial uses. Skelleftea saw its open data effort as a platform for co-created societal growth and focused on how to make government more open. Its goals were boosting citizen engagement with government, and changing existing power structures. The broader point is that the choice of programme objectives will influence how open data efforts are advanced.

173. van der Graff (2014) says cities should focus on “open access” not just open data, and offer new opportunities for citizens to get involved in the generation, use and integration of economic, social and environmental data.

174. Janssen (2011) distinguishes between calls for open data based on concepts of freedom of information (focused on boosting accountability and transparency of government and public participation), and those based on the reuse of public sector information (focused more on the economic value of information held by public sector agencies). Freedom of information is a less demanding standard than data reuse, since it only requires the release of data in any pre-existing format. She suggests that agencies should rethink their information policies and make one coherent plan that covers all request for access, actively making data available for public use and encouraging the development of works based on the data.

Other relevant inputs

175. In this section we briefly review some data on how New Zealand is going on open data in government, and review some feedback to the DFP in response to its public request for input on its Diagnose and Fix work.

Domestic assessments

176. New Zealand's open data programme has strong foundations:

- The 2011 Declaration on Open and Transparent Government requires core government agencies to release all high-value public data for reuse.
- The 2011 New Zealand Data and Information Management Principles set rules for release, including on formats, publishing and pricing.
- The New Zealand Government Open Access and Licensing framework (NZGOAL) solves licensing issues, giving government agencies clear guidance on legal issues around the release of open data.
- Several thousand datasets have been published from a wide range of agencies.
- There is a Register (http://data.govt.nz) of open data released, intended to make data easier to find.
- Officials report annually to Cabinet on progress with implementation of the Declaration, and release those reports publicly.

177. As of 4 November, there were 4,312 datasets registered on http://data.govt.nz. The 2014 report to Cabinet on progress against the Declaration reports that around two thirds of all datasets published by the public and non-public service departments feature on the Register.

178. This is an under-estimate of the total volume of public data produced by government agencies. Some agencies have just one Register record for multiple datasets, because agencies do not list all of their published datasets on the Register, and because there is much more data that is published by agencies on their websites that is not easily accessible as a dataset and is not referred to on the Register.

179. Echoing the findings from previous annual reports, the 2015 report describes ongoing progress with publication and valuable uses of data but also (paragraph 38) that "not all high value public data is routinely released" and when released "it is not always visible or published in fully open formats or licensed". Outside of central government, there is growing awareness but little action on open data. Even inside central government, progress is uneven: LINZ has published 42 per cent of all the datasets on the Register. Eight other agencies have published more than 100 datasets. Eighty four other agencies have released just one dataset, generally Chief Executive expenses and in PDF format.

180. The top ten datasets that users would like to see more open, according to an online survey by the Open Government Information and Data Programme (in LINZ) in June 2016, include:

- Weather
- Water quality
- Property, postal address and postcodes
- Resource consents
- Public transport
- Crime statistics
- Companies Register
- Government contracts and spending

181. The 2014 report to Cabinet says that only 10% of data requested by users on http://data.govt.nz has been released. In some cases this is because open data release is incompatible with business models that have been adopted by government agencies. In other case it seems that releasing data is not sufficiently high on agencies' lists of priorities.

International assessments

182. The 2015 Open Data Barometer (ODB) ranks New Zealand sixth amongst 92 countries, behind the US, USA, France, Canada and Denmark (and down from fourth place in 2014). New Zealand scores well for social and political impact, for its open data policies and for the readiness of civil society to take advantage of open data. It is held back by relatively poor data availability, especially a lack of machine-readable data, and by a lack of open data in particular areas: notably the companies register, government spending, public contracts, and public transport timetables. New Zealand also scores relatively poorly for the impact of open data on transparency and accountability of government, and on economic impacts.

Other feedback

183. The output from various DFP public engagements were shared with us for consideration in the course of our work.

184. Generally the barriers nominated were related to a lack of data being released, and the useability of data that is released:

- Respondents were particularly concerned about a lack of data at detailed geographical levels, a lack of data publishing by agencies as a matter of course (rather than requiring requests), a lack of timely data publication.
- Respondents also said it was too hard to find data and there was a lack of machine-readable data,

185. Some respondents thought a barrier was a lack of agency capability in making their data available. Related: one respondent focused on the systems and processes available for data publication calling for two mandatory common data transfer formats and processes:

- One for secured, raw, non-confidentialised, not-to-be-published unit record data, and
- One for unsecured, clean, confidentialised, publishable aggregate data.

Interviews

186. We have interviewed 14 people as part of the first phase of this project. What follows is a summary of what we learned from those discussions, organised by theme. These interviews were useful to develop and test our thinking, and to explore areas of agreement and disagreement. There is no quantitative element to this piece of our work.

187. The interviewees are a mix of DFP members and people working with data in the public and private sectors, and in academia. A list of names is in the Appendix. They were recruited from our networks and from asking interviewees for other names. We plan to do more interviews as part of the next phase of this work, more directly testing our thinking about barriers and potential solutions.

Value at stake

188. The people we spoke to thought that more data sharing would generate a lot of value for New Zealand, and in many different ways. One noted that so far “we are only scratching the surface of what is possible with data sharing”.

189. None of our interviewees had an overall view of the scale of value available. Some said it was not possible to estimate it in aggregate, although case studies were one way to get a sense of value in particular cases. One referred to the Bakker (2013) study of the value of the Census as an example of what was possible.

190. Pressed on the quantification point, one interviewee gave an example of collecting and sharing data across agencies on one type of government spending and said that it was easy to imagine that knowing how much the government was spending in this area was worth at least a one per cent improvement in the effectiveness of spending. And in practice the project cost much less than one per cent of total spending in any case. From the people we spoke to, this seemed to be how business cases for investment in data sharing were justified, i.e., based on comparing the spending requested with the threshold level of success required to justify that spending.

191. Another public sector interviewee nominated an improvement of ten per cent in the cost-effectiveness of government services, on the basis that 100 per cent increase in value seemed infeasible, and one per cent seemed too low.

192. Interviewees often said that there would be more value in some areas than in others. Some felt that the value from sharing of data generated by social sector agencies was likely to be higher than from sharing of data generated by economic agencies because the former was directly related to improving the effectiveness of government services.

193. More than one interviewee noted the value difference between a simple data supply and a more complex data integration. An example of the first case would be MBIE releasing tenancy bond data by suburb on its website. An example of the latter would be a joining of MBIE’s tenancy bond data with data on household

incomes from the Census for the same suburbs. The latter has much great value for analysis.

194. A couple of interviewees said that the ultimate value would come from the ability to entirely redesign our societal institutions for a world where data can be easily collected and shared. In particular, the specialised vertical silos that characterise our delivery of government services will no longer be the most efficient approach.

195. One interviewee argued that there was no reason to try to quantify the value at stake in any case. The question that faces us is whether to do more or less data sharing than we are doing now. And the answer to that question is clearly to do more. When we see the returns from new data sharing projects reducing, then we can conclude that most of the value has already been captured.

196. We were told of some areas where the value of additional data sharing efforts might be low. One example was the Regional Economic Activity Reports that have had a lot of investment already.

Types of sharing

197. We sometimes found it helpful in interviews to distinguish between different types of data sharing based on whether the collector and the user of the data were in the public or private sector. Several interviewees commented on how these different situations were playing out, as reflected in the table below.

Collecting agency	Using agency	Situation
Public	Public	Data sharing is evident and increasing. There is not a lack of will (because data sharing improves agency service delivery) but it is hard to do. Data standards and standard agreements could help make it easier.
Public	Private	Open data seems hard and little is happening outside of Statistics NZ and LINZ. Public data is crucial for some private sector uses, especially Census and location datasets. In some cases there are commercial barriers to release, e.g., NIWA, MetService, property data
Private	Public	Little happening visibly although there are examples of collaboration. Public agencies are reportedly averse to buying data. Private companies see little reason to share it with government. Might be helped if there were more public examples of success.
Private	Private	There is a market for data with established players. There are some interesting examples of the establishment of shared structures to enable pooling of industry data. Standardised agreements for value sharing could help reduce the fixed costs of negotiating sharing arrangements.

Motivations for sharing

198. Several interviewees noted a lack of motivation to share data both in the public and private sectors.

199. Public sector interviewees noted a sense of fear from officials about open data sharing. It could create trouble for officials in particular instances even if data sharing in general is supported at Ministerial level. Some public sector agencies fear that more open data release will lead to misuse of their data that will reflect badly on them.

200. Generally interviewees felt that these concerns could be allayed through careful engagement agency by agency. But this is time consuming and difficult work. And because the motivation for open data is so weak in general, even quite small barriers tend to be hard to surmount.

201. Interviewees noted that using data for policy was still something that was being pushed from the data side, rather than being actively driven by policy people seeking data to help their work.

Data governance, quality and standards

202. A common thread in the interviews was a lack of data governance amongst those collecting and sharing data. This includes, for example, ensuring the organisation knows what data it has and its quality (just as for any other significant organisational asset). There are existing international standards and voluntary codes that could be built upon both to lift the quality of oversight of governance internally, and to give agencies comfort that the organisation they are dealing with in a data sharing project is trustworthy.

203. Several interviewees pointed out that data quality was highly variable between datasets. Basic matters like a good data dictionary, quality metadata, and clarity on how changes in definitions over time have been dealt with are often missing. Data that an organisation itself uses in its administrative processes is higher quality (because a lack of quality has direct negative impacts for the organisation), whereas data that is collected but not used by the organisation can be of lower quality without it causing any concern for them. Perhaps there are ways to encourage those collecting data to think about the needs of others who might come to use their data later on.

Coordination, standards and processes

204. Many interviewees talked about ways to make data sharing easier through systematisation and coordination:

- Several argued for aligned definitions and methods for data collection so that where many organisations are collecting the same data, they do it in the same way. For example, there could be standards for how to structure and verify addresses or names that would mean that all organisations would do so in well-documented and understood ways.
- Several pointed out the benefits of standardised legal agreements for data sharing and simple ethical guidance materials. This could minimise the

effort for each data collector in developing their own agreements and making their own decisions about what is permissible. Standard agreements for splitting value between the members of a data supply chain could also be useful.

- Several suggested that there should be processes in the public sector to make release of open data easier. It should be a more standard part of data collection and analysis rather than a special activity that requires particular attention.

Privacy of personal information

205. Some of our interviewees saw an inconsistency between the existing rights-based approach to privacy and enabling wider data sharing and use. Some argued that our existing rights-based approach did not work in practice as well as it should: data subjects have to protect themselves against data security threats (through consents and passwords and verification) but still carry the risk of data breaches when things go wrong. Others noted that, although data subjects are notionally in control of use of data about them, the requirement for consent is more a legal protection for data collectors, not a practical one for data subjects. Nor do privacy rules provide any ethical guidance for those using data.

206. Some interviewees thought that an approach based on providing transparency when personal data is used might be more practical and useful than the existing requirements for consent for all purposes in advance.

207. Some interviewees posited a clear tradeoff between privacy and data sharing. Others argued that this was a false dichotomy and that it was possible to build privacy-friendly data-sharing frameworks. One noted that the Privacy Act is built from a perspective of restricting sharing and it would be helpful to have some legislation that was directly supportive of data sharing.

208. Several interviewees thought that anonymity and re-identification were likely to be more significant issues in New Zealand than in other countries given our small population.

What else might help

209. Interviewees made a range of other suggestions, including:

- Improvements to the IDI to make it more readily usable as an operational tool as well as a research device, e.g., to run standardised monthly reports on matters of particular interest.
- Developing some centralised systems to enable easy data sharing, which sound in principle like the idea of Data Commons that is being pursued as a Catalyst project.
- Making open data compulsory for public sector agencies and removing the restriction to just “high value” data.
- Treating everyone as a data user rather than assuming that data analysis is a specialised activity separate from developing policy advice or other roles.

- Figuring out ways to make data sharing have operational value for government agencies. This will give them direct business motivations to do more data sharing.

List of interviewees

Those mentioned together were group interviews.

- Aaron Zhu, Trace Research
- Aaron Beck and Alistair Ramsden, Stats NZ
- Alison Holt, Longitude 174
- Deb Potter, Che Tibby, Vince Galvin, Statistics NZ
- Ed Abraham, Dragonfly Data Science
- Eileen Basher, MBIE
- Erin Thomas, Loyalty NZ/Lab 360
- Julian Carver, Seradigm
- Lillian Grace, Figure.NZ (and DFP member)
- Mark Sowden, Statistics NZ
- Mike Parsons, Datamine
- Miriam Lips, Victoria University (and DFP member)
- Peter Ellis, MBIE
- Shaun Hendy, Auckland University

References

ACIL Tasman (2009), "Spatial information in the New Zealand economy: Realising productivity gains"

AnalysysMason (2014a), "Data Driven Innovation in Singapore"

AnalysysMason (2014b), "Data Driven Innovation in Japan: supporting economic transformation"

Australian Productivity Commission (2016), "Data Availability and Use", Productivity Commission Draft Report, October 2016

Bakker, Carl (2013), "Valuing the Census"

Barua, Anitesh, Deepa Mani, Rajiv Mukherjee (2011), "Measuring the Business Impacts of Effective Data", "Impacts of Effective Data on Business Innovation and Growth", and "Impacts of Effective Data on Operational Efficiency"

Brynjolffson, Erik, Lorin Hitt and Heekyung Kim (2011), "Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?"

Cap Gemini (undated), "The Digital Advantage: How digital leaders outperform their peers in every industry"

Charbonneau, Etienne and Gautam Nayer (2012), "Barriers to the Use of Benchmarking Information: Narratives from Local Government Managers"

Custers, Bart and Helena Ursic (2016), "Big data and data reuse: a taxonomy of data reuse for balancing big data benefits and personal data protection" de Montjoye, Yves-Alexandre, César A. Hidalgo, Michel Verleysen, Vincent D. Blondel (2013), "Unique in the Crowd: The privacy bounds of human mobility”

de Rosnay, Melanie Dulong, and Katleen Janssen (2014), “Legal and Institutional Challenges for Opening Data across Public Sectors: Towards Common Policy Solutions”

de Saulles, Martin (2007), “When public meets private: conflicts in information policy”

de Serres, Alain, Naomitsu Yashiro and Hervé Boulhol (2014), “An international perspective on the New Zealand productivity paradox”, New Zealand Productivity Commission Working paper, WP 2014/1

Dekkers, Makx, Femke Polman, Robbin te Velde, Marc de Vries (2006), “Measuring European Public Sector Information Resources: Final Report of Study on Exploitation of public sector information – benchmarking of EU framework conditions”

Deloitte Access Economics (2011), “The Connected Continent – How the Internet is transforming the Australian economy”, Report prepared for Google Australia

Deloitte Access Economics (2013) “Connected Small Business – How Australian small business are growing in the digital economy”

Dragos, Dacian and Bogdana Neamtu (2009), “Reusing Public Sector Information - Policy Choices and Experiences in some of the Member States with an emphasis on the Case of Romania

Englesman, Wilco (2007), “Information Assets and their Value”

Fioretti, Marco (2010), “Open Data, Open Society”

Forgeron, Jean-Francois (2015), “Legal Aspects of Big Data”

Glass, Hayden, Preston Davies, Eli Hefter, Gary Blick (2014), “The Value of Internet Services to New Zealand businesses”

Golle (2006), “Revisiting the uniqueness of simple demographics in the US population,”

Grimes, Arthur, Cleo Ren and Philip Stevens (2012), “The Need for Speed: Impacts of Internet Connectivity on Firm Productivity”

Hammond, Kristian (2013), “The Value of Big Data Isn’t the Data”, Harvard Business Review, May 01, 2013

Higgins, Emma, Mark Taylor, Paulo Lisboa and Farath Arshad (2014), “Developing a data sharing framework: a case study”

Hofman, Wout and Madan Rajagopal (2014), “A Technical Framework for Data Sharing”

Janssen, Katleen (2011), “The role of public sector information in the European market for online content: a never-ending story or a new beginning?”

Jetzek, Thorhildur, Michel Avital, and Niels Bjorn-Andersen (2015), “The Value of Open Government Data: A Strategic Analysis Framework”

Kaschesky, Michael and Luigi Selmi (2014), “7R Data Value Framework for Open Data in Practice: Fusepool”

Kitchin, Rob, Sandra Collins and Dermot Frost (2015), “Funding models for Open Access digital data repositories”

Koerten, Henk and Marcel Veenswijk (2013), “Public sector information reuse across Europe: Patterns in policy-making from an organizational perspective”

Koski, Heli (2011), “Does Marginal Cost Pricing of Public Sector Information Spur Firm Growth?”

La Valle, Steve, Eric Lesser, Rebecca Shockley, Michael S. Hopkins and Nina Kruschwitz (2011), “Big Data, Analytics and the Path from Insights to Value”

Lassinantti, Josefin, Birgitta Bergvall-Kåreborn and Anna Ståhlbröst (2014), “Shaping Local Open Data Initiatives: Politics and Implications”

Manterlero, Allesandro (2014), “Social Control, Transparency and Participation in the Big Data World”

McKinsey Global Institute (2011), “Internet matters: The Net’s sweeping impact on growth, jobs, and prosperity”

McKinsey Global Institute (2013), “Open data: Unlocking innovation and performance with liquid information”

Micus Consulting (2008), “The impact of broadband on growth and productivity – a study on behalf of the European Commission”

Missingham (2009), “Switch on the Data: Changes Needed for Access to Public-Sector Information”

Morris, Bonnie, Virginia Franke Kleist, Richard Dull Cynthia Tanner (2014), “Secure Information Market: A Model to Support Information Sharing, Data Fusion, Privacy, and Decisions”

MYOB (2011), “MYOB Business Monitor – Online Special Report”, July 2011

Nugroho, Rinanta Putri (2013), “Comparison of open data policies in different countries: Lessons learned for an open data policy in Indonesia”

Nugroho, Rinanta Putri, Anneke Zuiderwijk, Marijn Janssen and Martin de Jong (2015), “A comparison of national open data policies: lessons learned”

OECD (2013), “Exploring Data-Driven Innovation as a New Source of Growth”, Digital Economy Papers No. 222

OECD (2015), “Assessing government initiatives on public sector information”, Digital Economy Papers No. 248

Ohm (2010), “Broken promises of privacy: Responding to the surprising failure of anonymization”

Pang, Min-Seok, Gwanhoo Lee, William DeLone (2014), “IT resources, organizational capabilities, and value creation in public-sector organizations: a public-value management perspective”

Parycek, Peter, Johann Hochtl and Michael Ginner (2014), “Open Government Data Implementation Evaluation”

Petychakis, Michael, Olga Vasileiou, Charilaos Georgis, Spiros Mouzakitis, and John Psarras (2014), “A State-of-the-Art Analysis of the Current Public Data Landscape from a Functional, Semantic and Technical Perspective”

PIRA (2000), “Commercial exploitation of Europe’s public sector information”

Prime (2015), “Prime, R. (2015). How to measure the value of big data. Information Age, 21 August 2015.

PWC (2014), “Deciding with data: How data-driven innovation is fuelling Australia’s economic growth“

Rehman, Muhammad Habib, Victor Chang, Aisha Batool, The Ying Wah (2013), “Big Data Reduction Framework for Value Creation in Sustainable Enterprises

Rohunen, Anna, Jouni Markkula, Marikka Heikkila and Jukka Heikkila (2014), “Open Traffic Data for Future Service Innovation – Addressing the Privacy Challenges of Driving Data”

Sapere and Covec (2015), “Data Driven Innovation in New Zealand”

Scassa, Teresa (2014), “Privacy and Open Government”

Statistics New Zealand (2013), “Strong connection between ICT and business-growth activities”

Susha, Iryna and Ake Gronlund (2015), “Organizational measures to stimulate user engagement with open data”

Sweeney (2000), “Simple Demographics Often Identify People Uniquely”

Sweeney (2002), “Foundations of Privacy Protection from a Computer Science Perspective

Taylor, P (2012), “Extracting value from information”, The Financial Times, 26 May 2012

Ubaldi, B (2013), “Open government data: Towards empirical analysis of open government data initiatives”, OECD Working Papers on Public Governance, No. 22, OECD Publishing

Vaish, A, A Prabhakar, H Mishra, N Dayal, S Singh, and U Goel (2011), “Quantifying information dynamics through a new valuation system”

van der Graff, S (2014), “Smarten Up! Open Data, Toolkits and Participation in the Social City”

Varian et al (2002), “The Net Impact Study – the projected economic benefits of the Internet in the United States, United Kingdom, France and Germany”

Venkatraman, Archana (2010), “Open up to the public”

Vickery, Graham (2011), “Review of recent studies on PSI Re-use and Related Market Developments”

Viscusi, G, M Castelli, and C Batini (2014), “Assessing social value in open data initiatives: A framework”

Weiss, Peter (2002), “Borders in Cyberspace: Conflicting Public Sector Information Policies and their Economic Impacts”

Wiewiora, Anna, Glen Murphy, Bambang Trigunarsyah and Kerry Brown (2014), “Interactions Between Organizational Culture, Trustworthiness, and Mechanisms for Inter-Project Knowledge Sharing”

World Wide Web Foundation (2016), “Open Data Barometer”

Yang, Tung-Mou and Terrence Maxwell (2011), “Information-sharing in public organizations: A literature review of interpersonal, intra-organizational and inter-organizational success factors”

Zakaria, Abd Hadi, and Neil McBride (2000), “The commercialisation of public sector information within UK government departments”

Zang and Bolot (2011), “Anonymization of location data does not work: A large-scale measurement study”

Zuiderwijk, Anneke, Natalie Helbig, J. Ramón Gil-García, Marijn Janssen (2014), “Special Issue on Innovation through Open Data - A Review of the State-of-the-Art and an Emerging Research Agenda: Guest Editors’ Introduction”, Journal of Theoretical and Applied Electronic Commerce Research, Vol 9 Issue 2, May 2014