Statistics New Zealand

Unleashing the power of data to change lives

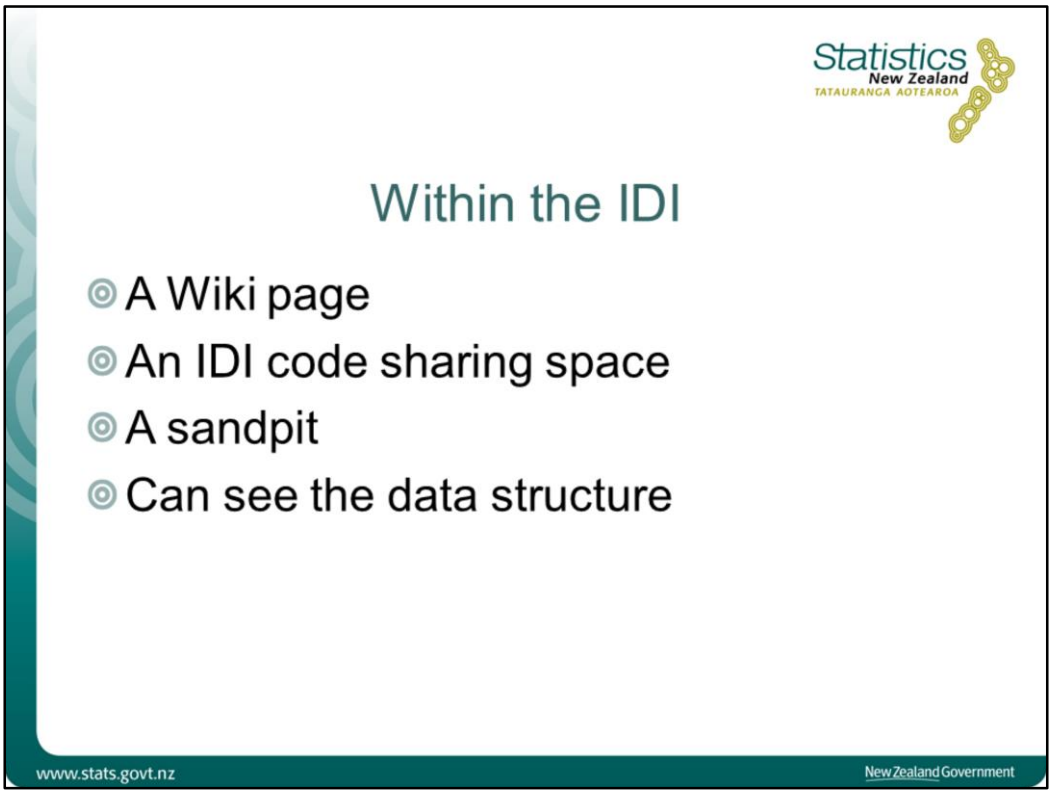
Getting started in the IDI

April 2015

www.stats.govt.nz

New Zealand Government

The IDI is a series of SQL databases. The data tables in the IDI are generally organised by the name of the agency that supplied the data (e.g. IRD, MSD, MBIE). A researcher will only have access to those tables that have been approved for their research. That approval is based on the information necessary for the work that they are doing.

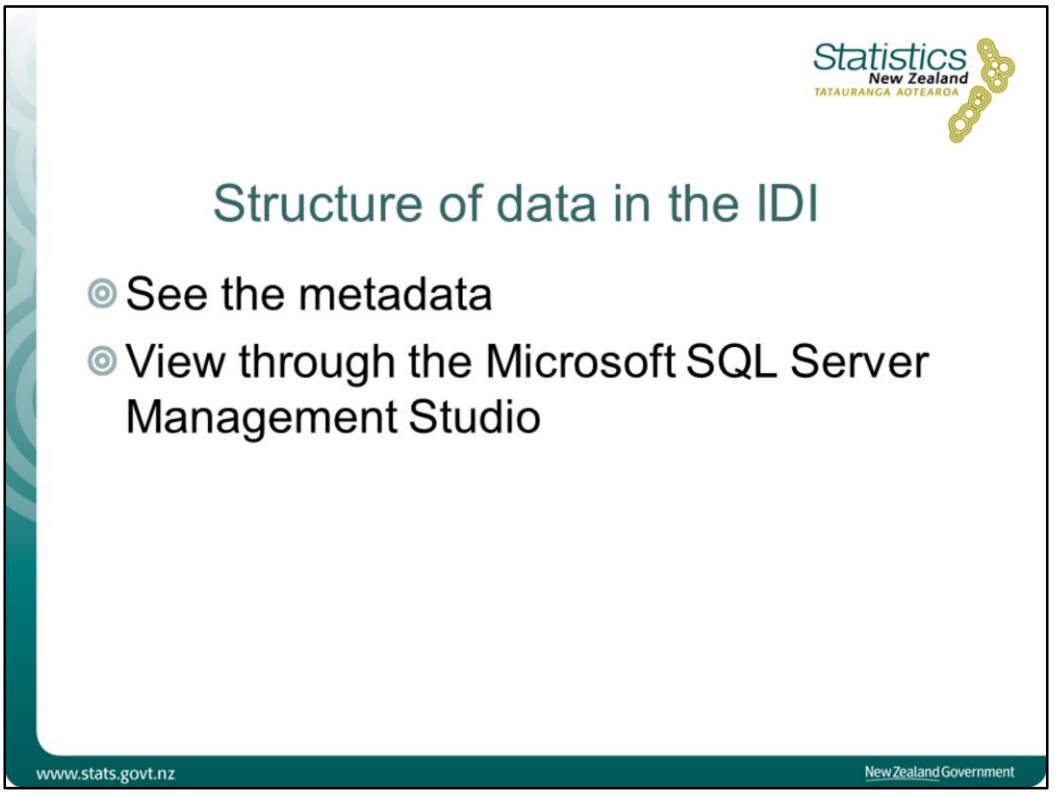


In addition to having access to the data tables in the IDI, the researcher also has access to some shared spaces. These include:

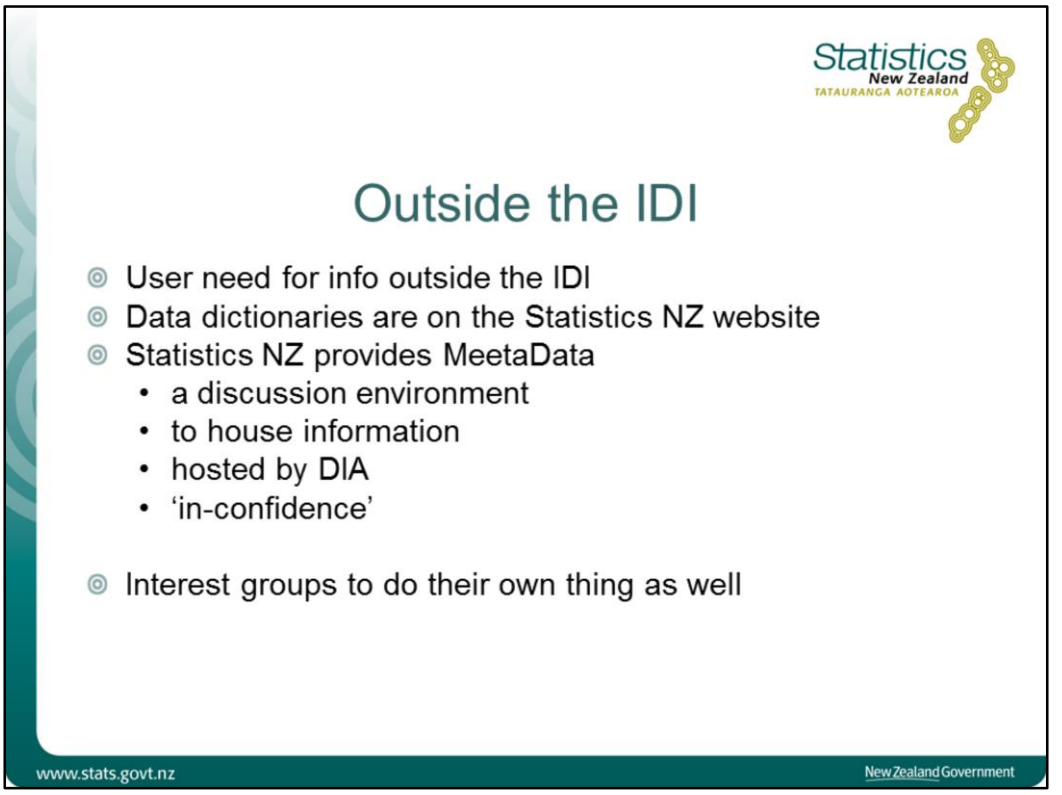
- A Wiki page containing useful information and resources
- An IDI code sharing space (accessed through the Wiki) where researchers can put code that they wish to share with all IDI users
- A sandpit where researchers can put tables, datasets, and programming code to be shared with members of their project team
- A metadata database that contains detailed metadata about specific collections (accessed through the Wiki)

The WIKI

- what the IDI is
- technical aspects of the IDI
- how to access and apply for access
- data quality information
- efficiency tips
- metadata (ie data dictionaries)
- note on using the using the sandpit
- forum for researchers
- discussion and code sharing area
- other useful information



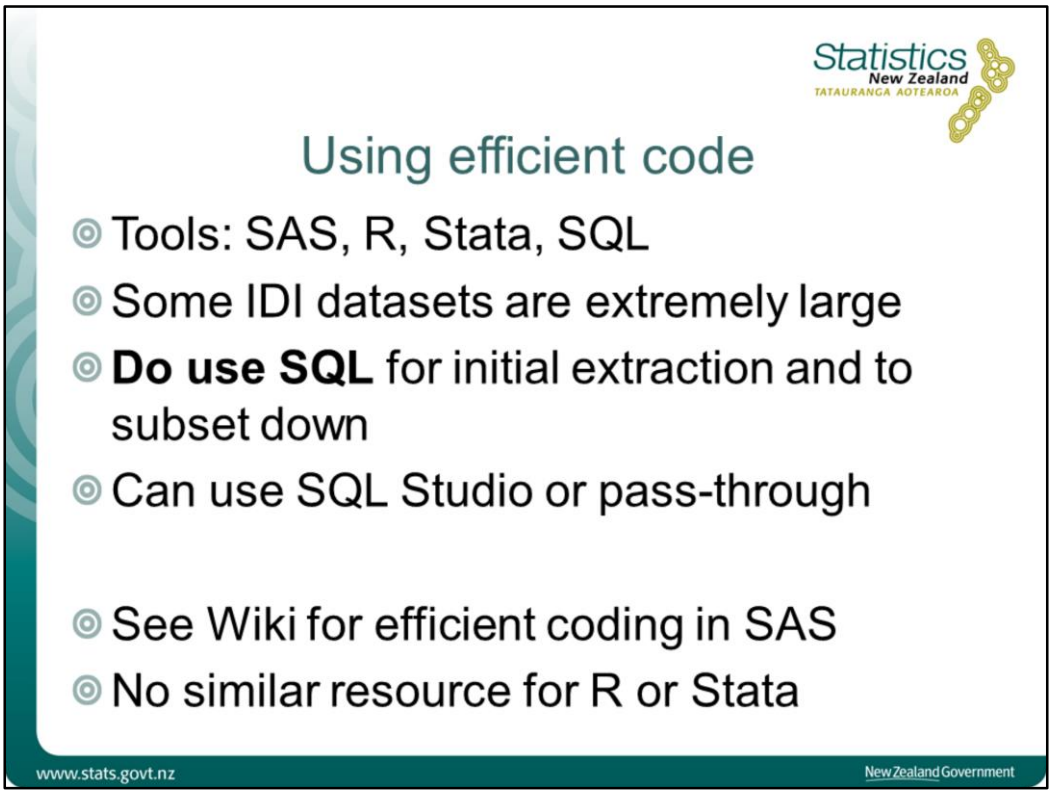
For researchers who have access to the IDI there are two main ways that they can learn about the structure of the data. First, they can examine the data dictionaries and other metadata available through the Wiki (discussed above). Second, they can view the data through the Microsoft SQL Server Management Studio.



Many researchers wish to have access to information about the IDI before they are granted access to the IDI. Having some prior knowledge about the IDI is important for researchers as it helps them to plan and cost their work. Researchers have also expressed a need to have access to an environment (outside the IDI) where they can discuss ideas related to their research. Some discussions may take place across a wide range of researchers while other discussions may take place within particular interest groups (e.g. Health researchers, or Justice researchers).

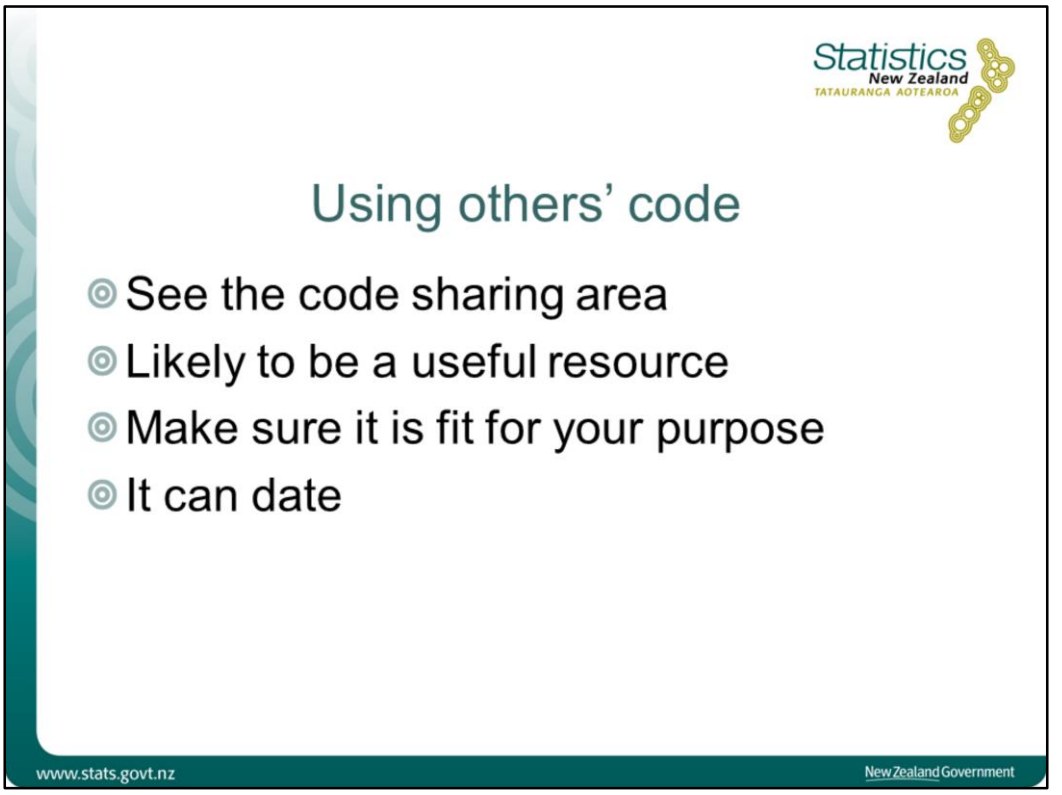
Statistics NZ has responded to this need for information and collaboration outside the IDI in two ways:

- data dictionaries are available on the Statistics NZ website
- MeetaData, an online collaboration space for microdata users to connect with each other, share knowledge, ask questions, and share code.



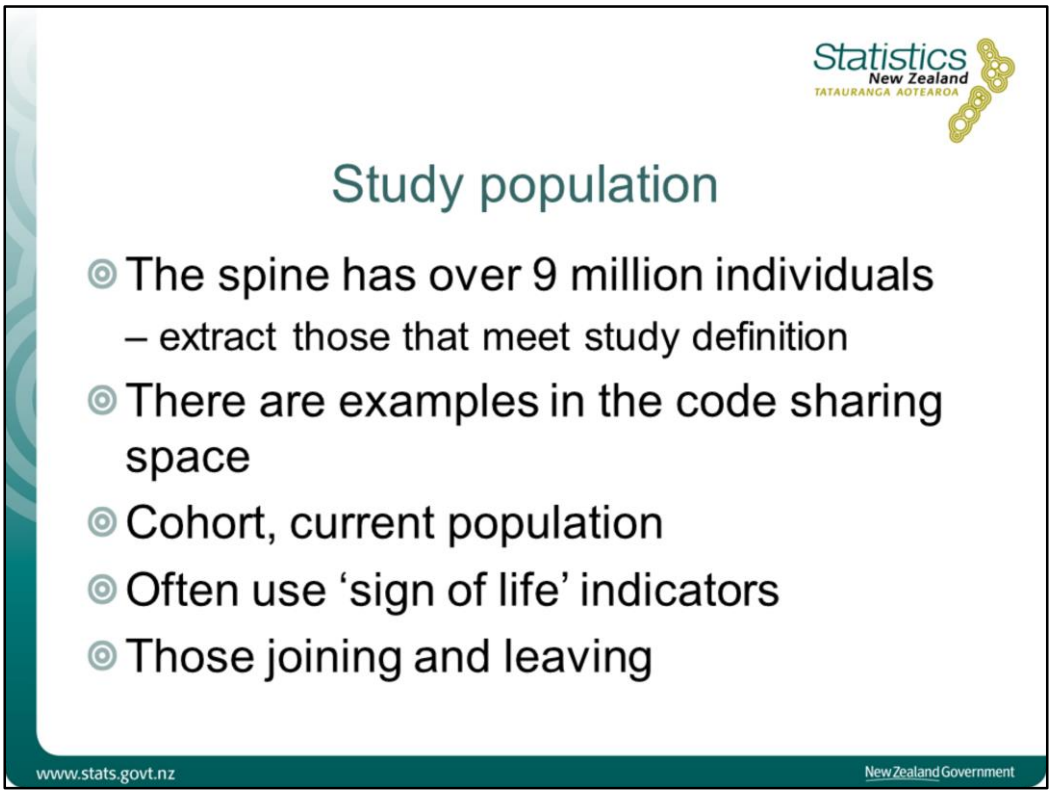
Datasets in the IDI are often extremely large and this requires researchers to be efficient in the way they write code – particularly when they first extract the datasets they want to use in their study.

Many programming/analytical languages can be used in the IDI. These include SAS, R, Stata, and SQL. Regardless of the tool used for the majority of the analysis, it is advisable to use SQL to do the initial data extraction from the IDI and to subset this down into a useable form. To do this efficiently, it is important to use SQL in such a way that the query is carried out directly on the IDI server. This means submitting the SQL code through the Microsoft SQL Server Management Studio or using an explicit pass through method from another tool (e.g. explicit pass through using SAS Proc SQL). The explicit pass through means that the code is submitted directly to the SQL server for processing.



The Statistics NZ Wiki page now has an area where researchers can post code they have used in their analysis (see IDI Research – IDI Researcher Code Sharing). The repository holds a lot of SAS code used to define study populations and to set up derived variables. The resource is likely to save considerable time and effort for researchers. Researchers should take adequate care to ensure that the code is fit for purpose.

Code which has been output-checked can also be shared on MeetaData.



Researchers usually wish to create a study population in the IDI. This is sometimes called a 'base', 'reference' or 'denominator' population. The study population is defined according to the needs of the research.

There are two common ways of defining a study population in the IDI. These are often referred to as the 'cohort population' and the 'current population'. It is not uncommon for both of these approaches to be used within the same study.

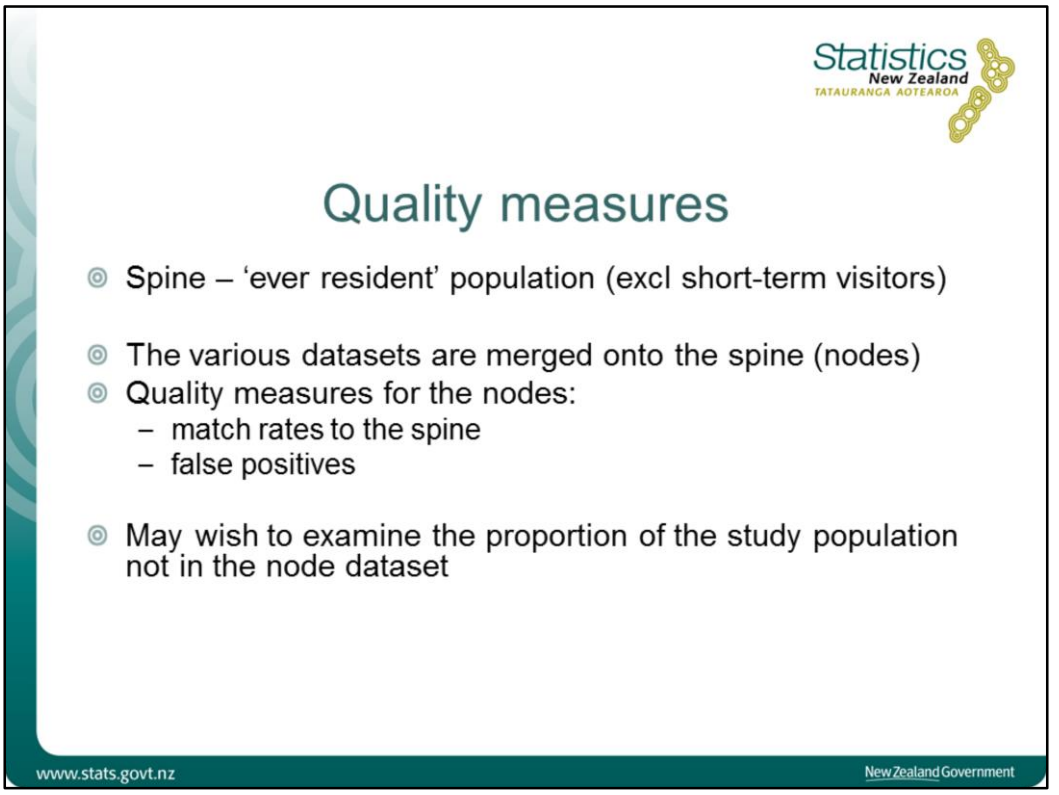
The cohort population defines a population as at a particular period and then follows that cohort through time (e.g. people born between 1 July 1990 and 30 June 1991).

The current population approach defines the study population, as closely as possible, to the current population of interest (e.g. people with permanent residence aged 0–25 as at December 2014 who had evidence of having lived in New Zealand in the preceding year).

The IDI spine contains more than 9 million individuals, far more than the current New Zealand usual resident population of approximately 4.5 million. The researcher will need to extract those individuals that meet their specific research needs.

When constructing a survey population the researcher might examine such things as: whether they made an income tax payment, whether they were a domestic student, whether they received a benefit, and whether they were part of the National Health population index. Such indicators are sometimes called 'sign of life' indicators.

Researchers need to consider how individuals join and leave the population. This would include considering how to deal with individuals who have died, those who have recently been born. and those who have arrived or left the country.



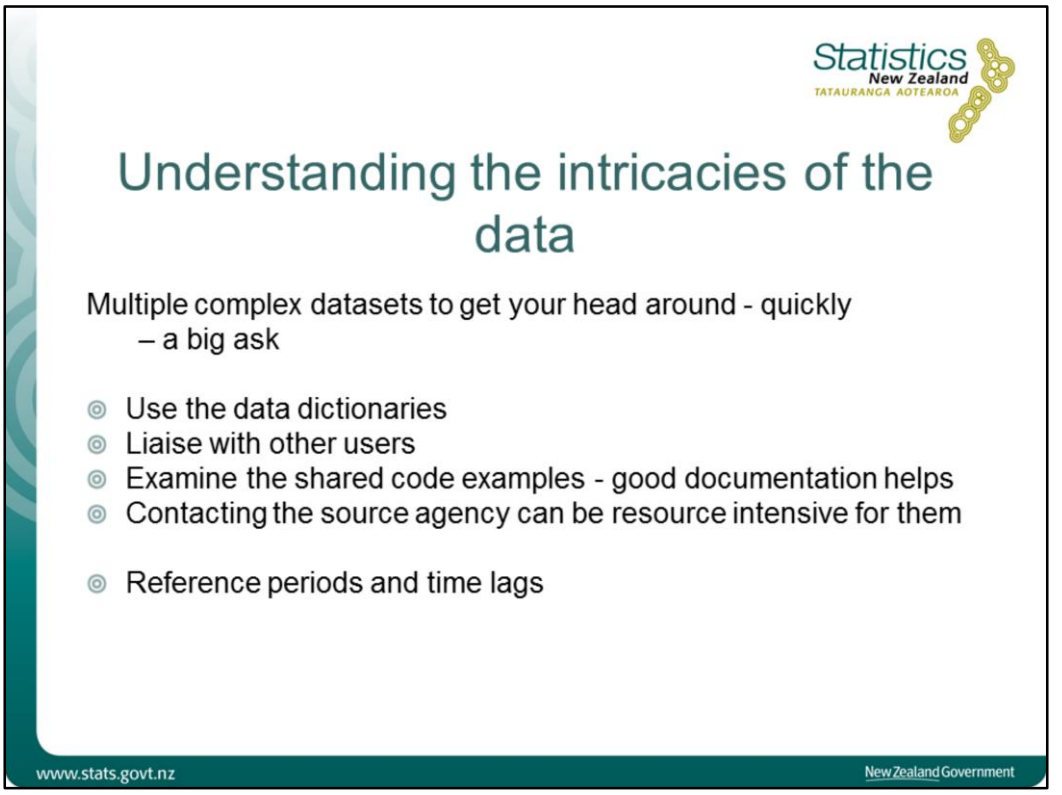
The IDI is based on the principle of being able to link the administrative records of New Zealanders. The ability to do this depends on having records for a large proportion of New Zealanders and being able to link the various records together at an individual level.

One of the fundamental building blocks of the IDI is the IDI spine. The spine is intended to be a list of the 'ever-resident' population (excluding short-term visitors). The spine is constructed from a small number of core datasets, namely: birth data (DIA), visa data (MBIE), and tax data (IR).

When constructing the spine, Statistics NZ examines a number of quality measures such as the link rates between the three contributing datasets, and the false positive rates.

Having constructed the IDI spine, Statistics NZ then links many other datasets onto the spine. The additional datasets are linked onto the database as a 'node'. Statistics NZ monitors the quality of the link between the spine and these additional node datasets. The node linking rates can be found in the Wiki under 'About the IDI – linking passes'. Researchers should be aware that linking with the node dataset is only attempted where there is sufficient data to think that a link may be able to be formed, that is, where key variables are missing, no link is attempted. The link rate that is quoted in the Wiki page only relates the part of the node dataset where a link was attempted. In particular the link rate is not necessarily the proportion of the node dataset that has linked to the spine.

Researchers may also wish to examine the proportion of a study population that is not in one or more of the node datasets. For example, sometimes a researcher may exclude observations from the study population where there is no data available from a particular node dataset. In this situation, the researcher may wish to report on the proportion of observations that are excluded because there was no link to the node dataset.



The IDI contains many datasets across a range of subject matter areas. While there are usually data dictionaries that accompany the datasets, there are many intricacies within the data that are hard to understand. To assist, researchers are encouraged to examine the metadata and the shared code. Another approach is to liaise with other IDI users, perhaps through an IDI collaboration environment such as MeetaData. In some instances it may be feasible to contact the agency that supplied the data, although this can potentially be very onerous for the agencies concerned.

One of the advantages of the IDI is that it provides a wide range of variables with which to analyse the data. Ultimately some judgement needs to be made about the variables that are reported. The results of any statistical analysis will help inform that decision. However, it is also suggested that researchers consider policy and operational requirements when making the decision about which variables to report on. It may be appropriate to include some variables even if they are found not to be significant in a statistical model. Such important variables might include age, sex, ethnicity, and possibly region.