Getting started using the IDI in the Statistics New Zealand Datalab

Using the datalab

Before you can work in the Datalab you will need to complete Statistics New Zealand's (StatsNZ) confidentiality training and sign a secrecy agreement.

StatsNZ has Datalab facilities available at their offices in Wellington, Auckland and Christchurch. These are available to all datalab users. Contact StatsNZ's microdata access team about accessing their datalabs (email - access2microdata@stats.govt.nz, phone - 04 931 4253).

Many institutions such as universities and government agencies have their own datalabs. Contact the relevant institution for details about accessing these.

The Datalab is a secure environment - nothing can be put into it or taken out of it without passing through StatsNZ's microdata access team. The computer that you use to access the IDI can't be used for anything else - it has no email and most websites are blocked. This means that you need to do some preparation in order to make the best use of your time in the Datalab. If you have code or other files that you want to use in the Datalab, email them to the microdata access team ahead of time so that they can put them in the Datalab environment for you. You are allowed to take a tablet or laptop into the Datalab, but you will not be able to transfer anything from that device into the Datalab environment. Use of it will be restricted to checking emails, searching the web, looking at your notes, etc.

The Datalab environment has SAS, SQL, STATA, and R Studio. If you need other software you will need to talk to the microdata access team well ahead of time (preferably when making your application).

See more from StatsNZ - Using the Datalab

Orientation

If you have not used the IDI before, I suggest you spend your first session in the Datalab getting to know the structure and content of the IDI.

Have a look around the IDI wiki. This should open automatically when starting Internet Explorer. There is also a link to the wiki from your Datalab folder. The wiki is only accessible from within the Datalab and contains data dictionaries, information about how to access IDI, how to pull the data into programs such as SAS and R, efficiency tips, and more. If you are interested, you can also have a browse through the code sharing area and the discussion board. These are linked from the wiki under the 'IDI research' heading.

SQL Management Studio

A good way to get a feel for the structure of the IDI and view the available data is to use SQL Server Management Studio. This is already available on Datalab computers. When opening, you will need to type in the server name you are connecting to. This can be found in the IDI wiki. You can use this in 'point and click' fashion, and there is no need to program in SQL to explore the environment. Data tables are found under the "Databases" menu on the left, and then under "Tables". The IDI is organised into 'refreshes', which are versions or updates. They are labelled "IDI_Clean_" followed by the date of the refresh. Usually when starting a project you will want to use the most recent refresh.

Within each refresh there are a series of data collections (schemas), and within each schema there are tables. The names of tables are displayed as schema.table. So, the table 'moe.enrolment' refers to the 'enrolment' table within the MoE (education) collection. Derived tables (tables created by IDI, rather than coming from a specific agency) are stored in the 'data' schema (eg data.personal_detail). In SQL viewer you can see the names of all tables, including those that you don't have access to (and cannot open). For those that you do have access to, you can view them by right clicking on the table name and selecting 'view top 1000 rows'.

Data dictionaries

Some tables contain many variables, and it's not always obvious what the variables or their values are. IDI data dictionaries are available through the wiki and have explanations of the content of IDI datasets. Some data dictionaries are also available outside the Datalab environment. The quality of the data dictionaries varies. If you have a question about the meaning of a particular variable, or where you might find something, don't hesitate to ask the microdata access team or other researchers. Sometimes it is also helpful to go back to the data dictionaries produced by the source agencies (eg for health or education). You can find these online.

Further variable information is also available from tables in the IDI_Metadata folder in the SQL Management Studio. All Datalab users should have access to the Metadata – if you find you do not, contact the microdata access team and request access. Note however that the IDI_Metadata does not contain information for all variables.

You can also explore the IDI structure through SAS or R, but it is less convenient. For example in SAS you can run a libname statement at the beginning of your code for the data collection(s) that you want to view (see the wiki for the format of the libname). This pulls the data to SAS and you can then view the tables in the left panel underneath the libname.

See more from StatsNZ - What data is in the IDI?

Extracting data

It is recommended that you make your initial data extractions using SQL. This can be done through SQL server, or from within SAS using the 'passthrough' method described on the wiki (under the heading 'An intro to SAS explicit passthrough queries'). It is also possible to extract data from the IDI using other means (eg a SAS data step) but it will be slower than using SQL. That said, if you only want to extract a small amount of data, it probably won't make much difference.

For R users, there is an example of how to connect to IDI using R in the code sharing area. Alternatively, you can make the initial extraction using SQL and then save a csv or similar file to read into R.

Where to save your work

Once you have completed the initial extraction, you can save your dataset and continue to use SAS/STATA/R as you normally would. If you are saving your dataset as an SQL table, you can store it in the sandpit (if you do the initial extraction through SQL server, this is your only option). SAS datasets, csv files and other formats can be stored in your Datalab project folder. You can find your project folders in the I: drive. The project folder can also be used to store code, notes, tables and write-ups, and any other documents you are using within the Datalab. If you request the microdata access team to transfer any files to the Datalab environment for you, they will put it into your project file. One Datalab folder is allocated to each microdata access application (or project) and is

shared with all others working on that project. If your application is large and contains several sub-projects it is suggested you set up sub-folders within the Datalab folder.

Some examples of information you might need, and where to find it

Age and date of birth

Age and sex can be found in the personal detail table. This table contains all individuals who have a record anywhere in the IDI. Month and year of birth is available, but not day of birth. If an individual has different sex or date of birth recorded in different data sources, IDI has an algorithm for deriving the values in the personal detail table (details available on the wiki).

Ethnicity

Recent refreshes of IDI (from September 2018 onwards) provide ethnicity in the personal detail table, drawn from the highest quality source available. Census is considered the highest quality, followed by DIA and then health. In older refreshes, this same information can be found in the source_ranked_ethnicity table. More details about this table are available from the VHIN ethnicity guide, on the Wiki or Meetadata.

For further information on ethnicity in the IDI see - Reid, G, Bycroft, C, Gleisner, F (2016). Comparison of ethnicity information in administrative data and the census. Retrieved from http://archive.stats.govt.nz/methods/research-papers/topss/maori-info-admin-data.aspx

Geographic location

Information about geographic location is recorded in several sources in the IDI. As of 2018, there are seven address sources: IRD, ACC, Health (NHI and PHO), Education, MSD (residential and postal), and Census. At any given time, an individual may have several different addresses recorded with different administrative providers (if, for example, they have moved house and updated their address with their GP, but not with the IRD). These might be different types of addresses, for example MSD has both residential and postal addresses. You will need to think about which sources and types are best suited to your needs.

IDI geocodes all address information that they receive to a co-ordinate point. The variables that contain this information are called 'snz_idi_address_register_uid' or similar. The coordinate points identify a land parcel (roughly equivalent in most cases to an address or dwelling). However, these have been encrypted so that they are not identifiable. The same encryption key has been used on different IDI datasets (so, for example, you can be sure that location '10076534' in health data is the same location as '10076534' in tax data). However, the encrypted locations cannot reveal the actual location of a dwelling.

Locations are also coded to statistical areas, such as meshblock, territorial authority, and regional council area. These are not encrypted.

Geographic location information is available in the source data collections, and is also combined into two tables in IDI. The 'address_notification_full' table contains a list of all address updates that have been recorded in all IDI data sources, along with a flag indicating the data source. The 'address_notification' table uses prioritisation to get a single geographic location for an individual at any given time. The prioritisation process is described in IDI geographic tables metadata on the wiki.

When geographic information is needed use the 'address_notification_full' table, and select the most recently updated address at the reference date. This method has been shown to provide a

more accurate address (based on comparisons against addresses recorded in census data) than a prioritisation method (see http://archive.stats.govt.nz/methods/research-papers/topss/quality-geo-info-idi.aspx). For further information see the VHIN guide on geographical information.

Leaving/entering NZ

You might be interested in identifying whether or not an individual was in NZ at a given date. The easiest place to get this information is from the person_overseas_spell table in the 'data' schema. This table lists periods spent overseas by each individual in the IDI, from 1998 onwards (migration data in IDI does not go further back than 1998). Each row in the table is one overseas spell (so an individual might have many rows in the table). For each overseas spell there is a start date, end date, and length (in days). There are default dates used for a long time in the past or a long time in the future for those arriving in the country for the first time, or those who have left but not returned. The table is derived from the MBIE migration tables, which are a list of all border movements in and out of New Zealand (data are collected from passports, not from the departure/arrival card that you fill out. IDI does hold the card data, but it is a different collection called 'New Zealand Customs Service'). If you need more information than what is in the person_overseas_spell table, you will have to get access to the MBIE and/or Customs tables (MBIE is better for some purposes, Customs for others, you will need to talk to StatsNZ about which collection would suit your needs better). But for many people, the information in the person_overseas_spell table will be enough.

Getting data out of the datalab

The Datalab is a secure environment and you cannot remove data, tables, graphs, etc without (1) applying confidentialisation in accordance with StatsNZ's rules, and (2) having it checked by StatsNZ checkers.

Important: There is a 5+ working day turnaround for checking and releasing outputs from the datalab. If you need some numbers, graphs or tables for a presentation or report, you need to plan ahead.

The process for confidentialising and releasing datalab outputs is described in the Microdata Output Guide. If you have any questions, don't be afraid to ask the microdata access team. SNZ has the right to revoke your datalab access if you break the rules, so it is in your interests to make sure you are following them correctly.

By Sheree Gibb, Vivienne Rijnberg and Andrea Teng

Version: Original 10 March 2016, last updated 29 January 2019.