Loading your own data into the IDI

Frequently researchers want to add data into the Datalab environment or link their own datasets to the IDI or LBD in order to answer specific research questions. This blog post gives some tips about the ways you can add external data to the IDI. StatsNZ has published some helpful advice about this process on their website. This post will build on this StatsNZ advice from an IDI data user perspective.

Ethics approval

The first step is to ensure you have ethics approval for your study. This is done by applying to the relevant committee at your institution, for example the University ethics committee. Ensure that the ethics committee has information on the data linkage and what consent you have for this data linkage. StatsNZ may also require you to contact the Health and Disability Ethics Committee to assess whether your project is in scope. Frequently these IDI data analyses are out of scope. StatsNZ will require evidence of this from HDEC (eg an email).

Requests to save external information in your Datalab folder

Whilst in the Datalab environment you may want to use external information to help with your research; for example statistical code, written documents or simple datasets. This type of information does not require StatsNZ to do any linkage to the other IDI and LBD datasets. You can simply email your files and request the Microdata Access team to copy them into your Datalab folders, giving the appropriate folder name. This is the recommended process for bringing in statistical code into the Datalab environment (eg SIAL code, format variables etc).

This process is not appropriate if you want to link individual record data to the IDI. There are privacy implications, consent considerations and a specific process developed by StatsNZ that has been designed for this process. This is described below.

StatsNZ IDI dataload process for linking a dataset to the IDI

In 2017, StatsNZ introduced an application process to load datasets into the IDI to both formalise and prioritise IDI ad hoc data loads. The process involves meeting with StatsNZ, completing the dataload application form to get StatsNZ approval and then getting a Memorandum of Understanding signed between StatsNZ and the data owners (Board level). An initial meeting with Stats and the data owners helps to ensure everyone is happy with what the project aims to achieve and understands what linked data is. The dataload application form requires uses to provide evidence of individual consent and consideration of the relevant privacy issues, as well as explaining the benefits of linking and a summary of the quality of the data. As well as applying for the dataload you will also need to submit a project application to use the IDI datalab.

There are two types of dataloads; full integration and ad hoc dataload. StatsNZ encourages data owners to give permission for the linked data to subsequently be used by other researchers. This is called full integration and can be done in a way that future researchers

have to get permission from the data owner to use the dataset in the IDI (as is currently done through the data application process for different government agency data custodians). In this way, future ad hoc data loads for the same data would not be required reducing the administrative costs to the data owners, StatsNZ and future researchers. The alternative is where data is loaded for one individual research project and this is called an ad hoc dataload. This is often more feasible if data owners are not comfortable to provide access for future research.

Data is usually transferred to StatsNZ by person to person transfer of an iron key (a secured USB drive). Access passwords should always be communicated separate to the transfer. The process by StatsNZ to link the data to the IDI can take upward of three months and depends on the prioritisation that the dataload is given and resources available at StatsNZ at the time of the transfer.

Data linking is done by StatsNZ via one of two ways; deterministically using government agency identifiers or probabilistically. Deterministic linkage requires the dataset to be submitted to StatsNZ with government agency identifiers (eg NHI). Probabilistic linkage is less exact and more resource intensive. A statistical programme links records based on similarity in names, sex, date of birth etc. Probabilistic linkage can sometime be avoided by sourcing government agency identifiers from other sources. For example if you don't have NHI numbers for your health research population you can contact the Ministry of Health to request they match your study participants to an NHI number for a small fee.

Once your data is linked it is saved into the sandpit schema for your Datalab project. The next step is to check the data for completeness and quality.

- Did all the records get linked as expected?
- How have missing data and zeros been treated?
- Each individual in the dataset should have a unique identifier. These should be the StatsNZ encrypted government agency identifiers and not the snz_uid, because that ID changes every refresh.
- How many records are present in the spine, and have been successfully matched to other datasets in the IDI?

To see what datasets are being loaded to the IDI see "Upcoming datasets for the IDI and LBD". This list may not be completely up-to-date or fully precise.

Linking data to the IDI can add significant value to a research project however the current process has a long lead in time and requires careful advanced planning to gain the full benefits. Full integration of new data is a way your research can support and contribute to future research and improve health and social policy and delivery in New Zealand.

Original post published 23rd May 2018, by Andrea Teng, Hayley Denison, Nevil Pierse