Stats NZ
Tatauranga Aotearoa

9 July 2021

Hiria Te Rangi
fyi-request-15512-4a14cd35@requests.fyi.org.nz

Dear Hiria,

Official Information Act request

Thank you for your Official Information Act request, dated 20 May 2021 for the following:

1. *I would like to see the data structure for IDI Clean as well as IDI Raw. I would also like to know which organisations have access to IDI Clean and who has access to IDI Raw or their employer.*

The Integrated Data Infrastructure (IDI) is a large research database. It holds data about people and households. The data is about life events, like education, income, benefits, migration, justice, and health. It comes from government agencies, Stats NZ surveys, and non-government organisations (NGOs). The data is linked together, or integrated, to form the IDI.

The IDI complements the Longitudinal Business Database (LBD), which holds linked data about businesses. Researchers use the IDI and LBD to gain insight into our society and economy. This research can help answer questions about complex issues that affect New Zealanders.

Both the IDI and the LBD are located in the Stats NZ Data Lab, a secure virtual desktop environment. The Data Lab is housed in the Stats NZ IT environment but it can be accessed via approved secure remote Data Lab rooms. Researchers that have been approved for access to the data for specific projects can use the remote data labs to work with the data.

You have asked "who has access to the IDI?"
Anyone can apply for access to use the data contained in the Data Lab, by going through an application and assessment process before approval. Once the application is accepted, researchers must complete a training course and sign legally binding documents to assure that the data will be used appropriately. Further information on the application process can be found on the Stats NZ website (https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure#how-apply).

The list of organisations that have, or have had access, to the IDI and LBD are listed in Appendix A. These organisations may have more than one project running concurrently in the Data Lab. A list of project summaries and published outcomes of projects are available on the Stats NZ website (https://cdm20045.contentdm.oclc.org/digital/collection/p20045coll17).

You have asked "what is the data structure of the IDI?"
The IDI is a large database that contains many different datasets that have been linked together. A list of all dataset sources available in the IDI is listed in the attached file 'Data in the IDI, March 2021' (also available https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/data-in-the-idi/). Each project application will only be allowed to access datasets that are relevant to the project objectives.

In the linking process a 'spine' is created that represents a base population for all individuals who have been resident in New Zealand. This spine is created from linking tax data (from IR), births data (from DIA) and migration data (from MBIE). Each dataset that makes up the IDI is then linked to the spine. This linking is completed by using a mix of deterministic linking (where unique identifiers like NHI or IRD numbers are present) and probabilistic linking (where non-unique identifiers such as name and date of birth are used). The resulting data structure is known as IDI Raw. The datasets then goes through a process of de-identification to create IDI Clean, which is used by researchers.

IDI Clean

The data that the researchers access in the Data Lab is known as IDI Clean. This is to differentiate it from the IDI Raw data that is used to create the IDI. IDI Clean data has been anonymised, meaning that any identifiable information about the people in the data sets, such as name, address or day of birth has been removed. By using the data that their project has access to, researchers are able to build populations of interest and investigate outcomes and points of interest across the data sets. To publish information from the Data Lab, researchers must go through a process of output checking where trained Stats NZ staff ensure that confidentiality measures have been applied to the data to maintain the privacy of people's information.

IDI Raw

IDI Raw is the name used for the data that Stats NZ receives and uses for the creation of the Integrated Data databases (IDI and LBD). This data is received from the data supplier in its original format and is stored securely in a separate system to the Data Lab. This data is standardised and cleaned, then goes through the linkage process each time the IDI is refreshed (quarterly). This data is only accessed by approved Stats NZ staff for either the purposes of creating the IDI and LBD, or for supporting the production of official statistics.

More information about Integrated Data can be found on the Stats NZ website www.stats.govt.nz. Whilst we feel that we have answered your questions, our Integrated Data team is also available to do onsite or video link presentations. You can contact contact them via access2microdata@stats.govt.nz.

If you are not satisfied with this response, you have the right, by way of complaint to the Office of the Ombudsman under section 28(3) of the Act, to seek an investigation and review of this response to your request.

Stats NZ intends to publish its response to your request made under the Act on the Stats NZ website. This letter, with your personal details removed, will be published in its entirety. Consistent with the Act, publishing responses increases the availability of information to the public and helps promote balanced public debate.

We would be happy to discuss any further questions you may have in relation to this request. If you would like to discuss this further, please contact me in the first instance.

Yours sincerely,

Stallah Valaau
Advisor – Office of the Chief Executive