

IDI Search

*A metadata search app for exploring
New Zealand's administrative linked data*

Tom Elliott

Collaborators

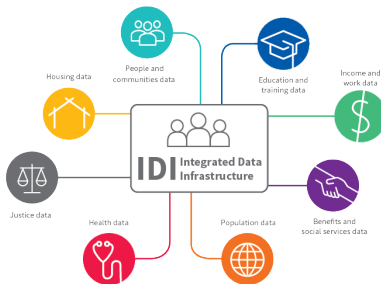
Barry Milne, Eileen Li, Andrew Sporle, and Colin Simpson

Developed by: Te Rourou Tātaritanga terourou.org
Ongoing support: iNZight Analytics Ltd inzight.co.nz



IPDLN Chicago
September 2024

The Integrated Data Infrastructure (IDI)



- ▶ Large research database, de-identified microdata
- ▶ Cross-sector research

What's in the IDI?

- ▶ $\approx 50,000$ variables, $\approx 1,000$ datasets
- ▶ Data dictionaries available in Data Lab
- ▶ “Is information on X available in the IDI?”
- ▶ Can we build a searchable index?

IDI data sets

- ▶ Admin datasets linkable at the individual level
- ▶ *Clean* contains the routinely (3x yearly) updated data
- ▶ *Adhoc* contains one-off or more frequently updated data (often with timestamped names)

⇒ Generate a list of variables from the schema

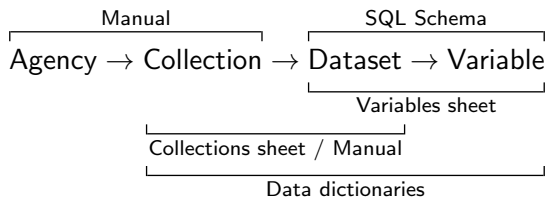
Data dictionaries

- ▶ Excel workbooks
 - ▶ Detailed descriptions
 - ▶ Variable names, coding information, etc.
- ▶ Independent update/maintenance across agencies
 - ▶ Inconsistent formats, structures, typos/errors, etc.
 - ▶ Duplicated templates ⇒ duplicated errors
- ▶ Only available to approved, existing researchers

Other data sources

- ▶ Manually curated datasets by us
- ▶ Bridge the gap between IDI schema and data dictionaries
- ▶ Lists of collections without dictionaries
- ▶ Renamed datasets/variables
- ▶ Agency names and associated collections
- ▶ Regex patterns for date-stamped naming conventions

Data hierarchy



Data preparation



- ▶ free, open-source
- ▶ can read/write many formats
- ▶ powerful data manipulation tools
- ▶ easy to script

Deployment



- ▶ cloud hosted (planetscale)
- ▶ structured data
- ▶ text search



- ▶ Hosted on Vercel (free!)
- ▶ Easy to set-up and deploy
- ▶ Javascript framework (ReactJS)

Read data dictionaries

- ▶ **R** script to parse Excel workbooks
- ▶ Extract collection, dataset, and variable metadata
- ▶ Detect issues/errors - modify script or fix manually
 - ▶ After updating, sent back to Stats NZ
 - ▶ More complex issues returned to Stats NZ to fix
- ▶ Identify issues e.g., duplicate variables

Read SQL variable list

- ▶ All refreshes and adhoc lists
- ▶ Contains database, table, variable name
- ▶ Some info about variable type/size

Combine with metadata

- ▶ Link variables and datasets (some regex)
- ▶ Link datasets with collections (some manual)
- ▶ Link collections and agencies
- ▶ Store other metadata (e.g., name changes)
- ▶ Clean up duplicates, formatting, special characters
- ▶ Write to MySQL database

[Search](#)
[Clear](#)


Data Supply Agencies (25)

[CSV](#) [JSON](#)

Name
Accident Compensation Corporation
Auckland City Mission
Page 1 of 13



Collections (106)

[CSV](#) [JSON](#)

Name	Agency	
ARCOS	University of Auckland	●
Centre of Innovation and Entrepreneurship Participati	University of Auckland	●
NZ Rugby Representatives	NZ Rugby	
Page 1 of 36		

Datasets (1046)

[CSV](#) [JSON](#)

Name	Collection / Agency	
Client acc_clean.clients	IDI ACC Injury Claims data Accident Compensation Corporation	
Claims acc_clean.claims	IDI ACC Injury Claims data Accident Compensation Corporation	
Claims historic acc_clean.claims_historic	IDI ACC Injury Claims data Accident Compensation Corporation	
Medical codes acc_clean.medical_codes	IDI ACC Injury Claims data Accident Compensation Corporation	
Addresses acc_clean.addresses	IDI ACC Injury Claims data Accident Compensation Corporation	
<div><div></div><div>Page 1 of 210</div><div></div></div>		

Variables (50049)

Name	Dataset / Collection
1st_readmit_date 1st_readmit_date	IDI UoO REGIONS care National Stroke Register
1st recurrent stroke date 1st recurrent stroke date	IDI UoO REGIONS care

Welcome to the IDI Search App

The IDI Search App allows researchers to search for variables that are available in the IDI and, in some cases, metadata about these variables. The app uses data from IDI variables and Data Dictionaries shared with us by Stats NZ. The data are stored in a database which can then be searched using the web app. For help navigating the app, click **Help** in the top right corner.

Use the search box to enter terms to filter. To search multiple terms, prefix each word with a plus (+) sign. For example, to search for records that contain both the words "income" and "employment", enter "+income +employment". See [Help](#) for more information.

App updated 27 August 2024 [see changelog](#)
Database updated 4 April 2024
Latest refresh April 2024

Proudly supported by



Conclusion

Restricted data and data dictionaries

We grant access to the following data on a case-by-case basis where a research project meets access criteria. This ensures approved researchers only get access to the IDI data essential for their project.

[Apply to use microdata for research](#)

Email access2microdata@stats.govt.nz to get the data dictionary for a dataset.

Alternatively, IDI researchers can explore what data is in the IDI along with their basic metadata through the [IDI Search | What's in the IDI? \(terourou.org\)](#) web app developed by [Te Rourou Tātaritanga](#).

- ▶ 10-30 users per day (200 users per month)
- ▶ Listed as a go-to resource for new researchers
- ▶ Database funded by Stats NZ
- ▶ Data dictionary coverage increased, quality improved
- ▶ Mentioned in 'Intro to the IDI' training

Future Work

- ▶ Automate data dictionary update/processing
- ▶ Secure funding for ongoing, routine updates
- ▶ New features



`idisearch.terourou.org`
`github.com/tmelliott/idi-search-app`