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

Tom Elliott

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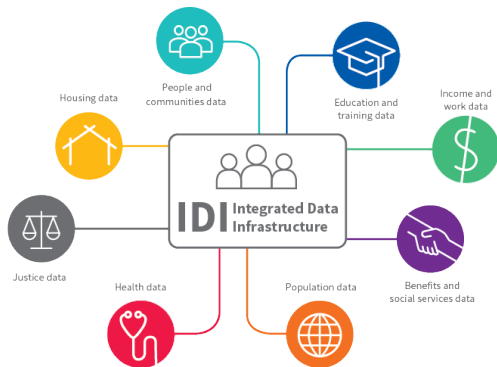Barry Milne, Eileen Li, Andrew Sporle, and Colin Simpson

IPDLN Chicago
September 2024

# The Integrated Data Infrastructure (IDI)



An integrated database containing de-identified longitudinal microdata about people, households, & businesses.

# What's in the IDI?

- ▶ $\approx 50,000$ variables, $\approx 1,000$ datasets
- ▶ Government agencies, surveys, and non-government organisations
- ▶ Linkable at the individual level

# The problem: access to metadata

- ▶ Data dictionaries contain metadata about collections of datasets
- ▶ Independent maintenance across agencies
  - ▶ Inconsistencies, typos, etc.
  - ▶ Duplication
- ▶ Only available to approved, existing researchers
- ▶ Not easily navigable, difficult to keep up with changes

Introduction
000

Creating a solution
●00000

IDI Search
00000

Summary
0000

## Why a search app?
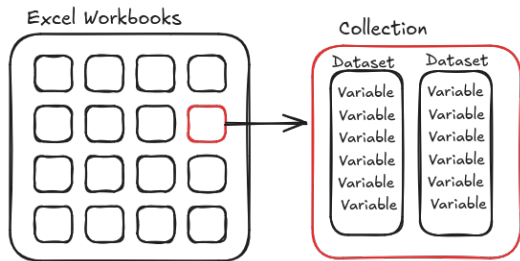
▶ A publicly available, searchable database
▶ What data is available?
▶ What has changed?

⇒ Increase utility of a valuable database

Introduction
○○○

Creating a solution
○●○○○○

IDI Search
○○○○○

Summary
○○○○

## Data sources

- Database schema
- Data dictionaries
- Manually curated metadata
  - Agency names and associated collections
  - Renamed datasets/variables
  - Datestamped tables

Introduction
000

Creating a solution
000●000

IDI Search
00000

Summary
0000

# Read data dictionaries



Excel Workbooks

Collection

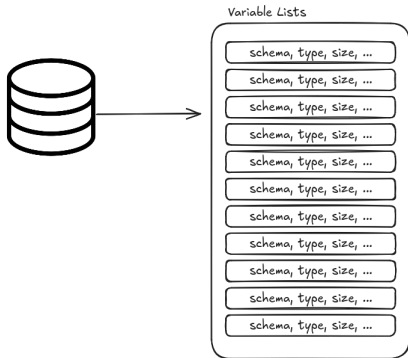Dataset   Dataset

- Search workbook for keywords
- Sanitise data
  - Minor errors fixed
  - Otherwise rejected
- Returned to Stats NZ (usually fixed!)

Introduction
ooo

Creating a solution
ooo●oo

IDI Search
ooooo

Summary
oooo

# Read SQL variable list



Variable Lists

schema, type, size, ...
schema, type, size, ...
schema, type, size, ...
schema, type, size, ...
schema, type, size, ...
schema, type, size, ...
schema, type, size, ...
schema, type, size, ...
schema, type, size, ...
schema, type, size, ...
schema, type, size, ...

▶ Stats NZ generate at each update

▶ Standard IDI output process

▶ Table and variable IDs (schema)

Introduction
000

Creating a solution
000000
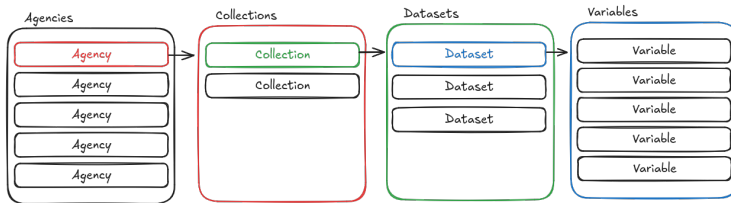
IDI Search
00000

Summary
0000

# Combine with metadata
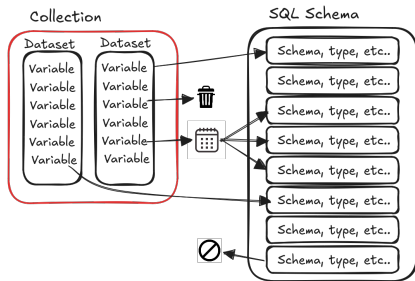


- ▶ Variables in SQL schema missing from dictionaries
- ▶ Variables in dictionaries missing from SQL schema (deleted? error?)
- ▶ Datestamped variables

Introduction
○○○

Creating a solution
○○○○○●

IDI Search
○○○○○

Summary
○○○○

# Final clean data



▶ Text fields cleaned/formatted

▶ Duplicates removed

▶ Written to MySQL database

Introduction
○○○

Creating a solution
○○○○○○

IDI Search
●○○○○

Summary
○○○○

# Deployed solution: IDI Search app

▶ UI: NextJS (Javascript framework)

▶ MySQL text search

▶ Cross-linking for easy navigation

▶ `idisearch.terourou.org`

| ▽ | Enter search term to filter results | Search | Clear | ⚙ |

## Welcome to the IDI Search App

**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 Participation | University of Auckland | ● |
| NZ Rugby Representatives | NZ Rugby | |

| « | ‹ | Page | 1 | of 36 | › | » |

**Datasets (1046)**   ☁ CSV  JSON

| Name | Collection / Agency | |
| --- | --- | --- |
| Client<br>acc_clean.clients | IDI ACC Injury Claims data<br>Accident Compensation Corporation | ● |
| Claims<br>acc_clean.claims | IDI ACC Injury Claims data<br>Accident Compensation Corporation | ● |
| Claims historic<br>acc_clean.claims_historic | IDI ACC Injury Claims data<br>Accident Compensation Corporation | ● |
| Medical codes<br>acc_clean.medical_codes | IDI ACC Injury Claims data<br>Accident Compensation Corporation | ● |
| Addresses<br>acc_clean.addresses | IDI ACC Injury Claims data<br>Accident Compensation Corporation | ● |

| « | ‹ | Page | 1 | of 210 | › | » |

**Variables (50049)**   ☁ CSV  JSON

| Name | Dataset / Collection | |
| --- | --- | --- |
| 1st_readmit_date<br>1st_readmit_date | IDI UoO REGIONS care<br>National Stroke Register | ● |
| 1st_recurrent_stroke_date<br>1st_recurrent_stroke_date | IDI UoO REGIONS care<br>National Stroke Register | ● |
| ABGROUP<br>abgroup | IDI MOE PISA 2009<br>IDI MOE PROGRAMME FOR INTERNATIONAL STUDENT ASSESSMENT (PI... | ● |

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

**inzight analytics**

▽ | housing | | Search | Clear | ⚙

## Data Supply Agencies (0)    ⟳ CSV JSON

| Name |
| --- |
| No data |

«  ‹    No data    ›  »

## Collections (7)    ⟳ CSV JSON

| Name | Agency | |
| --- | --- | --- |
| IDI Census of Population and Dwellings data 2013 | Statistics New Zealand | ● |
| IDI Census of Population and Dwellings data 2018 | Statistics New Zealand | ● |
| IDI General Social Survey Ongoing Supplements | Statistics New Zealand | ● |

«  ‹    Page 1 of 3    ›  »

## Datasets (19)    ⟳ CSV JSON

| Name | Collection / Agency | |
| --- | --- | --- |
| pilot housing survey 1819<br>clean_read_branz.pilot_housing_survey_1819 | *Not Available* | |
| Individual<br>cen_clean.census_individual_2018 | IDI Census of Population and Dwellings data 2018<br>Statistics New Zealand | ● |
| Dwelling<br>cen_clean.census_dwelling_2018 | IDI Census of Population and Dwellings data 2018<br>Statistics New Zealand | ● |
| IDI COR Programmes<br>cor_clean.ra_programmes | Offender Management Periods<br>Department of Corrections | ● |
| IDI ACM Housing and Statistics<br>acm_clean.acm_housing | IDI Auckland City Mission data<br>Auckland City Mission | ● |

«  ‹    Page 1 of 4    ›  »

## Variables (85)    ⟳ CSV JSON

| Name | Dataset / Collection | |
| --- | --- | --- |
| acm_has_HNZ_app_ind<br>acm_has_hnz_app_ind | IDI ACM Housing and Statistics<br>IDI Auckland City Mission data | ● |
| acm_has_housing_status_code<br>acm_has_housing_status_code | IDI ACM Housing and Statistics<br>IDI Auckland City Mission data | ● |
| acm_has_pays_rent_to_code<br>acm_has_pays_rent_to_code | IDI ACM Housing and Statistics<br>IDI Auckland City Mission data | ● |
| acm_ose_housing_status_code<br>acm_ose_housing_status_code | IDI ACM Outreach Service Episodes<br>IDI Auckland City Mission data | ● |

---

Close ✕

DATASET
# Dwelling

In collection: **IDI Census of Population and Dwellings data 2018**
Data Supply Agency: **Statistics New Zealand**

The 2018 Census dwelling table includes variables, derived variables and source indicators for variables where applicable across topics: population structure, location, housing, transport.

### Linking variables  ⓘ

⊗ snz_cen_dwell_uid (10)

⊗ snz_idi_address_register_uid (45)

### Variables in this dataset (71)    ⟳ CSV JSON

Sort by: Dictionary order | Name

| Name | Dataset / Collection | |
| --- | --- | --- |
| cen dwl sector of llord imp flag<br>cen_dwl_sector_of_llord_imp_flag | Dwelling<br>IDI Census of Population and Dwellings data... | ● |
| cen dwl wkly rent amnt imp flag<br>cen_dwl_wkly_rent_amnt_imp_flag | Dwelling<br>IDI Census of Population and Dwellings data... | ● |
| Dwelling ID<br>snz_cen_dwell_uid | Dwelling<br>IDI Census of Population and Dwellings data... | ● |
| Online form indicator - dwelling<br>cen_dwl_mode | Dwelling<br>IDI Census of Population and Dwellings data... | ● |
| Bilingual paper form indicator - dwelling<br>cen_dwl_blngul_ind | Dwelling<br>IDI Census of Population and Dwellings data... | ● |
| Confidentiality key<br>cen_dwl_conf_key | Dwelling<br>IDI Census of Population and Dwellings data... | ● |
| Census night number of occupants<br>cen_dwl_nbr_occupants_code | Dwelling<br>IDI Census of Population and Dwellings data... | ● |
| Main types of heating used - single/combi<br>cen_dwl_heat_app_sngl_comb | Dwelling<br>IDI Census of Population and Dwellings data... | ● |
| Main types of heating used (total respons)<br>cen_dwl_heating_app_code | Dwelling<br>IDI Census of Population and Dwellings data... | ● |
| Number of heating fuels used | Dwelling | |

| ▽ | snz_sex_code | Search | Clear | ✕ |

**Advanced search configuration**

**Include results from:**
- ☑ Refreshes
- ☑ Adhoc

**Exact search:**
- ☑ Only show results that exactly match the search string

### Data Supply Agencies (0)
⊕ CSV JSON

**Name**

| « ‹ | No data | › » |

### Collections (0)
⊕ CSV JSON

| **Name** | **Agency** | |
| « ‹ | No data | › » |

### Datasets (0)
⊕ CSV JSON

| **Name** | **Collection / Agency** | |
| « ‹ | No data | › » |

### Variables (33)
⊕ CSV JSON

| **Name** | **Dataset / Collection** | |
|---|---|---|
| msd_snz_sex_code<br>msd_snz_sex_code | IDI MSD Customer Aliases Linking Information<br>IDI Ministry of Social Development Client Identity data | ● |
| nia_links_snz_sex_code<br>nia_links_snz_sex_code | IDI POL NIA Links<br>IDI NZ Police NIA Links data | ● |
| nzcvs snz sex code<br>nzcvs_snz_sex_code | matching | ● |
| nzta_snz_sex_code<br>nzta_snz_sex_code | IDI NZTA DLR Historic<br>IDI NZ Transport Authority Drivers Licence and Motor Vehicles Regis... | ● |
| nzta_snz_sex_code<br>nzta_snz_sex_code | IDI NZTA Drivers Licence Register<br>IDI NZ Transport Authority Drivers Licence and Motor Vehicles Regis... | ● |
| nzta_snz_sex_code<br>nzta_snz_sex_code | IDI NZTA Motor Vehicle Register<br>IDI NZ Transport Authority Drivers Licence and Motor Vehicles Regis... | ● |
| otfs_snz_sex_code<br>otfs_snz_sex_code | IDI OTFS Client Identities<br>IDI Oranga Tamariki Family Start data | ● |
| partner snz sex code<br>partner_snz_sex_code | IDI HLFS Survey of Working Life 2012<br>IDI HLFS Supplements | ● |
| resp parent snz sex code | IDI HLFS Childcare Child 2009 | ● |

---

Close ✕

VARIABLE
## msd_snz_sex_code

In dataset: **IDI MSD Customer Aliases Linking Information**
In collection: **IDI Ministry of Social Development Client Identity data**
Data Supply Agency: **Ministry of Social Development**

Sex variable concorded to the Statistics NZ standard.

Note: Organisations have a value of 3.

### SQL Information

| **Field Name** | msd_snz_sex_code |
|---|---|
| **Table Name** | msd_clean.msd_customer_alias |
| **Type** | varchar |
| **Size** | 100 |
| **Database** | IDI Clean |

### Refresh availability

| ⊘ Mar 2024 | ⊘ Oct 2023 | ⊘ Jun 2023 | ⊘ Mar 2023 | ⊘ Oct 2022 |
| ⊘ Jun 2022 | ⊘ Mar 2022 | ⊘ Oct 2021 | ⊘ Jul 2021 | ⊘ Apr 2021 |
| ⊘ Oct 2020 | ⊘ Jul 2020 | ⊘ Jan 2020 | | |

### This variable may have a different name in other refreshes:

- **msd_snz_gender_code**

▽  moe_nsi_sex_snz_code                                    Search   Clear   ✕

**Advanced search configuration**

| Include results from: | Exact search: |
| --- | --- |
| ☑ Refreshes | ☑ Only show results that exactly match the search string |
| ☑ Adhoc | |

### Data Supply Agencies (0)                                    ⊕ CSV JSON

| Name |
| --- |
| No data |

« <                                                              > »

### Collections (0)                                              ⊕ CSV JSON

| Name | Agency |
| --- | --- |
| No data | |

« <                                                              > »

### Datasets (0)                                                 ⊕ CSV JSON

| Name | Collection / Agency |
| --- | --- |
| No data | |

« <                                                              > »

### Variables (1)                                                ⊕ CSV JSON

| Name | Dataset / Collection | |
| --- | --- | --- |
| moe nsi sex snz code<br>moe_nsi_sex_snz_code | MOE National Student Index (NSI) data<br>MOE National Student Index (NSI) data | ● |

« <                          Page 1 of 1                          > »

Close ✕

VARIABLE
# moe nsi sex snz code

In dataset: **MOE National Student Index (NSI) data**
In collection: **MOE National Student Index (NSI) data**
Data Supply Agency: **Ministry of Education**

## SQL Information

| Field Name | moe_nsi_sex_snz_code |
| --- | --- |
| Table Name | moe_clean.nsi |
| Type | varchar |
| Database | IDI Clean |

## Value Codes ⓘ

| Code | Label |
| --- | --- |
| 2 | female |
| 1 | male |
| null | unknown |

## Refresh availability

| ⊘ Mar 2024 | ⊘ Oct 2023 | ⊘ Jun 2023 | ⊘ Mar 2023 | ⊘ Oct 2022 |
| --- | --- | --- | --- | --- |
| ⊘ Jun 2022 | ⊘ Mar 2022 | ⊘ Oct 2021 | ⊘ Jul 2021 | ⊘ Apr 2021 |
| ⊘ Oct 2020 | ⊘ Jul 2020 | ⊘ Jan 2020 | | |

Introduction
○○○

Creating a solution
○○○○○○

IDI Search
○○○○○

Summary
●○○○

# Outcomes

▶ 10-30 users per day ( 200 users per month)
▶ Listed by Stats NZ as a go-to resource

## 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.

Introduction
○○○

Creating a solution
○○○○○○

IDI Search
○○○○○

Summary
○●○○

# Outcomes

▶ Low-cost hosting funded by Stats NZ

▶ Easily updatable when new dictionaries, refreshes available

▶ Data dictionary coverage increased, quality improved

Variables:

Mar 2024 refresh     20298 (66%)

Adhoc     28632 (21%)

Datasets:

Mar 2024 refresh     444 (89%)

Adhoc     585 (13%)

Introduction
○○○

Creating a solution
○○○○○○

IDI Search
○○○○○

Summary
○○●○

## Next steps

- ▶ Automate data dictionary update/processing
- ▶ Secure funding for ongoing, routine updates
- ▶ New features
  - ▶ flag new datasets/variables
  - ▶ showing additional information e.g., application codes
  - ▶ extract information from other sheets in dictionaries

**Dr. Tom Elliott**
iNZight Analytics Ltd.
`tom@inzight.co.nz`

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