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

Introduction

The Integrated Data Infrastructure (IDI)

- Large research database
- De-identified microdata about people and households
- Cross-sector research $\rightarrow$ insight into society/economy
- Data Lab: secure facility providing access to the IDI

Data dictionaries

- Contain *metadata* about data in IDI
  - Detailed descriptions
  - Variable names, coding information, etc.
- Stored in Excel workbooks
  - Independent update/maintenance across agencies
  - Inconsistent formats, structures, typos/errors, etc.
- Audience: existing researchers
- Difficult to find general information
- Not available to new researchers developing research proposals

A searchable index

- New users often ask
  Is information on X available in the IDI?
- Can we construct a searchable index?
- How do we go about parsing the data dictionaries?

The data

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

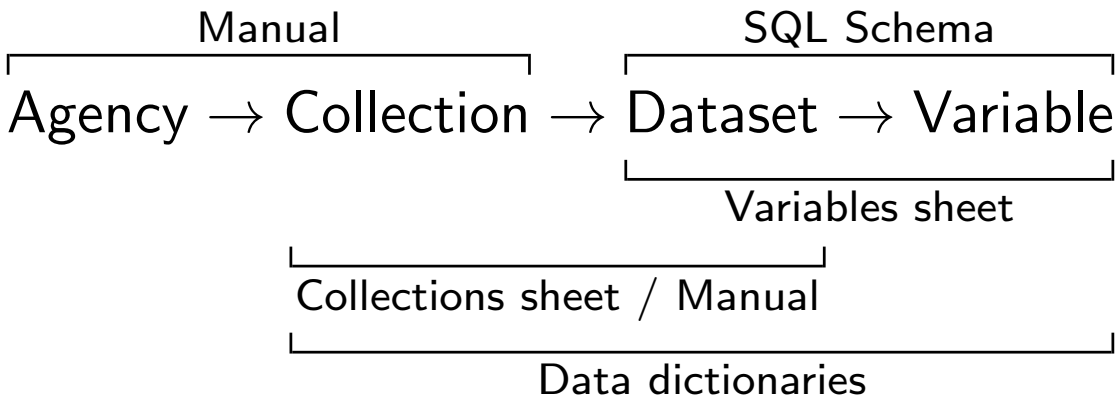
- Individual Excel workbooks per *collection* (one or more related datasets)
- Collection and dataset(s) description
- Spreadsheet of variables, descriptions, codings, etc.

Other 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



Tools

Data wrangling

- **R** statistical software
  - free, open-source
  - can read/write many formats
  - powerful data manipulation tools
  - easy to script

App Hosting

- **MySQL** database
  - cloud hosted (planetscale)
  - structured data
  - text search
- **NextJS** web framework
  - Hosted on Vercel (free!)
  - Easy to set-up and deploy
  - **ReactJS** - Javascript framework for interactive applications

The wrangling process

Demo

Conclusion