Integrated Data at Stats NZ

Stats NZ

- Stats NZ is the public service department of New Zealand charged with the collection of statistics related to the economy, population and society of New Zealand.
- Stats NZ manages the IDI and the LBD - two large research databases built from multiple data sources.

Hamish James

- General Manager – Customer Channels at Stats NZ.
- Leads team responsible for customer facing services and products, including New Zealand's Integrated Data Infrastructure.
- Began career working on quantitative history projects at the University of Otago.
- Spent a number of years in the UK, working at the UK Data Archive at the Arts and Humanities Data Service.
- Spent the last 14 years working in a variety of roles related to information management, strategy and customer support at Stats NZ.

Outline of presentation

- What are the IDI and LBD?
- How we operate the IDI and LBD
- How the IDI and LBD are being used
- Matching and linking data - challenges
- Discussion

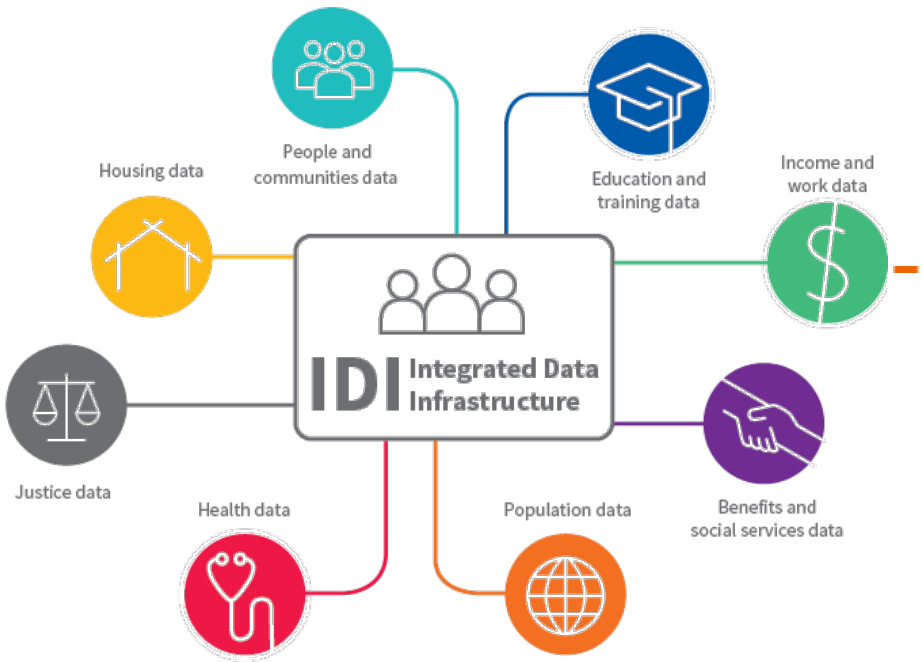
What are the IDI and LBD?

- Stats NZ has two large integrated databases containing de-identified longitudinal microdata. These can be used for research about issues that affect New Zealanders.
- The IDI contains data about people and households.
- The LBD contains data about businesses.

Integrated Data at Stats NZ

Integrated Data Infrastructure (IDI)

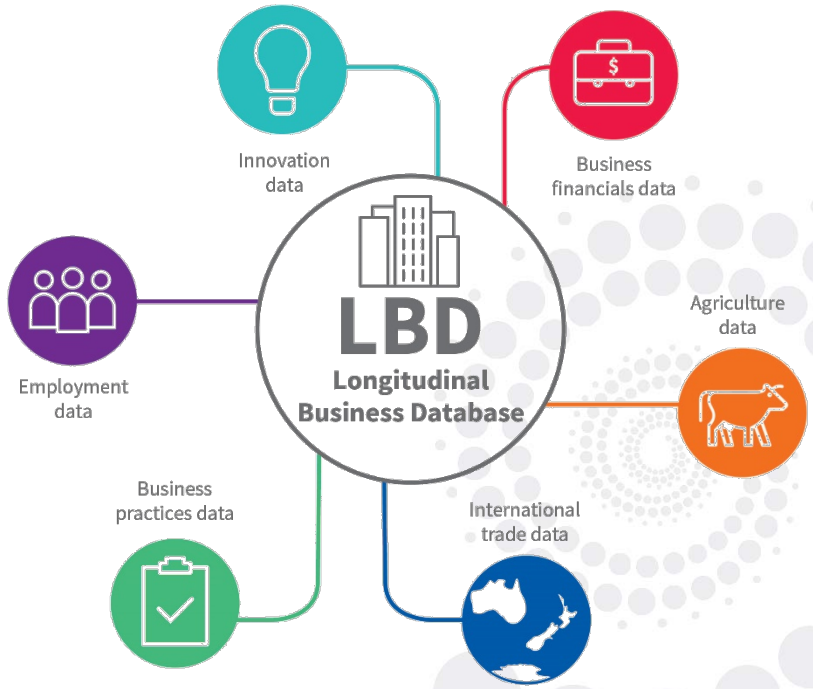
An integrated database containing de-identified longitudinal microdata about people & households.



IDI and LBD are linked through tax data

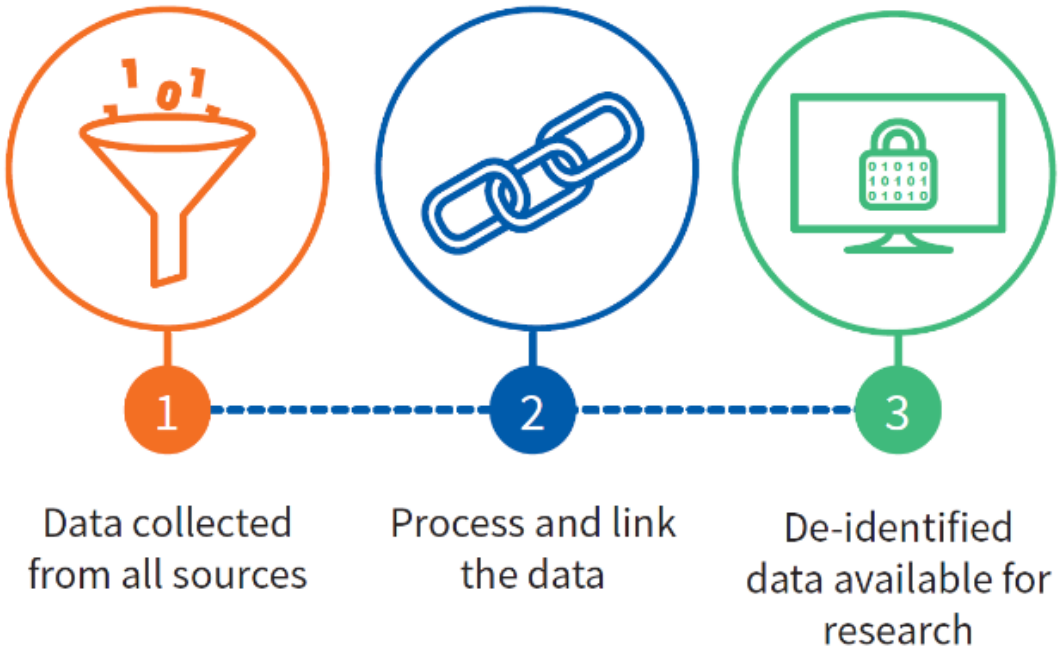
Longitudinal Business Database (LBD)

An integrated database containing de-identified longitudinal microdata about businesses.



How we operate the IDI and LBD

Flow of data in the IDI and LBD



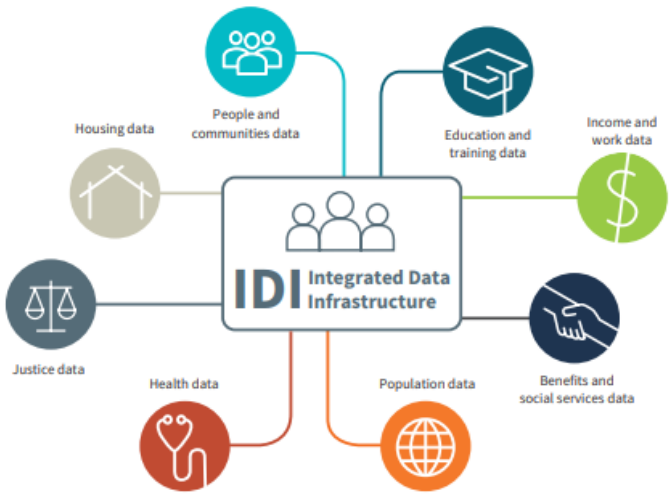
Data collected from all sources

Data collected
from all sources

Data in the IDI September 2018

Stats NZ Tatauranga Aotearoa

Stats NZ's Integrated Data Infrastructure (IDI) is a large research database containing de-identified microdata about people and households.



The IDI contains person-centred microdata from a range of government agencies, Stats NZ surveys including the 2013 Census, and non-government organisations. For more information about data in the IDI, see **www.stats.govt.nz/integrated-data/integrated-data-infrastructure**

The Longitudinal Business Database (LBD) complements the IDI with microdata about businesses. For more information about data in the LBD, see **www.stats.govt.nz/integrated-data/longitudinal-business-database**

Health data

- B4 School Checks – from 2011
- Cancer registrations – from 1995
- Chronic conditions – from 2007
- General medical services claims – from 2002
- Health tracker – 2006-13
- Laboratory claims – from 2003
- Mortality – from 1988
- Immunisation – from 2006
- National non-admitted patient collection – from 2007
- Pharmaceuticals – from 2005
- PHO enrolments – from 2003
- Population cohort demographics and addresses – from 2004
- Mental health and addiction – from 2008
- Publicly funded hospital discharges – from 1988
- National Needs Assessment and Service Coordination Information System (SOCRATES)
- Maternity – from 2003

Education and training data

- Early childhood education participation – from 2008
- Primary education – from 2007
- Secondary education – from 2004
- Tertiary education – from 1994
- Industry training – from 2001
- Targeted training – from 2001
- Adult competency assessments – from 2014

Benefits and social services data

- Benefits – from 1990
- Youth services – from 2004
- Children's Action Plan – from 1996
- Working for Families – from 2003
- Child, Youth, and Family – from 1991
- Student loans and allowances – from 1992
- ACC injury claims – from 1994
- Family Start – from 2008

Justice data

- Recorded crime: offenders – from 2009
- Recorded crime: victims – from 2014
- Court charges – from 1992
- Sentencing and remand – from 1998

People and communities data

- Auckland City Mission – from 1996
- Migrant Survey – from 2012
- Driver licence and motor vehicle registers
- Longitudinal Immigration Survey of NZ – 2005-09
- General Social Survey – 2008-2016
- Disability Survey – 2013
- Te Kupenga – 2013

Population data

- Border movements – from 1997
- Visa applications – from 1997
- Departure and arrival cards – from 1997
- 2013 Census
- Births, deaths, marriages, and civil unions – from 1840

Income and work data

- Tax and income – from 1999
- NZ Income Survey – from 2006
- Household Labour Force Survey – from 2006
- Survey of Family, Income, and Employment – 2002-10
- Household Economic Survey – from 2006

Housing data

- Tenancy – from 2000
- Social housing – from 1980

Stats NZ operates a five-safes environment, balancing privacy and confidentiality with data insights.

For information about applying to use the IDI or to learn about how we keep the data safe, see **www.stats.govt.nz/integrated-data**

De-identified data available for research

De-identified data available for research

How is the data kept safe?

We operate within a 'five safes' framework to ensure that access to the IDI and LBD is only provided if all of the following conditions can be met:

ID Tikanga framework (in development)

Safe people	Pūkenga (Expertise, Skills)	Whakapapa (Relationships)
Researchers can be trusted to use data appropriately	Researchers can demonstrate an awareness of and intention to work with data in culturally appropriate ways	Researchers have existing relationships with the communities the data comes from
Safe Projects	**Pono (Truth, Validity)**	**Tika (Correct, Accuracy, Fairness)**
The project has a statistical purpose and is in the public interest	Level of accountability to community of research is explained	Research should be part of a body of work that contributes towards better outcomes for Māori and NZrs
Safe Settings	**Kaitiaki (Guardians)**	**Wānanga (Repositories of knowledge)**
Ensuring the data is secure and preventing unauthorised access to the data	Decision-makers of the project are identified and Māori are involved in decision-making	Institutions have established systems, policies and procedures to ensure data is used in culturally appropriate and ethical ways
Safe Data	**Wairua (Spiritual essence of people)**	**Mauri (Life force principle)**
Personal information is not identified	Māori community objectives align with project research objectives	Level of transformation of the data from its original collection purpose is explained
Safe Output	**Noa (Ordinary, Unrestricted)**	**Tapu (Restricted, High sensitivity)**
Stats NZ results do not contain identifying results. Outputs must be confidentialised.	Accessibility of data and awareness of the impact on Māori	Sensitivities in the use of data are identified including privacy issues for whānau and identifiable community groups

April 19_Second iteration

How the IDI and LBD are being used

Researchers from:

- government agencies
- Universities
- NGOs
- …and more

Studying issues like:

- Vulnerable children
- Education and employment outcomes
- Impact of health conditions
- Business productivity
- …and more

Researchers currently using the IDI and LBD

There are currently 550 researchers using the IDI for 280 different research projects.

Some examples of research projects that have been conducted using data from the IDI:

- What happened to people who left benefit system during the year ended 30 June 2014 – *Ministry of Social Development, 2018*
- Impact of head injury on economic outcomes – *Victoria University of Wellington, 2019*
- Costs of raising children in New Zealand – *BERL, Business and Economic Research Ltd, 2019*

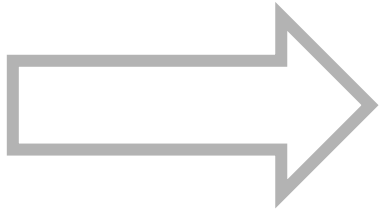
Case Study:

How Integrated Data Helps... Shine a light on the Gender Pay Gap

In work commissioned by the Ministry for Women, researchers from Auckland University of Technology (AUT) and Waikato used multiple methods to examine the gender pay gap.

The insights

- Researchers found a minimal gap between men and women for lower wages, but approximately a 20% gap at the top end.
- The average woman earns 4.4% lower hourly wages as a parent than if she hadn't had children, but there was no significant effect of parenthood for men.
- They found that even after accounting for a wide range of factors, close to 80% of the gap was unexplained.



Integrated data in action

Insights from Integrated Data have helped with many initiatives to help improve the gender pay gap.

Case Study:

How Integrated Data Helps... Child wellbeing

Social Workers in Schools (SWiS)

SWiS is a community social work service provided in most decile 1-3 primary and intermediate schools, and kura kaupapa Māori.

The Insights

- General pattern of improvements in students' outcomes in school and kura after the service was introduced.
- Indications that SWiS had an impact on stand-downs and suspensions from school, care and protection notifications, and police apprehensions for alleged offending.

Integrated Data in action

Using the Integrated Data Infrastructure, the study compares how students did before and after the SWiS programme expansion.

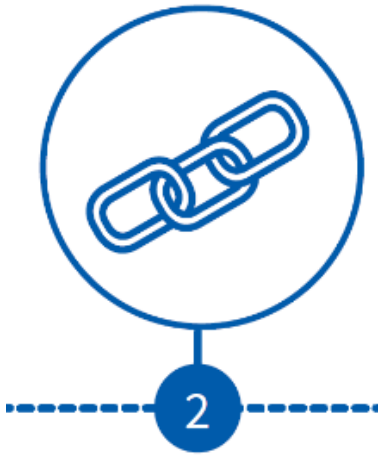
Benefits and limitations

- Researchers can tackle previously ‘unanswerable’ questions
- Longitudinal view
- Cross-sector view
- Geographical views
- Reduced research cost and burden

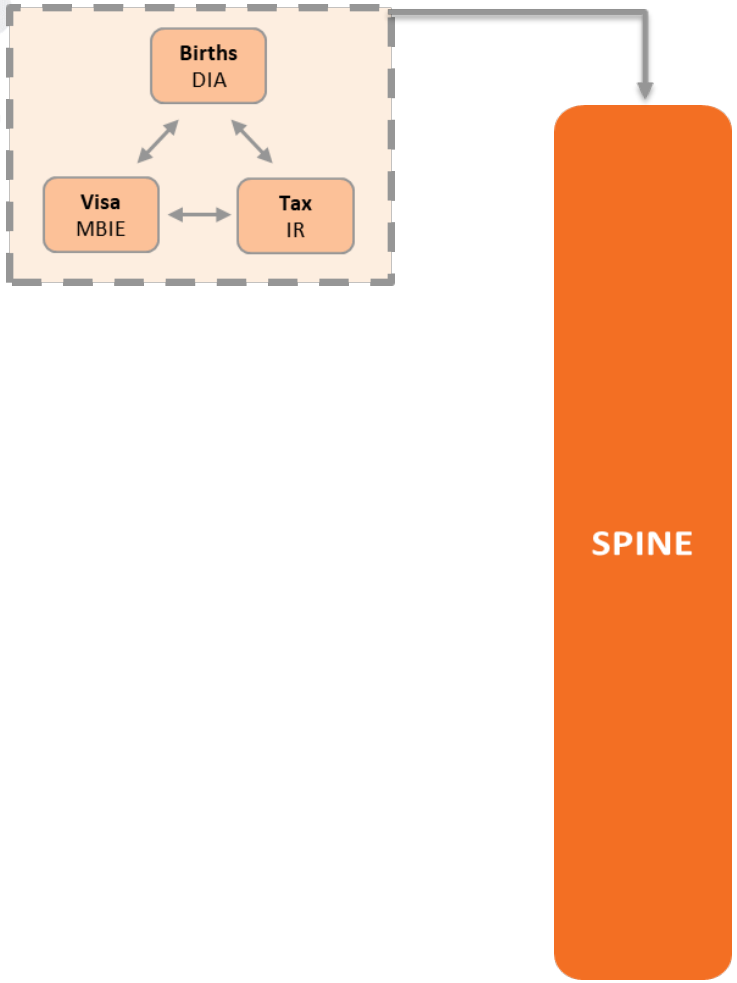
BUT…

- Administrative data quality issues
- High time and skill investment
- Small number studies limitations

Process and link the data

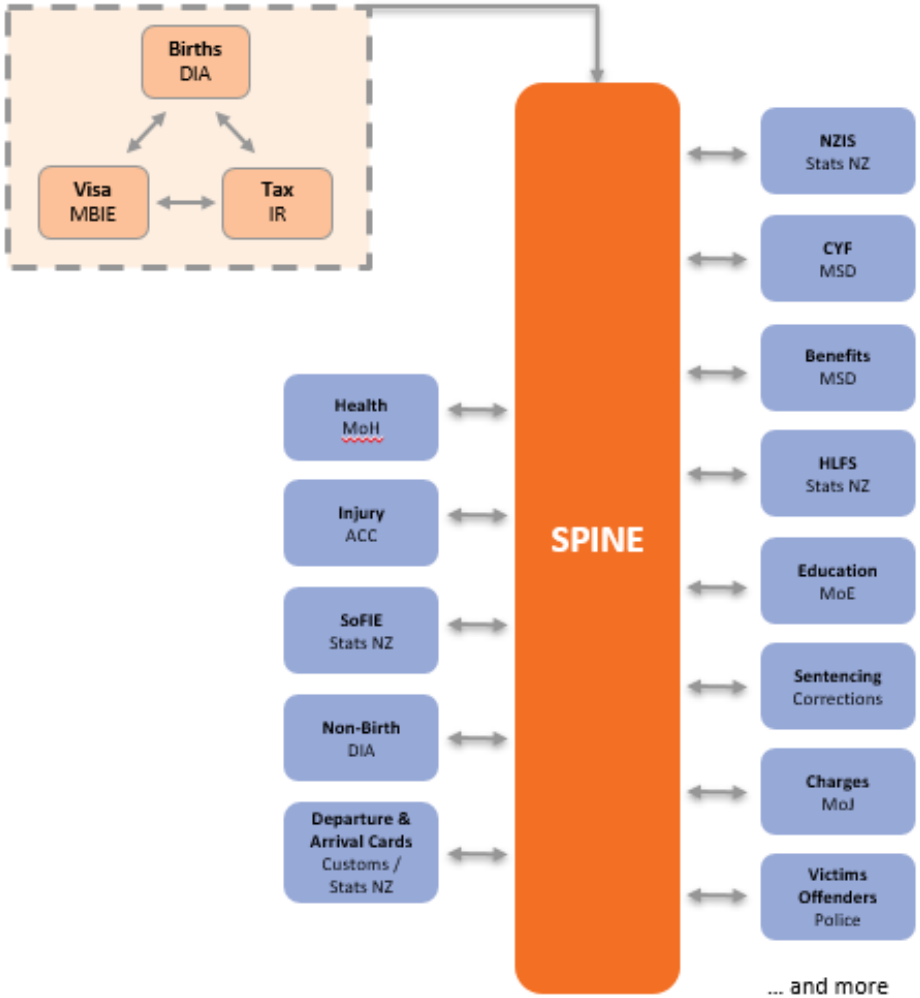


Linking datasets together



1. Spine created by linking births, tax, and visa data together

Linking datasets together



1. Spine created by linking births, tax, and visa data together

2. Other sources (aka nodes) are linked with the spine.

Two types of linking

Deterministic linking

Links records in different datasets based on a shared unique identifier (e.g. IRD number in employment and student loans).

Probabilistic linking

Best match based on key identifying variables such as name, business name, address, and date of birth.

Probabilistic matching

- Probabilistic record matching is so called because it relies on calculating scores or weights based on probabilities.
- The method involves measuring the agreements between the ‘linking variables’ in the two records, and also the disagreements.
- Linking variables are used to compare two records.
- A score or weight is calculated from the number of agreements minus the number of disagreements, and used to determine whether the record pair should be regarded as truly linked or not.

Probabilistic matching - example

Record	First Name	Last Name	Sex
A	Claire	Parker	M

Record	First Name	Last Name	Sex
B	Claire Mary	Jones	F

True Rec	First Name	Last Name	Sex
C	Claire Mary	Parker	F

No real data is used in examples

Comparison functions

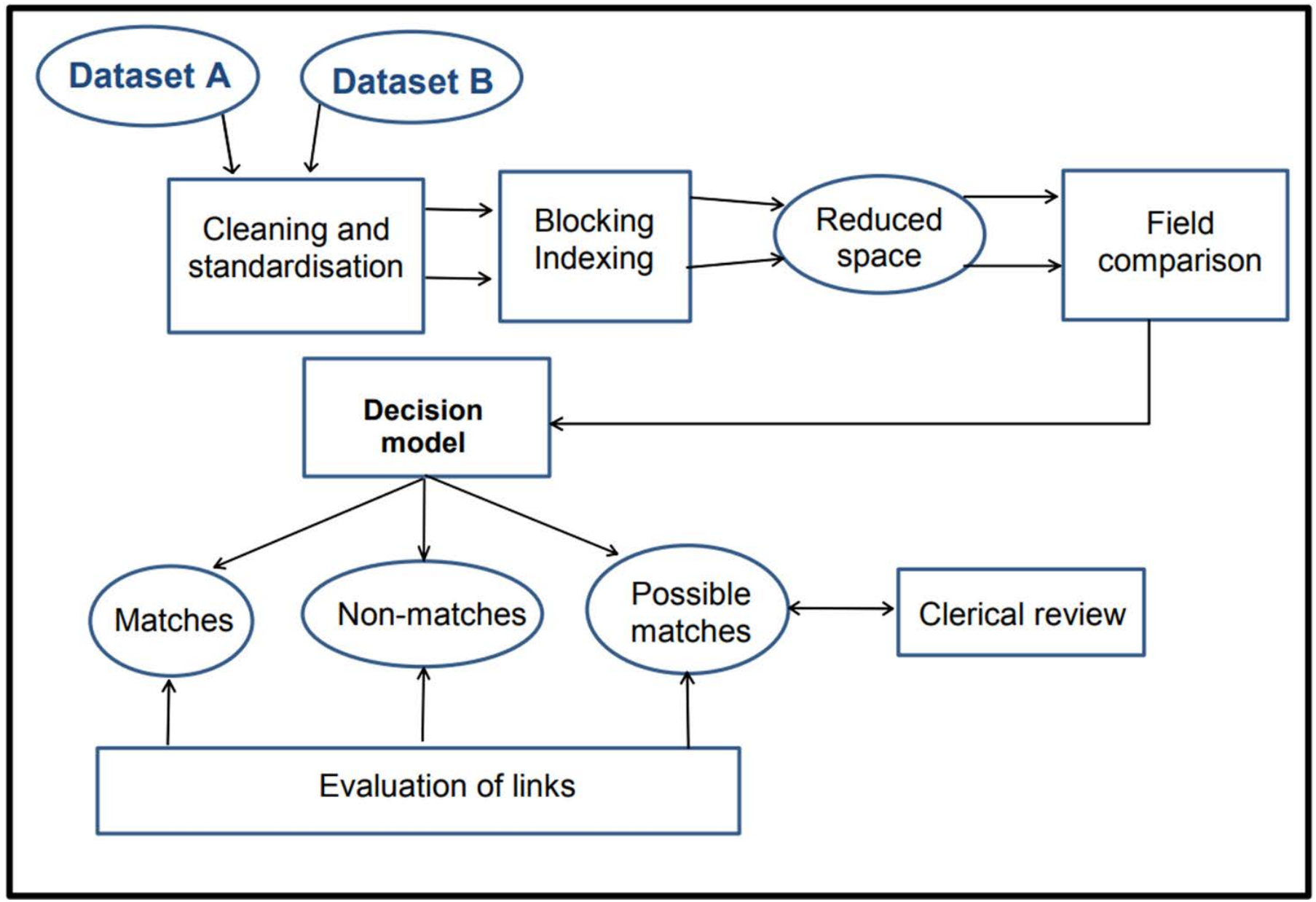
A way of comparing values to see if they're similar.

A comparison function for date might check for similarity between two dates, including by swapping the day and the month around to see if that gives a match

A comparison function for names might check for similarity using a sounding function to account for different spellings (e.g. SOUNDEX

- Edit distance comparisons such as Jaro-Winkler distance

Schematic representation of the record linkage process



Challenges with data in the IDI

Notable issues with admin data

- Admin data doesn’t have good coverage at certain ages. For example, DIA birth records only have parents' birthdates digitized after 1990.
- People may give different answers in different datasets - the same person may self-identify differently in Health vs Education data
- Even when using deterministic matching techniques, people can have more than one unique identifier. For example, you get a new IRD number if you go bankrupt.

Messy Data

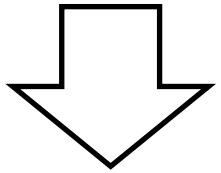
Admin data is often untidy. It can contain strange characters in places they're not meant to be, spelling mistakes and transcription errors.

FIRST NAME	LAST NAME	DOB
BUILDER	SMITH	1983-08-23

OCCUPATION
BOB

For example, Bob competed a survey for Stats NZ. Without checking, he accidentally entered his first name under occupation and vice versa.

FIRST NAME	LAST NAME	DOB
SARA	JONES	1992-05-02



FIRST NAME	LAST NAME	DOB
5ARA	J0NES	1992-05-02

Another example would be a name that has a number entered in error when transcribing survey results.

No real data is used in examples

Metadata for the IDI and LBD

- Because most admin data is intended for operational use or case management, there is very little metadata that travels with it.
- Ideally, we would like to receive both data dictionaries and encyclopaedic contextual information, but for most datasets the information is outdated or missing.

Changes in data over time

Because admin data is not curated in the same way that, for example, survey data is, it can be hard to managed changes in data over time.

IRD (tax) data was originally formatted as the receipt of paper forms submitted, however, as IRD has moved to capturing electronic transactions the *format* of the data has changed substantially.

While IRD can work through these changes, they have significant impacts for all downstream users of the admin data.

A lack of common data concepts

Different data collections express similar variables the same way

- A variable called “address” might be either the a postal or residential address, or a mix of both.
- A variable called “gender” may actually be “sex”, or vice versa

Different data collections express the same variable different ways

- Some collections have separate fields for first name, middle names and last name.
- Some collections have one field for the whole name
- Date formats are sometimes not even standardised within a single supply

Non-standard variable formats

Jane Abigail Smith was born in New Zealand, 17th April 1982.

The birth record would look something like this:

BIRTH

First Name	Last Name	Sex	Date of Birth
Jane Abigail	Smith	F	17041982

Jane's parents are listed on her birth certificate, but their DoBs will not be digitised. This means it is difficult to link the parents listed here back to *their* original birth records.

To maximise the linking opportunities we would standardise this record by

- Uppercasing all text
- Ordering all names alphabetically
- Standardising sex from "M" and "F" into "1" for male and "2" for female
- Standardising the date format into yyyy-mm-dd

BIRTH

First Name	Last Name	Sex	Date of Birth
ABIGAIL JANE	SMITH	2	1982-04-17

No real data is used in examples

Stable and Non-stable attributes

Almost all attributes about a person can change during their lifetime

- They may change their last name if they marry or enter a civil union
- They may alter their name or go by a nickname in some data collections
- They may change their gender

Even Date of Birth – which ostensibly cannot change, can easily be expressed in a different format, perhaps by mixing up the day and the month. It can also be erroneously reported for migrants or refugees to New Zealand.

Changes in surname

Jane Smith gets married on 30 September 2007 to an American immigrant named Ashley Elliott Jones.

The standardised visa record would look something like this:

VISA

First Name	Last Name	Sex	Date of Birth
ASHLEY ELLIOTT	JONES	1	1980-06-23

Jane Smith decides to change her name to that of her partner's, and starts paying tax under her married name. This means that the tax record is trying to link to a birth record that have different surnames.

BIRTH

First Name	Last Name	Sex	Date of Birth
ABIGAIL JANE	SMITH	2	1982-04-17

TAX

First Name	Last Name	Sex	Date of Birth	Address
ABIGAIL JANE	JONES	2	1982-04-17	123 Jam Street

No real data is used in examples

A lack of shared data definitions

Address is a good example of a variable for which there isn't a common definition. Addresses can be expressed in different ways by the different people at different times

2/43 Toast Road
No. 2, 43 Toast Rd

Changes in address

Jane Smith completes the new HES survey in 2008. She and her husband have just moved house, so the address she gives for HES is different to that on her IRD record.

TAX

First Name	Last Name	Sex	Date of Birth	Address
ABIGAIL JANE	JONES	2	1982-04-17	123 Jam Street, Auckland.

STATS NZ HES

First Name	Last Name	Sex	Date of Birth	Address
ABIGAIL JANE	JONES	2	1982-04-17	456 Nutella Ave, Green Bay, Auckland 0642

The addresses here are in similar formats – most of the addresses received by Integrated Data do not have much consistency. Even small changes in format can make it hard to do address matching.

No real data is used in examples

Errors in Date of Birth

Jane Jones and Ashley Jones have a daughter whom they name Mary-Elizabeth Joy Jones.
Mary-Elizabeth Jones was born 1st February 2009.

The standardised birth record would look something like this:

BIRTH

First Name	Last Name	Sex	Date of Birth
ELIZABETH JOY MARY	JONES	2	2009-02-01

Mary-Elizabeth Jones starts school in 2014. By this point, she only goes by the first name “Mary”.
Her father enrols her, and accidentally formats the date incorrectly.

Now the education record must try and link with the original birth record.

EDUCATION

First Name	Last Name	Sex	Date of Birth
MARY	JONES	2	2009-01-02

No real data is used in examples

A lack of shared data definitions

Admin data may force into certain categories, use non-standard classification or use old classifications that don't map well.

Mary Elizabeth may decide that she would prefer to identify as a non-binary gender.

Not all data collections have an appropriate category for them to select.

There may also be confusion over whether a data collection is asking for sex at birth or gender as chosen.

- How could collection of key linking variables be standardised across admin and survey sources?

- If they cannot be standardised, what techniques could be used to deal with the discrepancies when trying to link?