The IDI prototype spine's creation and coverage

Statistics New Zealand Working Paper No 16–03

Andrew Black

Crown copyright ©
This work is licensed under the Creative Commons Attribution 4.0 International licence. Except for any photographic images, or any content with a specific copyright statement, you are free to copy, distribute, and adapt the work, as long as you attribute the work to Statistics NZ and abide by the other licence terms. To ask permission to reuse a photographic image, send a request to info@stats.govt.nz. Please note you may not use any departmental or governmental emblem, logo, or coat of arms in any way that infringes any provision of the Flags, Emblems, and Names Protection Act 1981, or would infringe any provision if the use occurred within New Zealand. Use the wording 'Statistics New Zealand' in your attribution, not the Statistics NZ logo.

Disclaimer
The Statistics New Zealand Working Paper series is a collection of occasional papers on a variety of statistical topics written by researchers working for Statistics New Zealand. Papers produced in this series represent the views of the authors, and do not imply commitment by Statistics NZ to adopt any findings, methodologies, or recommendations. Any data analysis was carried out under the security and confidentiality provisions of the Statistics Act 1975.

Liability statement
While all care and diligence has been used in processing, analysing, and extracting data and information in this publication, Statistics New Zealand gives no warranty it is error free and will not be liable for any loss or damage suffered by the use directly, or indirectly, of the information in this publication.

Citation
Black, A (2016). *The IDI prototype spine's creation and coverage.* (Statistics New Zealand Working Paper No 16–03). Retrieved from www.stats.govt.nz.

ISBN 978-0-908350-54-4 (online)
ISSN 1179-934X (online)

Published in July 2016 by
Statistics New Zealand
Tatauranga Aotearoa
Wellington, New Zealand

Contact
Statistics New Zealand Information Centre:
info@stats.govt.nz
Phone toll-free 0508 525 525
Phone international +64 4 931 4610

www.stats.govt.nz

Contents

List of tables and figures

Tables by chapter

Figures by chapter

1 Abstract

The Integrated Data Infrastructure (IDI) is a collection of linked datasets that allows evaluation and research on the pathways, transitions, and outcomes of people. There is one specific dataset, called the ‘spine’, to which all of the other datasets link. It is imperative it covers the target population as much as possible. A spine with poor coverage severely limits research use of the IDI.

New data collections are continuously being added to the IDI. This allows Statistics New Zealand to periodically assess the spine, and a chance to decide on the make-up of the spine. Before 2015, the spine was created from only one source – Inland Revenue data.

This paper describes the process to create the new spine, which is termed the IDI prototype spine. We used the target population for the new spine and the spine assessment criteria to decide on the three data sources for the prototype spine: tax, births, and visa data. These were combined to create a singular spine dataset.

We then evaluated the prototype spine by looking at its composition, over- and under-coverage, and making a link-rate comparison with the previous Inland Revenue-only spine.

The results suggest the prototype spine is a significant improvement, particularly for the coverage of an ‘ever-resident New Zealand population’.

Further work is required to work on known areas of over- and under-coverage, and on new data collections available since the prototype spine was introduced, such as health and Census 2013.

2 Background about the IDI and its spine

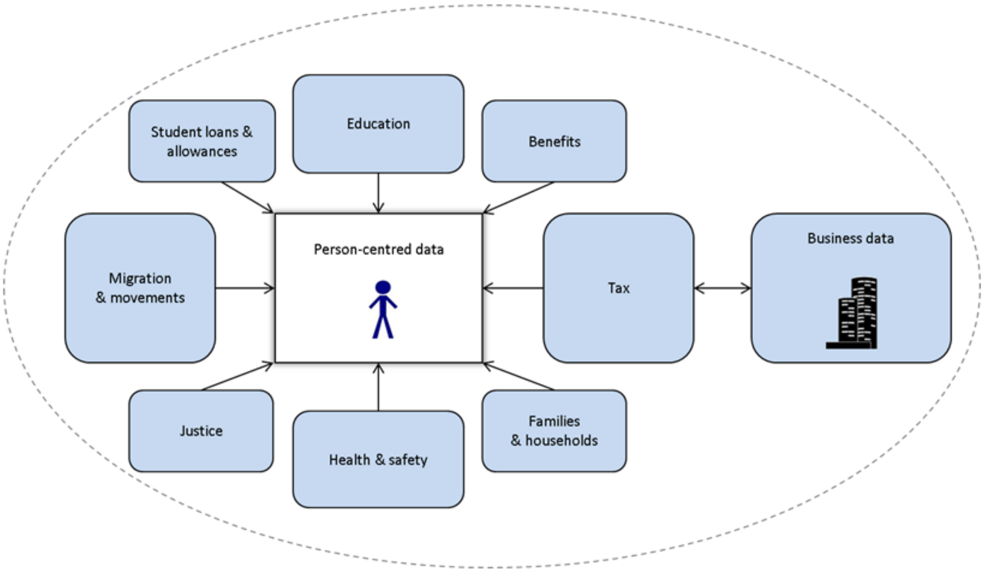
The IDI

Statistics New Zealand's Integrated Data Infrastructure (IDI) is a large longitudinal dataset constructed by linking administrative and survey data sources at the individual (person) level. It covers a wide range of data, including tax, education, health, justice, and benefits.

The IDI can be used for policy evaluation and research, and to produce government and non-government statistical outputs on the pathways, transitions, and outcomes of people.

Figure 1

Integrated Data Infrastructure



See How researchers are using the IDI for information on research projects using IDI data (Statistics NZ, 2015a).

Statistics NZ is required by law to protect the information we collect. Access to the IDI is carefully monitored and controlled by Statistics NZ in order to protect the privacy and security of individuals. Once access is granted, the data available in the IDI is de-identified – personal identifying information such as names, addresses, birth-day [month and year are available], and unique identifiers are removed. All research is checked before it is released to ensure no information that identifies individuals or entities is published or disseminated.

See Microdata access protocols for information about accessing the IDI and the principles against which research projects are assessed (Statistics NZ, nd).

History of the IDI

In 1997, Government directed that "where datasets are integrated across agencies from information collected for unrelated purposes, Statistics NZ should be custodian of these datasets in order to ensure public confidence in the protection of individual records" (Cabinet minutes, 1997).

Since then, Statistics NZ has undertaken a number of projects that integrate datasets supplied by different government agencies, including Linked Employer-Employee Data (LEED) and a Student Loans and Allowances integrated dataset. However, these datasets existed independently, which meant the usability of the data was limited (Statistics NZ, 2012).

In 2011 Statistics NZ began creating a prototype of the IDI to allow efficient linkage of individual and business-level data. The IDI prototype consolidated the existing individual integrated datasets.

This enabled research and statistical outputs on the transitions and outcomes of people through:

- the secondary and tertiary education systems
- the labour market
- the benefit system
- movements in and out of New Zealand.

The IDI prototype increased the flexibility to respond to changes and development in source-agency administrative datasets, and to update statistical processes and outputs. It was enhanced and replaced by the IDI in December 2012.

In 2013 the IDI was extended to include extra data sources, including the justice sector (Statistics NZ, 2013). Ongoing expansion in 2014 and 2015 included a broader range of data, in particular an increase in data from the social sector. For example, extended benefit and education data were added while health, and births, deaths, and marriages data were introduced.

Structure of the IDI

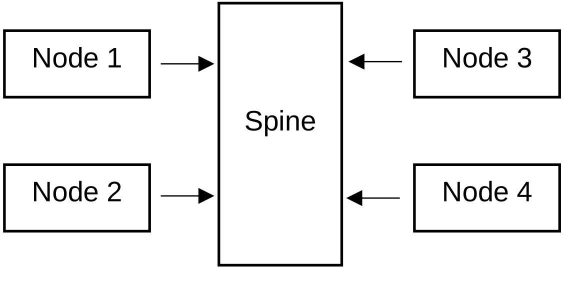
The IDI consists of a central ’spine’ and a series of ’nodes’. The spine is the primary person-level dataset that all other person-level datasets are linked to. These other datasets, called nodes, are the datasets created from different sectors (eg health and education). Each node is linked to the spine, but they are not generally directly linked to each other.

Note: The IDI currently has two instances where nodes are directly linked to each other – education data to student loans data, and education data to migration data.

Figure 2 shows a simplified example of four nodes linking to a central spine.

Figure 2

Simplified spine and node example



At the end of 2014, the IDI spine consisted of anyone who had interacted with Inland Revenue since 1999. This includes anyone who has, since then:

- received income from any source, including a job, benefit, pension, or investments

- joined KiwiSaver
- applied for a student loan or child support
- started a company or gone into business
- filed a tax return
- applied for Working for Families Tax Credits (partners and children will also be included) (Inland Revenue, 2015).

Tax data was used as a spine for two reasons.

- It was one of the first datasets available in the IDI that had good coverage of the New Zealand population.
- Many of the early uses of the IDI focused on employment outcomes and people's future income from particular programmes.

As the IDI expanded, both as a data source and in researcher use, demand to link children in the IDI increased. Because children are characteristically absent from tax data it was necessary to assess whether tax data alone was appropriate for use as the spine. New data sources had also recently become available.

The remainder of this paper describes the process of choosing, creating, and evaluating a new IDI spine.

Structure of linked administrative data sources internationally

International literature often focuses on creating a statistical population register from administrative data – to use administrative data for census purposes. Europe has led the way – there are established population registers and greater public acceptance of using personal data for statistical purposes (Kukutai et al, 2014).

For example, in 2015 the Office for National Statistics (ONS) in the United Kingdom created the first version of a Statistical Population Dataset (SPD), an attempt to create a 'usual resident population' that can be used to create population estimates. This was done primarily by linking the National Health Service Patient Register, Department for Work and Pensions Customer Information System, and data from the Higher Education Statistics Agency. This represents a similar population to that New Zealand would obtain by combining the health, tax, and higher education data.

The ONS applied rules to the linked data to create multiple versions of the SPD. They compared the SPD population estimates (by location, age, and sex for 2011, 2013, and 2014) with official census estimates in 2011 and official mid-year population estimates in 2013 and 2014. The ONS was limited by the data available, so future releases may include new data that has since come available (Office for National Statistics, 2015).

In Ireland, a person activity register (PAR) was created. This summarises each person's annual activity in key public administration systems, including births, benefits, education, and employment. One purpose of the PAR is to enable longitudinal analysis across administrative systems. A key difference from New Zealand is that each Irish citizen has an official personal identification number (PIN) when engaging with the public sector (Dunne, 2015).

The Nordic countries have population registers and a PIN, which makes linking much easier (United Nations Economic Commission for Europe, 2007). Netherlands, which also has a population register and a PIN, achieves a linking rate of almost 100 percent – in most cases using the PIN alone (Bakker et al, 2014).

However, the approach to creating linked administrative data sources must be very particular to the country's situation. It depends on factors that include:

- whether an existing population register is available
- whether a PIN is available
- which datasets are available at the time of creation
- the individual quality of the datasets available.

New Zealand does not have an existing population register or PIN. The datasets available and the quality of those datasets are discussed in chapter 3.

The IDI is primarily designed to be used for microdata research, rather than being a population register. This paper therefore details information about creating a linking population spine using probabilistic linking, rather than using an existing population register or PIN.

What makes a good spine

The spine should be a complete list of uniquely identified members of the target population – it should include every person in the target population once and only once. This enables researchers to derive meaningful subpopulations and creates a robust environment to integrate data from diverse sources.

From a practical perspective, the spine should capture the maximum coverage possible while minimising the number of sources included in the spine. This is because each extra source added to the spine increases the cost and complexity – because it needs to be linked to all existing sources to identify common people between the datasets.

The target population for the IDI spine is broadly an 'ever-resident' population. It includes people born in New Zealand, permanent residents, people with visas that allow them to reside, work, or study in New Zealand (including international students and temporary workers), and those who live and work here without requiring a formal visa (eg Australians). It excludes short-term visitors (eg tourists).

The target population definition attempts to define a spine that includes most research populations of interest. Researchers are expected to subset the spine to find their population of interest, by using variables from the datasets available in the IDI.

As the target population is an ever-resident population, it differs from other international structures that are set up mainly for census purposes – which require resident population at a specific point in time. The ever-resident population is necessary to allow linkages to be made between sources across decades simultaneously. New Zealand's approach is for the census to use the spine as a starting point for determining who is living in New Zealand at a point in time (Gibb et al, 2016).

Because this project was the first time that multiple sources were investigated to create a spine, the new spine was termed the 'prototype spine'. In theory, any data in the IDI could be used to create the prototype spine. Therefore we needed to establish some criteria to assess the different data sources. We used similar criteria as were used to evaluate data sources for population estimates (Statistics NZ, 2011).

Table 1

Assessment criteria for potential prototype spine sources

Criterion	Elaboration
Simplicity	How easy is it to document, explain and understand
High coverage	Coverage of dataset compared to target population
Timeliness	Time lag, periods the data covers
Unique identifiers	Quality and coverage of unique identifiers to help remove duplicates
Variables	Quality and availability of name, sex and date of birth to link to other sources
Consistency	Consistency across period covered, how often resupplied

3 Choosing datasets for the prototype spine

We assessed the tax data that was acting as the spine against the criteria, along with data from the Ministry of Education, births data from the Department of Internal Affairs (DIA), and movements and visa data from the Ministry of Business, Innovation and Employment (MBIE). MBIE's movements' data has each person's border movements (in and out of New Zealand) since 1997, while the visa data is a subset of the movements' data that is limited to people accepted for an entry visa other than a visitor's or transit visa.

At the time of creating the prototype spine, data available from the Ministry of Health was restricted to the working-age population only. Therefore we did not assess this data against the criteria. Since the creation of the prototype spine, the full health population has come available and is being assessed in the same way – as a possible future spine data source.

Assessing the data sources

Table 2 summarises our evaluation of the data sources against the assessment criteria. In some cases, previous use of the datasets or the metadata provided the result. In other cases, we required extra investigation, particularly for the newer datasets and to assess coverage.

Table 2

Evaluation of potential prototype spine sources against criteria

Criteria	Tax	Education	Births	Movements	Visas
Simplicity	Good	Good	Good	Good	Moderate
Coverage	Moderate	Poor	Moderate	Poor	Moderate
Timeliness	Good	Moderate	Good	Good	Good
Unique identifiers	Good	Moderate	Good	Moderate	Good
Variables	Good	Moderate	Good	Good	Good
Consistency	Good	Moderate	Good	Good	Good

The existing spine (tax data) achieved a 'good' rating on all criteria except coverage. Specifically, tax data has low coverage of younger people and new migrants.

Education and births data could both help add children to the prototype spine. However, it is clear from the evaluations that the births data had better coverage of children (especially children under five years) than the education data. As the births data covers a reference period much longer than any other data source, it would also add a population of people who hadn't worked in New Zealand.

Education data was not used as a prototype spine source for a couple of reasons. Firstly, we know there are duplicates in the data source. The same person can have multiple national student numbers (NSNs). NSNs are used in the education sector to identify students.

Secondly, the effective start of the reference period for education data is 2003, when full-name data first became available. While this source would improve the coverage of younger people, it does not cover a long reference period (as the births data does) so would not add older populations of non-workers.

The main problem with the movements' data is over-coverage when compared with the target population, which results in a 'poor' rating in table 2. Over-coverage means there are people in the movements' data who are not in the target population.

Because movements' data includes everyone who has entered or exited New Zealand since 1997, this dataset would add far more visitors and tourists to the prototype spine than the residents we are interested in. However, visa data only includes people who were accepted for visa types we are interested in, including working and resident visas. This fits much more closely to the target population than the whole movements population does.

We concluded that no single data source would be good enough to act as the prototype spine. For example, having a prototype spine made up of only the births data would result in our missing migrants. Therefore we needed to link several data sources to achieve better coverage.

Data sources in the prototype spine

We decided the prototype spine would be based on the tax, births, and visa datasets. Specifically, the visa data is anyone accepted for a visa to enter New Zealand, other than on a visitor's or transit visa.

While the births data includes births back to 1840, the spine only used births from 1920. This was due to reasons such as the reference periods of other data sources available, and research use of the IDI. We decided that adding pre-1920 births was more likely to lead to incorrect links than add research value to the IDI.

The prototype spine is therefore is made up of:

- tax data from 1999
- births data from 1920
- visa data from 1997.

Together, these three data sources conceptually have good coverage of the target population, without having too much over-coverage.

4 Methodology used to create the IDI prototype spine

After choosing three sources for the prototype spine we needed to somehow combine them into a single prototype spine dataset, ready for nodes to link to.

Appending the three datasets together could result in a dataset that covers the whole target population; however, it would not uniquely identify members of the target population. For example, anyone born in New Zealand and who has worked in New Zealand would be in the dataset twice. This becomes a problem when nodes are linked to the prototype spine, because the same person might link to their birth record in some instances and their tax record in other instances. Incomplete information would be available for research.

Data integration is the way to resolve this issue. Specifically, probabilistic linking – in most cases there is not a unique identifier on both datasets that can allow exact linking to occur. Unlike the Nordic countries, New Zealand does not have a PIN that is used across government systems. When linking, occasionally a common unique identifier is available. However, when linking the spine sources together, no unique identifiers were common across the complete datasets.

We used three probabilistic linking projects to create the prototype spine – tax to births, tax to visa, and births to visa.

The process to link each pair of datasets together is much like the process to link any two datasets in the IDI. It involves cleaning and standardising the data before attempting to probabilistically link the datasets in a series of passes.

Name, sex, and date of birth were used in all three linking projects. For the tax and visa data, passport number was also used, as in the tax data a passport number is available for some people. The process to link two datasets in the IDI is described in a report on linking methodology in the IDI (Statistics NZ, 2014).

See Data Integration Manual (Statistics NZ, 2015b) for information on how weights are calculated when deciding whether two records are linked.

Combining links

Once we had linked tax to births, tax to visa, and births to visa in this process, we used these links to create the prototype spine – a spine that aims to be made up of uniquely identified members of the target population.

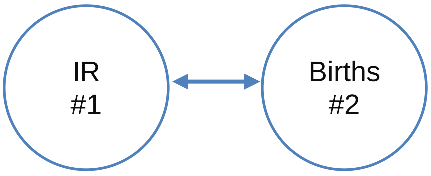
To do this, we used the links we had made between the three prototype spine sources to combine records on the spine. However, some of the links we created could have had contradictions. The general process combines records where there are no contradictions and resolves contradictions with business rules. We describe this through a series of examples.

The simplest example is where two people were linked between two datasets (figure 3a; for example in Inland Revenue tax (IR) and in births.

Figure 3

Examples of combining linking results

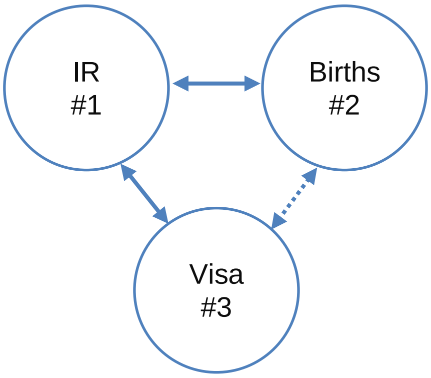
Figure 3a



A spine record is created that combines IR #1 and Births #2 into the same record.

Another possibility is someone from one dataset linking to both the other two sources (figure 3b).

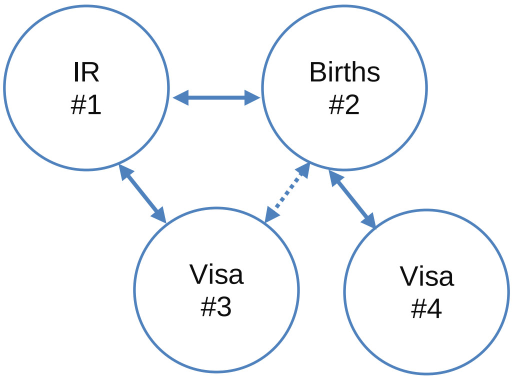
Figure 3b



A spine record is created that combines IR #1, Births #2, and Visa #3 into the same record. Despite not making a link between Births #2 and Visa #3 we have an implied link between these two people (shown by the dotted line in figure 3b) because they both linked to the same IR person.

The most complicated example is if we create links that contradict each other (figure 3c).

Figure 3c



As in figure 3b, this means there is an implied link between Births #2 and Visa #3. However, this contradicts the link we made between Births #2 and Visa #4.

We needed a way of resolving these conflicts. We did this by defining a priority for which links are accepted and which are rejected. The current priority is: IR–visa, IR–births, and lastly births–visa, meaning the births–visa link is the most likely link to be rejected.

To decide the priorities, we considered the quality of the data – so the links accepted were likely to be those of the highest quality. In the example above, it means we reject the link created between Births #2 and Visa #4. A spine record is created that combines IR #1, Births #2, and Visa #3 into the same record.

If someone from any of the three datasets didn’t link to either of the other two, they would still be on the prototype spine but would not be combined with any other records. In figure 3c, this means Visa #4 would be on the spine, but not linked to any IR or birth records.

The prototype spine also needs to have a process to deal with different information coming from each data source. We may link someone’s records together despite their name, sex, or date of birth being slightly different. The spine retains different versions of a person’s name, sex, and date of birth to be available to link to each node.

5 Evaluating the IDI prototype spine

Make-up of the prototype spine

Once the prototype spine is created we gain a view of its make-up, in terms of the number of records coming from each of the three data sources. For example, figure 4 shows the proportion of people in the prototype spine from the IDI refresh finished in August 2015. In total, the prototype spine includes approximately 9.2 million records. Figure 4 is to scale (Micallef and Rodgers, 2014).

Figure 4

Composition of the prototype

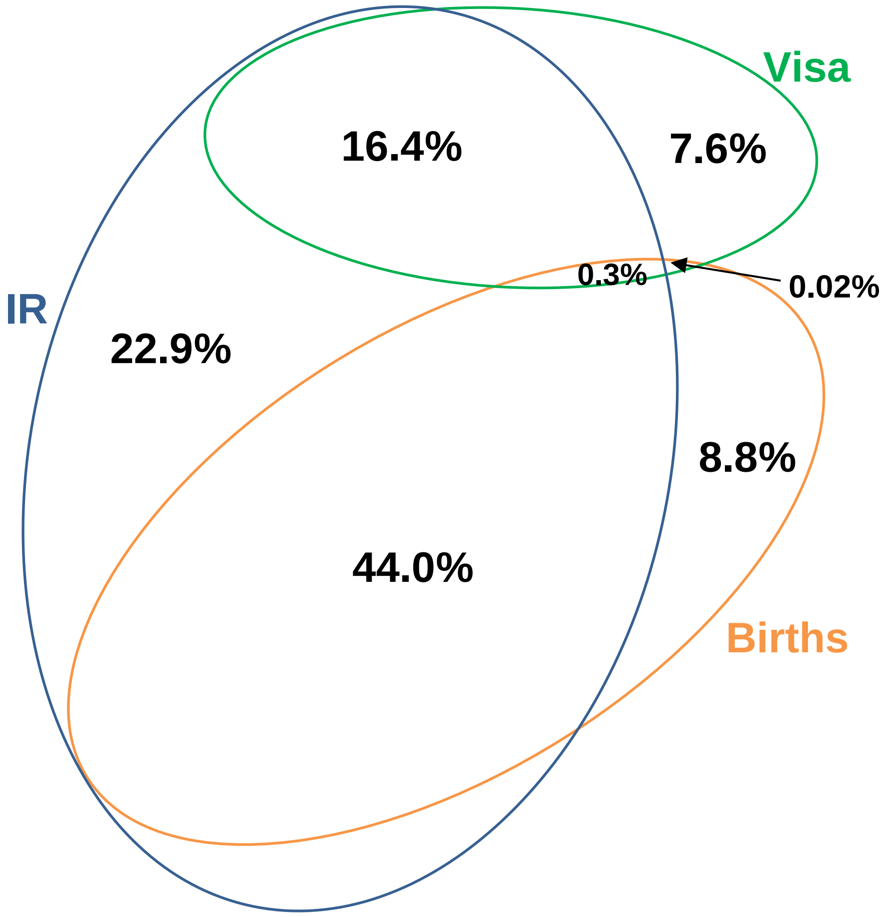


Figure 4 shows the largest overlap for the prototype spine sources is between IR and births (44.0 percent) while visa and births (0.02 percent) has the lowest overlap. Just 0.3 percent of prototype spine records overlap all three sources.

The make-up of the prototype spine for the number of people from each source isn’t surprising, with a large overlap between IR and births – this is the people who were born in New Zealand and later have interaction with the tax system. A large overlap between IR and visa sources isn’t surprising either, because a large number of people get a visa to work in New Zealand.

What may be surprising is the overlap between births and visa – people being born in New Zealand but later requiring a visa to enter the country. However, since 2006, children born in New Zealand only acquire citizenship at birth if one of their parents is a New Zealand citizen or is entitled to reside indefinitely in New Zealand, the Cook Islands, Tokelau, or Niue. This includes Australian citizens or permanent residents – they are able to reside in New Zealand indefinitely (Department of Internal Affairs, nd). This can explain the small overlap between the births and visa data sources. The bulk of the links for people born before 2006 were those entering New Zealand with a returning resident visa.

The areas of the spine made up from only one source are also interesting. There are good reasons why people may appear in one dataset and not another. For example, someone may be born in New Zealand but move out of the country before they interact with the tax system.

However, the 22.9 percent of people who are only in the tax data is interesting. We would expect most people who interact with New Zealand's tax system to have been born here or have been granted a visa to enter.

We investigated this issue with the data available and discovered three points.

- A group is from Australia, Cook Islands, Tokelau, and Niue, and does not require a visa to enter the country. See under-coverage for this group (below).
- 2013 Census data indicates 25.2 percent of New Zealand's population was born overseas, including over 300,000 people who had lived in New Zealand for over 20 years. They won't be in the visa data, which starts in 1997 (Statistics NZ, 2015c). Others who are only in the tax data will have left the country before the 2013 Census.
- A large number did not have any travel records since 1997 and did not appear to have worked in New Zealand since 1999. We were uncertain of the reliability of these records, which warrants further investigation.

In investigating the 7.6 percent recorded in only the visa data, we found the link rates to the tax data for certain visa types (permanent resident, returning resident, resident, and work visas) were higher than for other visa types (student, limited, and diplomatic visas).

Over- and under-coverage in the prototype spine

We can also think about potential causes of over- and under-coverage in the prototype spine when compared with our target population. Over-coverage exists when we include people in the prototype spine who don't meet our target population; under-coverage means we have people who are in our target population but aren't in the prototype spine.

The target population for the IDI prototype spine is broadly an 'ever-resident' population. It includes people born in New Zealand, permanent residents, people with visas that allow them to reside, work, or study in New Zealand (including international students and temporary workers), and those who live and work here without requiring a formal visa (eg Australians). It excludes short-term visitors (eg tourists).

Over-coverage

Over-coverage in the prototype spine may not be directly related to the linking; for example people who:

- worked in New Zealand but were only short-term visitors
- were non-residents but paid tax in New Zealand
- were accepted for a visa but didn't actually come to New Zealand
- were born in New Zealand but left soon after.

Another source of over-coverage in the prototype spine is duplicates. This can happen either because there are duplicates in the individual prototype spine sources or we failed to make a link between prototype spine sources. One reason for failing to make a link would be a person having different information on each source. A common example is a woman having her birth surname in one source and a different married name in the other source, which results in us being unable to link the two sources.

Under-coverage

A number of reasons exist for under-coverage in the prototype spine that are not directly related to the linking; for example:

- migrants who arrived in New Zealand before 1997 and haven't interacted with the tax system since 1999
- non-registered births for people who haven't interacted with the tax system since 1999
- migrants into New Zealand who are not required to have a formal visa and haven't interacted with the tax system since 1999.

A source of under-coverage that is due to linking is links made between the sources that are incorrect. This results in one record representing what should be at least two separate people in the prototype spine.

We investigated the quality of the linking to create the prototype spine for the proportion of links that were incorrect (false positives) and for how many links we didn't make that we should have (false negatives). The process for assessing incorrect links is more developed – the process involves clerical review of a sample of the links made.

Assessing the level of missing links is harder, as attempting to find them requires either linking to alternative datasets or using alternative linking methodologies. For the overall prototype spine, the rate of missed links was estimated at 1 percent to 4 percent while the rate of incorrect links was estimated at under 1 percent.

A closer look at migrants not needing a visa

A particular area of prototype spine under-coverage that we looked at in detail was for migrants into New Zealand who are not required to have a formal visa. People from Australia, Cook Islands, Tokelau, and Niue can enter New Zealand and work without getting a visa. Most of these individuals who we would want on the prototype spine have worked in New Zealand and therefore appear through tax data. However, those without an IRD number will not appear in the prototype spine, even though some have been resident in New Zealand for a period.

We would add over 4.3 million people from Australia, Cook Islands, Tokelau, and Niue to the prototype spine if we added anyone from those countries who had ever entered New Zealand. However, most of these were likely be tourists, so adding them would cause a huge amount of over-coverage in the spine. A linking investigation, with this group added to the spine, found we would lose a large number of existing links to nodes. It was clear that adding this whole group to the prototype spine was not an option worth pursuing.

We investigated adding only people from Australia, Cook Islands, Tokelau, and Niue who had been in New Zealand for longer than a certain period of time. When looking at the link rate by different lengths of stay in New Zealand, there was no obvious time period that did not add significant over-coverage to our ever-resident target population.

The under-coverage analysis for these migrants led to deciding between two options – excluding all additional people from the prototype spine or including those over a certain length of stay.

Because we cannot identify the group that does fit the target population, the level of under-coverage remedied by adding a subset of this group would be more than offset by the over-coverage caused. While Inland Revenue data is a prototype spine source, under-coverage in the Australia, Cook Islands, Tokelau, and Niue population is minimised.

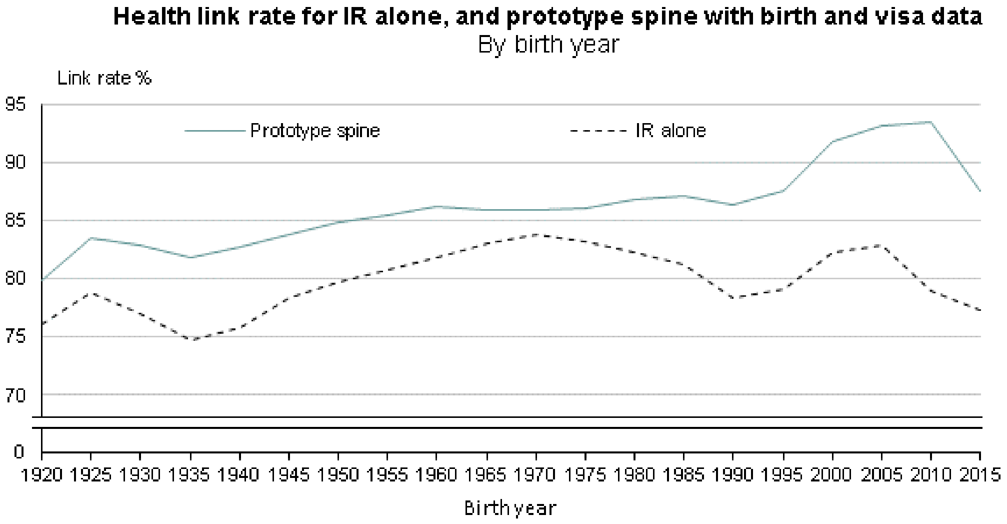
Significant changes to the prototype spine would require reassessing the decision.

Link rate comparison

Another way to assess how well the prototype spine is functioning is to examine how well the nodes can be linked to it. We can compare the link rates for the prototype spine with those for the previous tax (IR) spine.

Figure 5 shows the link rate for health across different age groups, to show the effect of adding birth and visa data to the prototype spine. We see the link rate to health is better for the prototype spine at all age groups than it was to IR alone.

Figure 5



Source: Statistics New Zealand

Comparing the health link rates for IR and the prototype spine, it is clear the prototype spine coverage is much better. It has improved the link rate for all age groups, particularly some older and younger age groups. Younger age groups now have the highest link rates for health.

While this shows the improvement the prototype spine has over using Inland Revenue data alone as the spine, there could be a more optimal spine created from other sources that would be better for the IDI.

Assessment based on spine criteria

In determining the data sources to make up the prototype spine, we assessed potential data sources against six criteria: simplicity, coverage, timeliness, unique identifiers, variables, and consistency. We can use these criteria to assess and compare the prototype spine (based on three sources) and the existing spine (just tax data).

Table 3

Assessment of tax and prototype spine against criteria

Criterion	Assessment	
	Tax spine	**Prototype spine**
Simplicity	Good	Moderate
Coverage	Moderate	Good
Timeliness	Good	Good
Unique identifiers	Good	Good
Variables	Good	Good
Consistency	Good	Good

The prototype spine is more complex than the tax spine because it is made up of three datasets and requires probabilistic data integration to create a single dataset.

The main improvement in the prototype spine is in coverage. Results in figure 5 show significant improvement across all age groups, particularly for children. Despite areas of over- and under-coverage, the prototype spine's coverage (compared with the target population, 'ever-resident New Zealand population') is much closer to meeting the target.

The prototype spine also inherits the good aspects of the three data sources in its unique identifiers, variables, and consistency. For example, the prototype spine can include variables from different sources to link to nodes, including different names and extra variables such as country of birth.

Using the prototype spine

Feedback from researchers using the prototype spine has been positive. For example, it shows significant improvements for census transformation purposes, bringing the constructed administrative population closer to the estimated resident population (Bycroft, 2015).

Next steps in assessing the prototype spine

A next step for continuing development of the prototype spine is to explore alternate spine sources.

At the time we created the prototype spine, available Ministry of Health data was restricted to the working-age population only. Now full data is available we are investigating people in the health data who didn't link to the prototype spine and assessing them against the target population. Health data seems a logical dataset that could be used as a spine source in future.

Further investigation should also take place into the spine's make-up, particularly using the data available to understand why people would be in one dataset but not in either of the other two. This can extend to the people in the nodes who do not link to the spine.

In the IDI refresh that finished in February 2016, we added 2013 Census data to the IDI for the first time. This will allow detailed investigations into how the current prototype spine covers the resident population of New Zealand at a specific point in time.

6 Discussion

The results of this investigation suggest the prototype spine is an improvement over using Inland Revenue data as the single source. Over 15 percent of people in the prototype spine are not from the tax source.

While there are known areas of over- and under-coverage, the population coverage is improved by introducing the multi-source prototype spine. This improvement is largest for children, as was expected, although there was significant improvement for all age groups. Therefore, while conceptually we thought the prototype spine would better cover the target population than the tax data alone, we can conclude it also does so in practice.

Using probabilistic linking to create the prototype spine introduces extra complexity. Over- and under-coverage in the prototype spine can come from the individual sources as well as from the process used to link the three sources together. This means our aim to have each person in the target population in the prototype spine once, and only once, may not be met. The known under-coverage of people from Australia, Cook Islands, Tokelau, and Niue was investigated in detail.

We can also assess the prototype spine based on the criteria we used to select spine sources. While the prototype spine is more complex than one made using only one source, coverage is improved significantly. The prototype spine also allows us to use extra variables in linking to nodes and, particularly with the births data, includes a records source that goes back far beyond most other data sources available.

While the results and feedback suggest the prototype spine is an improvement over using Inland Revenue data as the single source, it is not perfect. Further investigations and having researchers using the prototype spine will discover areas for improvement.

References

Bakker, F, van Rooijen, J, & van Toor, L (2014). The System of social statistical datasets of Statistics Netherlands: An integral approach to the production of register-based social statistics. *Statistical Journal of the IAOS*, 30, 411–424.

Bycroft, C (2015). Census transformation in New Zealand: Using administrative data without a population register. *Statistical Journal of the IAOS,* 31 401–411.

Cabinet meeting minutes (1997). CAB (97) M 31/4. [Electronic copy not available].

Department of Internal Affairs (nd). Changes to citizenship by birth in New Zealand from 2006: Frequently asked questions. Retrieved from www.dia.govt.nz.

Dunne, J (2015). The Irish Statistical System and the emerging Census opportunity. *Statistical Journal of the IAOS* (31) 391–400.

Gibb, S, Bycroft, C, & Matheson-Dunning, N (2016). Identifying the New Zealand resident population in the Integrated Data Infrastructure. Retrieved from www.stats.govt.nz.

Inland Revenue (2015). IRD numbers. Retrieved from www.ird.govt.nz.

Kukutai, T, Thompson, V, & McMillan, R (2015). Whither the census? Continuity and change in census methodologies worldwide, 1985–2014. *Journal of Population Research*, *32*(1), 3–22.

Micallef, L, & Rodgers, P (2014). euler*APE*: Drawing Area-proportional 3-Venn Diagrams Using Ellipses. PLoS ONE 9(7): e101717. doi:10.1371/journal.pone.0101717. http://www.eulerdiagrams.org/eulerAPE

Office for National Statistics (2015). ONS Census Transformation Programme – Administrative Data Research Report: 2015. Retrieved from www.ons.gov.uk.

Statistics NZ (2011). Evaluation of alternative data sources for population estimates. Retrieved from www.stats.govt.nz.

Statistics NZ (2012). Integrated Data Infrastructure and prototype. Retrieved from www.stats.govt.nz.

Statistics NZ (2013). Introduction to the Integrated Data Infrastructure 2013. Retrieved from www.stats.govt.nz.

Statistics NZ (2014). Linking methodology used by Statistics New Zealand in the Integrated Data Infrastructure project. Retrieved from www.stats.govt.nz.

Statistics NZ (2015a). How researchers are using the Integrated Data Infrastructure. Retrieved from www.stats.govt.nz.

Statistics NZ (2015b). Data integration manual: 2nd edition. Retrieved from www.stats.govt.nz.

Statistics NZ (2015c). 2013 Census: Profile and summary reports. Retrieved from www.stats.govt.nz.

Statistics NZ (nd). Microdata access protocols. Retrieved from www.stats.govt.nz.

United Nations Economic Commission for Europe (2007). Register-based statistics in the Nordic countries: Review of best practices with focus on population and social statistics. Retrieved from www.unece.org.

IDI disclaimer

The results in this paper are not official statistics, they have been created for research purposes from the Integrated Data Infrastructure (IDI), managed by Statistics New Zealand.

The opinions, findings, recommendations, and conclusions expressed in this paper are those of the author(s), not Statistics NZ.

Access to the anonymised data used in this study was provided by Statistics NZ in accordance with security and confidentiality provisions of the Statistics Act 1975. Only people authorised by the Statistics Act 1975 are allowed to see data about a particular person, household, business, or organisation, and the results in this paper have been confidentialised to protect these groups from identification.

Careful consideration has been given to the privacy, security, and confidentiality issues associated with using administrative and survey data in the IDI. Further detail can be found in the Privacy impact assessment for the Integrated Data Infrastructure available from www.stats.govt.nz.

The results are based in part on tax data supplied by Inland Revenue to Statistics NZ under the Tax Administration Act 1994. This tax data must be used only for statistical purposes, and no individual information may be published or disclosed in any other form, or provided to Inland Revenue for administrative or regulatory purposes.

Any person who has had access to the unit record data has certified that they have been shown, have read, and have understood section 81 of the Tax Administration Act 1994, which relates to secrecy. Any discussion of data limitations or weaknesses is in the context of using the IDI for statistical purposes, and is not related to the data’s ability to support Inland Revenue’s core operational requirements.