Building on Aotearoa New Zealand's Integrated Data Infrastructure

Dr. Craig Jones, Stats NZ
Anna McDowell, Stats NZ
Vince Galvin, Stats NZ
Building on Aotearoa New Zealand's Integrated Data Infrastructure

Dr Craig Jones¹, Anna McDowell², Vince Galvin³

¹Deputy Government Statistician and Deputy Chief Executive, Stats NZ  
²General Manager, Customer Channels, Stats NZ  
³Chief Methodologist, Stats NZ

Abstract
Stats NZ established the Integrated Data Infrastructure (IDI) to enable researchers to understand population-level trends in people’s experience of government services (e.g. social welfare, education, health and justice). The IDI was established in 2011, building off earlier experiences linking employer and employee data (which itself subsequently formed the basis of another linked database known as the Longitudinal Business Database). Data in the IDI form an ‘ever resident’ New Zealand population of around 9 million people. Access to the IDI is available for bona fide researchers so long as they meet established criteria to show that they will keep the data safe and work with it in a culturally appropriate way. This paper gives a brief history to the development of the IDI (and the LBD) and shares some lessons learned along the way for the benefit of others starting down this integrated data journey.

Introduction
Aotearoa New Zealand has a long history of developing the infrastructure and skills to deliver evidence-informed policy. We have a well-connected and internationally competitive scientific community, world-class academic institutions and we historically ‘punch above our weight’ on many indices that lead to a strong and vibrant analytical community (OECD 2020). We have one of the world’s longest-running longitudinal cohort studies, with the participants of the Dunedin Multidisciplinary Health and Development Study now entering their 45th year. And we have also pioneered methods and cross-government approaches to extract benefit from the masses of administrative data held by government agencies to inform policy decisions.

Stats NZ has been at the forefront of some of these national efforts, particularly the use of integrated administrative data for research purposes. As Aotearoa New Zealand’s national statistical office, we collect and hold a tremendous amount of data relating to our people, economy, environment, and society. Some of these data are used to produce official statistics that help government, the academic community, businesses, our indigenous Māori population, and the broader population to make good decisions. For the last quarter of a century we have also been on a journey to leverage wider benefit from the data underpinning this official statistics system, with a focus on data integration.

As the history below will briefly overview, Stats NZ first began to integrate data in earnest in 1997. There were several advances following this, but our integration capability reached a new level with the development of a prototype integrated data infrastructure (IDI) in 2011. Through a series of iterations, what started as ‘a working prototype’ has now become ‘a prototype working’ and has grown to become a critical part of government operations in Aotearoa New Zealand.

In this article, we will outline some of the origins of Stats NZ's integrated data journey, overview the infrastructure we now have in place, some of the innovative ways this infrastructure is being used to shape policy and inform government operations, and give some lessons we have learned along the way.
Stats NZ’s integrated data programme

Structure
Stats NZ has two integrated databases containing de-identified longitudinal microdata. The IDI contains data about people, while the Longitudinal Business Database (LBD) contains data about businesses.

The IDI is built around a central linking ‘spine’ that all other datasets are linked to. The concept underpinning the spine is to get the largest population coverage with the fewest linking variables. This is important because the spine is essentially re-built whenever the IDI is refreshed (currently three times per year). Introducing more complexity into this initial linking process tends to increase the variability in linking outcomes from one refresh to the next.

The IDI spine is created by linking births, tax records and migration data to create an ‘ever-resident’ population. Once the spine has been created, the remaining data sources are linked to the spine on a one-to-one basis. The remaining data sources are not usually linked with one another to avoid an overly complicated tangle of links that are not required for most analytical purposes. The data sources that are linked to the spine include information about:

- Health (e.g. universal health screening prior to school entry, hospital discharge, immunisation, public health enrolments)
- Social welfare and social services (e.g. receipt of social welfare payments, injury claims, experiences within the child protection system)
- Education and training (e.g. early learning participation, school enrolment and achievement, tertiary education outcomes)
- Social housing receipt
- Income and employment
- Justice (e.g. victimisation, legal prosecutions, sentencing)
- People and communities (e.g. driver licencing, labour force and social surveys conducted by Stats NZ) and
- Population data (e.g. previous Censuses of Population and Dwellings, visa applications, births, deaths and marriages)

While the analysts who create the spine require access to identifying information to create the linkages, all datasets are stripped of information that could be used to identify individual people before being accessed by researchers. Stats NZ has strict protocols in place to check data before being outputted from the IDI to ensure there is no spontaneous recognition of people in statistical outputs. These safeguards are outlined in more detail below.

The Longitudinal Business Frame is a longitudinal representation of the Business Register maintained by Stats NZ. The Business Register is the primary frame used to identify businesses sampled for our

---

1 The resulting linkage contains more than 9 million people, which is far more than the current resident population of around 5.12 million. Many individuals in the IDI spine are former residents of New Zealand who have since left or died. While this over-coverage may sound problematic, it poses few analytical problems because the populations of interest are usually constrained by additional linking variables used in the course of individual projects (e.g. if the analyst is interested in current taxpayers, recipients of social welfare payments, or those involved in the criminal justice system). There are also a range of other causes of over- and under-coverage that are discussed in depth by Gibb et al. (2016).
business activity surveys each year. The Longitudinal Business Frame forms the backbone of the LBD and captures longitudinal information on all economically significant firms operating in Aotearoa New Zealand since 1999.

The information held within the LBD includes location, industry, business type, institutional sector, and parent-subsidiary relationships. Linked to this backbone are a range of additional datasets based around agriculture (NZ’s largest economic sector), international trade, business practices, financials, and information on business innovation (e.g. research and development surveys, and government assistance provided to stimulate innovation). The LBD is currently refreshed annually.

The IDI and the LBD are themselves linked via employee tax information. This enables an extremely rich potential to understand the longitudinal relationship between an individual’s socio-economic circumstances, their journey through employment and work, and the resulting economic activity at the business level.

Data quality
Stats NZ uses two types of linking to create the IDI and the LBD. The IDI spine is created using probabilistic linkage, by linking tax data to births data, births to visa data, and visa to tax data based on identifying variables common to these datasets (date of birth, first and last name, sex and address). Where there are unique identifiers available to link additional datasets into the IDI these unique identifiers are linked directly.

Probabilistic linking means that we match two records based on a probability that they are indeed the same person. Our integrated data are primarily used for research purposes and are not used to make decisions about individual people, which affords us a higher tolerance for imperfect matching. In the March 2021 IDI refresh, linkage rates with central government agencies ranged from around 80 per cent up to a high of 98.3 per cent. Linkage rates tend to be much higher where data quality is essential for the operation of the government services upstream from the IDI (e.g. the 98.3% linkage rate was for driver licencing and motor vehicle registration data). We use a reasonably conservative matching process and false positive rates (where two people are incorrectly identified as being the same person) are lower than 2% for almost every dataset linked into the IDI.

Another way of assessing the quality of the data within the IDI is to construct an estimated resident population from administrative IDI records (an 'IDI-ERP') and compare this to the estimated resident population derived from the five-yearly Census of Population and Dwellings where we attempt to enumerate every person in the country through household survey collection. The methods for selecting the IDI-ERP have been iteratively improved, with the most recent adjustments.
implemented in the use of the IDI-ERP in the 2018 Census of Population and Dwellings. Comparisons against an interim 2018 benchmark population showed that the IDI-ERP distributions tracked closely to the benchmark for distributions by age and sex nationally and for ethnic groups and territorial authorities (Stats 2019b). While the analysis is yet to be finalised, initial estimates suggest that the total IDI-ERP is likely to be around 1% lower than the official ERP in 2018.

Safe and ethical use of integrated data

The success of our integrated data programme is entirely dependent on the safe and ethical use of the data it contains. Stats NZ has two frameworks to guide decision making in this area. The first is known as the Five Safes Framework. All applicants and their proposed research must meet the ‘five safes’ conditions before access to the IDI is granted:

- **Safe people** – researchers are vetted and must commit to use data safely before they can access the data. They must pass referee checks, attend confidentiality training, sign a contract where they agree to follow our rules and protocols, have capability to use the data, and sign a declaration of secrecy under the Statistics Act 1975. The declaration is a lifetime agreement to keep data confidential. Researchers who break our protocols can be banned, blacklisted, or prosecuted.

- **Safe projects** – to gain access to integrated data, researchers must have a project they can demonstrate is in the public interest. Research projects must focus on finding insights and solutions to issues that are likely to have a wide public benefit. The IDI and LBD cannot be used for individual case management, such as making decisions about a specific person or family.

- **Safe settings** – a range of privacy and security arrangements are in place to keep data physically safe. Data can only be accessed through a secure virtual environment known as a Data Lab, and only in research facilities approved by Stats NZ. A variety of additional security layers also protect the information, including locating the IDI and LBD on a separate server that is not connected to the Internet, computers are not connected to a network, and there are no USB ports or printing facilities so users cannot take information in or out of the Data Lab without it being checked first by Stats NZ staff.

- **Safe data** – people’s identity is protected. Data held within our integrated data environment has had identifying information removed, and researchers only get access to the data they need. For example, a researcher granted access to the IDI will only have access to the specific datasets they need for their research project; they cannot see all information in the IDI. Data that is available to researchers is de-identified. Numbers that can be used to identify people are encrypted.

- **Safe output** – all information is checked to ensure it does not contain any identifying results. Researchers must confidentialise their results. We then check all outputs before they can be released from the Data Lab, to ensure information is grouped in a way that makes it impossible to identify individuals. Results that could potentially identify individuals are not released under any circumstances. We provide guidance to give researchers the methods and rules they need to confidentialise their results.

The Five Safes Framework has been extremely effective in keeping data safe from privacy breaches, and for ensuring Stats NZ maintains the social license to continue operating the programme. However, we are also acutely aware that people’s expectations around how their data are used are changing. High profile cases of unethical use of personal information such as the Cambridge Analytica scandal have honed people’s attention to ethical use of their information, not just the need to keep it private.
In Aotearoa New Zealand we have an additional layer of consideration – to ensure our use of data meets the expectations arising from the agreement the Government has with the indigenous Māori people, or tangata whenua (people of the land). For many Māori, data are a taonga (an item of significant cultural value which needs to be treasured). Objects that are taonga are afforded specific protections under Te Tiriti o Waitangi, the Treaty signed by the Queen of England and most Māori Rangatira (tribal chiefs) in 1840. Te Tiriti is the founding document of modern Aotearoa New Zealand and sets out certain expectations of the Government, along with rights and protections for Māori.

As with other parts of the world who have been impacted by colonisation, there is a growing movement toward data sovereignty in Aotearoa New Zealand. Data sovereignty usually refers to an understanding that data is subject to the laws of the place where it is collected and stored. In our context, it is also increasingly taken to mean that data about Māori is subject to Māori governance.

Working with Associate Professor, Maui Hudson (University of Waikato) Stats NZ has developed a data access framework to move us closer to realising this ambition. Ngā Tikanga Paihere was developed in 2018 with the intention to guide appropriate use of microdata in the IDI, focusing on how data about Māori and other under-represented populations is used for research purposes. The framework draws on 10 tikanga (Māori world concepts), which align with Five Safes Framework but guide researchers to understand the cultural competencies and practices they will need to consider in conducting research involving priority populations. Ngā Tikanga Paihere encourages researchers to consider the people behind the data, and the experiences and context for them. Researchers are required to plan more around the way they engage with relevant community groups, the advice they seek in ensuring their research is culturally sensitive, as well as ensuring that important and useful findings from the research reach the relevant community groups in ways that make sense to the community of interest. The tikanga are not described in detail here but examples include:

- Pūkenga | skills and expertise – this creates an expectation that researchers can demonstrate their intention to work in culturally appropriate ways, and that they will establish relationships with relevant communities before undertaking research
- ‘Tapu’ and ‘noa’ | ‘sacred’ and ‘ordinary’ – this tikanga recognises that some data are of such sensitivity that risks need to be carefully managed when using those data, while other data may be much more readily accessible. These forces need to be carefully considered and weighed up by researchers in their use of data relating to Māori and other populations of interest.

Stats NZ actively uses Ngā Tikanga Paihere to assess whether research projects have the appropriate cultural safeguards in place to conduct research in a way that will be beneficial to Māori and other priority populations (e.g. Pacific Island communities, or people with disabilities). Response to the introduction of Ngā Tikanga Paihere has been generally positive. It has highlighted for some research groups the need to grow their cultural capability as well as expanding the types of engagement and involvement that community groups have within the research undertaken in the IDI.

Some research projects require adding new data into the IDI for linking to the broader database. For Stats NZ to consider adding data to the IDI, it must also go through our Data Ethics and Privacy Assessment (DEPA). The DEPA is based on the Ngā Tikanga Paihere framework and the Privacy Act 2020 and is important for every data integration project. It helps Stats NZ to assess and mitigate impacts on privacy, ensure that data is treated in responsible and culturally appropriate ways, assess if a proposed data integration is consistent with relevant legislation, and consistent with Stats NZ’s data integration principles. The DEPA form asks for information on the source of the data and collection method(s), the original intended use, wider benefits of the data, legal and privacy obligations, sensitive variables, cultural responsiveness, and relationship with clients.
History of the IDI

At one level, Stats NZ has been on a data integration journey since annual statistical reports were first published by the Government Printer in the 1850s. The advent of modern computing has, of course, enabled this at an ever-increasing scale. It’s hard to imagine that in the early 1980s, the Department of Statistics was the only Government Department that had its own mainframe. Each floor would have 3-5 terminals with analysts allowed 30-minute segments testing their programme before putting it in the batch system to run overnight. Analysts could sort and merge files with up to 300,000 records overnight.

With the widespread adoption of microprocessor computing in the 1990s came an enormous growth in digital production and storage of data. However, while this increased capability and capacity to link data, it was still incredibly difficult to do so. Any sort of data-heavy processing involving manual steps and very long wait times to process files. The current refresh of the IDI and LBD results in around seven billion rows of data, requiring about 650GB of storage. By modern standards this is not a large amount of data, but certainly well beyond the compute capabilities in the late 1990s. While there was some interest at a policy level in how data integration could help to gain system-wide insights, the tools were not necessarily at our disposal to make this happen.

There was a very acute focus on cross-agency data integrated in the mid-late 1990s in Aotearoa New Zealand, culminating in a Cabinet directive stating that Stats NZ was the right place to carry out this work. Stats NZ is not an operational agency and the independence of the Government Statistician provided the neutrality required to undertake data integration safely. Some small-scale data integration happened in the 1990s and into 2000 but it was the introduction of the Linked Employer–Employee Data (LEED) project in 2005 that really marked the beginning point for our current integrated data programme.

To establish the LEED data, New Zealand looked to international experience (linked employer-employee datasets were already established in parts of Europe and North America). This culminated in the New Zealand Conference on Database Integration and Linked Employer-Employee Data held in 2002. The conference successfully answered three main questions for the New Zealand context

• The integrated data could be built in New Zealand, while maintaining confidence in the linkages and data quality
• The value added from this kind of approach was potentially enormous, with some key issues about the economy likely unable to be addressed without integrated data
• With a careful and deliberate approach, confidentiality and privacy issues can be addressed.

Learning from the international experience provided New Zealand with a solid foundation to take our integrated data programme forward.²

The LEED project in New Zealand, involved linking employee tax data with a database of New Zealand businesses, and an integrated dataset on student loans and allowances. These early integrations were for specific purposes, often linking only two or three datasets, with each kept in a separate environment for security purposes. Linking student information to labour market and business data enabled analysts to quantify the impact of government funding for tertiary education on labour market participation and broader economic outcomes.

In the ensuing years a range of other bespoke linkages were undertaken, requiring an increase in linkage capability and capacity inside Stats NZ. The IDI prototype was established in 2011, which consolidated the previously separate integration projects. The early prototype included economic,

² Lane and Maloney (2002).
education, social welfare, migration, and business data. The IDI prototype increased the flexibility to respond to changes and development in source agency administrative datasets, and to update statistical processes and outputs.

This prototype coincided with a significant push at the political level toward evidence-informed policy development and delivery. The Government at that time had recently been elected for a second term. In 2012 they launched the Better Public Services programme, setting targets for the public sector to achieve in order to deliver improved results for New Zealanders. The Better Public Services advisory group encouraged an increase in cross-agency co-operation, particularly on complex issues that span agency boundaries.

The then Minister of Finance and future Prime Minister, Hon. Bill English, had a particularly strong focus on using information to better understand the needs of vulnerable populations, adjusting services to address those needs, and using quantitative insights to measure the effectiveness of those services. These insights were then used to inform the next set of investment decisions. This policy approach became formally known as the Social Investment through the term of that Government. This approach and the report on evidence-based policy to deliver Better Public Services required improved capability across government to share and use existing datasets.

Funding to further expand the IDI was agreed to in 2013, largely due to the Better Public Services programme and the growth in the Social Investment model across different agencies. To that point, much of the focus for Social Investment had been on the social services sector (e.g. Social Investment Unit, 2017). The early driver was to identify and evaluate policies that reduced people’s long-term welfare dependency, which the Government of the time saw as key to improving the broader set of social outcomes for this group of vulnerable people. The Government saw this in mainly fiscal terms – investing to reduce the liability associated with the benefit system and grow the economy.

The broad intent of the Social Investment approach was continued with the change in Government in 2017. A new ministerial portfolio was created to provide evidence-based information, tools, guidance and products to social agencies, with a stronger emphasis on social wellbeing as opposed to purely fiscal drivers. The IDI has become no less relevant in this context and Stats NZ continues to develop our capability to deliver the information to inform this policy direction.

Around this time, Stats NZ was also formally mandated as the functional leader of the government data system. The Public Service Commission, who is the employer of public service Chief Executives, nominated the Government Statistician as the Government Chief Data Steward (GCDS). This role has enabled Stats NZ to work in partnership with government agencies to set the strategic direction for government’s management of data and unlock the value of data for all New Zealanders. The GCDS role also leads the public sector’s response to new and emerging data issues. This role was strengthened in September 2018 through a Cabinet mandate empowering the GCDS to set mandatory standards and guidelines for the collection, management and use of data by government agencies and direct agencies to adopt common data capabilities (e.g. data tools, linking infrastructure, or other sharing platforms).

This mandate has the potential to re-shape the way agencies manage and share data to further amplify the integration capabilities that have already been established.

Value of integrated data

To date, our integrated data programme has focussed mainly on its use as a research tool to understand complex social and economic issues in more depth, to inform policy, help with the targeting of resources, and for impact evaluation. There are also emerging use cases, where the
statistical outputs from the IDI to directly allocate resources, and to respond to emergency events. As such, it has become an integral component in the decision-making processes of central government.

There are countless examples of innovative uses of integrated data but three recent illustrations are representative of recent trends in use:

- **COVID-19 Complex Contagion Model.** At the beginning of the COVID-19 pandemic, researchers from the University of Auckland undertook a research project using the IDI where they built an individual-based network that is representative of the NZ population, including schools, workplaces, family structure, community groups, movements between and within cities, etc. This network model was used as the basis to run detailed complex contagion models including spatial and occupational effects for the COVID pandemic, under different scenarios. This project was regularly submitting outputs from the Data Lab directly to the National Crisis Management Centre, and the models garnered a high level of media interest with one of the project leads, Prof. Shaun Hendy, regularly speaking to media about the models findings. The detailed population data in the IDI was key to this project as it enabled the researchers to parameterise the synthetic networks, combined with what they knew about the topological properties of real-world networks.

- **Modelling social outcomes.** The Ministry of Social Development uses the IDI to build and develop the Social Outcomes Model. This is a detailed model that gives a broad view of what happens to people today, and what might happen to people in the future, which helps MSD understand current needs and what their needs may be in the future so they can support them with better services. This model looks at income, housing, health, and wellbeing. Output from this project have been feeding into policy development for MSD that ensures the right support is reaching the right people and funding is distributed effectively across support services. The IDI is the only tool that makes this project possible, as it requires linked data for a broad range of factors that covers the whole population, and the IDI is the only place that this is available.

- **Using integrated data to better target school-based equity funding.** The negative effects of socio-economic disadvantage on educational achievement are well documented. Most western countries, including New Zealand, fund schools with higher concentrations of disadvantaged families at a higher rate to mitigate the impact of this disadvantage. In New Zealand, this has been done by looking at the socio-economic characteristics of the neighbourhoods where children live, aggregating this to the school level, and then ranking schools according to their decile of disadvantage. Different funding rates are applied depending on the schools’ decile ranking. This method works reasonably well but there are some distortions within the system. For example, because there is a high degree of school choice in New Zealand, schools that attract the wealthiest children from the poorest neighbourhoods have artificially inflated disadvantage scores. To improve this system, the Ministry of Education developed a student-focussed statistical model based on IDI data to identify relationships between individual measures of disadvantage and a measure of educational success. The model found that previous decile-based calculations were a blunt instrument for identifying the level of disadvantage in a school. There are very large variations in the proportion of disadvantaged students within a given school decile ranking. The student-based measure was found to be a much more accurate way of targeting funding based on the actual level of disadvantage within a given school. The New Zealand Government invested an initial tranche of funding to transition to this new funding model in the 2021 national Budget.
These are but some of the many current uses of the IDI, which continue to grow year-on-year. There are currently 350 active projects in the IDI, with 815 researchers attached to these projects. Since current record-keeping began in 2016, 1,770 researchers have gone through the confidentiality training required before access to the Data Lab is approved. The ‘working prototype’ really has become ‘the prototype working’.

Lessons learned
While we view our integrated data programme as an overwhelming success, we have learned a lot of lessons along the way. We offer seven observations for others embarking on this journey.

1. **Start with a use case and let it grow**

Many projects in government and elsewhere start as good ideas but fail to deliver the intended value. There are myriad reasons for this, but they are often poorly planned, can be captured by scope creep, and become the solution to all organisational problems. We started with a simple use case, to understand simple relationships between education and training, and labour market outcomes. Not only did this provide immediate benefit and constrain the scope, it also enabled supplier agencies to build confidence in the linking process by trying things first. As benefits were achieved and none of their worst fears were realised, supplier agencies became more comfortable with putting their data in the IDI.

The other important aspect to this was letting the researchers identify where value could be found, rather than presupposing what data people might want to include. Early use cases in the Health sector gave their researchers the opportunity to test processes at their end to understand what they were signing up for. Once agencies got their own researchers using the data, the feedback loops closed more readily and the comfort levels picked up as well. And so it grew.

2. **Think about future uses up front and design the infrastructure accordingly**

Notwithstanding lesson #1, allowing the programme to grow organically from a series of use cases can create problems downstream. In our own case, we now have a highly utilised system that is still designed as a prototype. Many of the processes have been added to as the programme has evolved and we spend a lot of time cleaning and processing data between each refresh. The infrastructure investment has not kept pace with the growth in utilisation and a major reinvestment in the underpinning process and technology is required. Three key design principles worth considering are: (a) ensure your design is scalable, (b) work with a modular approach where different component parts can be slotted in and out of your processing pipeline as required and (c) look to partner where you can. The last principle is important to ensure the user community has some ownership in the success of the product as well.

3. **A clear authorising environment matters, and a ministerial champion makes it even better**

The political support from the Minister of Finance, Hon. Bill English, helped some government chief executives to be less risk averse about putting their data in the IDI. Without this support we certainly would not have progressed as fast as we did, and we may not have reached our current state of maturity at all. Having a ministerial champion gave agencies an understanding of why cross-agency linkage was important, and it also helped them to form a plan for how and when they could make it happen. Ministers were putting pressure on agencies to go further and faster, which helped to build momentum. The other key driver was the Better Public Services policy and the associated drive on Analysis for Outcomes. This gave us the funding to grow the IDI and at the same time the New
Zealand Treasury were tasked with helping with the analytical capability across government. Having both of these things happening at the same time was very helpful for our work with supplier agencies.

4. **Safe integration requires a strong regulatory environment**

Legislation is a critical enabler for successful data integration. The Statistics Act 1975 and the Privacy Act 1993 worked in combination to enable Stats to develop and use integrated data without the need for further legislation. The Statistics Act 1975 enables anonymised data to be used for bona fide research if the Government Statistician is satisfied that the research has the necessary knowledge, experience and skills to use that data. Likewise, the Privacy Act (which was updated in 2020) enables the use of information within a framework that protects New Zealander’s privacy.

Stats NZ has recently implemented policy work to modernise our legislative framework. There was little or no digitisation of personal information when the 1975 legislation was written and there is not a single reference to the word ‘data’ contained in the Act. The proposed legislation would modernise the language used and address some key inefficiencies and barriers to effective integration. One key proposal is to enable the Government Statistician to require the collection of information from other agencies for statistical purposes, even if that information is not required for the operations of the agency itself. This proposal recognises the increasing direction successive governments are taking to develop more system-focussed governance.

5. **Be transparent and consistent in how the rules are applied and be prepared to adapt**

The success of any integrated data programme will rest and fall on trust. Data suppliers need to trust that data will not be used in a way that could compromise their operations or in contravention of their legislation. The integrated data team need to trust that researchers are using data in accordance with the access agreements in place. Members of the public need to trust that their data is used safely and ethically by government.

To maintain this trust, we have a very clear set of rules that govern access to and use of the IDI and we are careful to hold all users to the same standard. However, these were not in place from the outset. Processes have evolved as the uses have changed and as the trust relationships with researchers have evolved. We were prepared to review our systems and processes, which are often the touch points that can be annoying for customers. For example, we have made changes to the output checking process, as well as the project application process to streamline and strengthen them over time.

6. **Data quality is a shared responsibility**

Data quality is an ongoing issue for us, but it has improved over time. In our experience, the best way to get agencies interested in improving their data quality is to stimulate them into becoming users of their own data. That said, we still spend most of the time between refresh cycles going back and forth with agencies to clean data in preparation for integration. Expectations around data quality have matured over time but upstream accountability for data quality should be made clear with data suppliers up front. System assets need system accountability. A key pain point will be metadata (data that describes the data in the IDI). This is a key issue for researchers who use the IDI and cannot be understated. This will probably never be perfect, but it does need investment and focus from data suppliers and the integrating entity.
7. Maintain an active network of analysts

Researchers who came into the IDI environment who deeply understood the vagaries of data from their own field (e.g. health) quickly realised that data from other fields (e.g. labour market) were just as idiosyncratic. They realised that they needed to share knowledge across disciplines and those who collaborated more tended to produce the highest quality research.

Stats NZ has worked to create a user community around the IDI. The Data Lab Community Group, which is comprised largely of technical leads, is facilitated by Stats NZ but chaired by an external member. There is an active mailing list, user forums, and an online community where researchers can discuss their research and share code. Research using the IDI is published to ensure that other researchers can see what has been done in the past and share their experiences with the broader IDI community. Stats NZ also helped to establish a Government Analytics Network (GAN) and supported a wider community of interest known as the NZ Data Analytics Forum (which involves private sector interests as well). The GAN has been most effective when there are concrete cross-agency problems to solve, often involving linking data.

Networks have also sprung up within fields; the Virtual Health Information Network being a good example of an active and really useful self-generated user group.

References


