Statistics New Zealand
TATAURANGA AOTEAROA

Statistics New Zealand's Experience of Data Integration:

A Policy and Legislative Perspective

Authors,

Andrew Hunter

Vince Galvin

Eleisha Jewell

Nairn MacGibbon

newzealand.govt.nz

Contents

1 Introduction and Summary

Statistics New Zealand has used administrative and survey data in a range of data integration projects over the last 10 years. This work has been influenced by and interacted with New Zealand's policy and legislative environment in a number of ways. These interactions have included the challenge of designing integration and privacy-related protocols in a rapidly evolving user and provider environment at times before their consequences can be fully assessed. Tensions between privacy and archival legislation have needed to be managed as well as ensuring that the diverse requirements of the legislation under which individual data sources are collected are honoured in the arrangements put in place in Statistics New Zealand.

This paper outlines some of Statistic's New Zealand's more significant experiences in creating output databases by integrating multi source data, and reflects on the policy and legislative aspects of this work. These dimensions are discussed under five headings;

- Policy environment
- Legislative environment
- Integration work done in Statistics NZ
- Lessons learnt
- Future challenges.

The paper concludes that data integration will be an important part of Statistics New Zealand's future work programme. This will require a good understanding of the environment within which integration exercises will take place and active management of the risks and associated issues. The demand for detailed, comprehensive information continues to increase and the question of the extent to which administrative data can meet these needs should be answered before alternatives are considered.

2 Policy Environment

Since the early 2000s, successive New Zealand governments have emphasised the need for the public service to take a 'Whole of Government Approach.' One of the stated aims of this approach is that agencies will increasingly integrate their information holdings and services with those of other government organisations, so that people do not have to approach several agencies in turn to get the services they need. The goal is to improve service delivery, people's experience of government, and enable administrative savings and reduced compliance costs.

Government policy development and evaluation is becoming increasingly targeted and complex. Traditional aggregate statistics compiled and disseminated by National Statistics Offices (NSOs) are not in themselves sufficient to inform users about what is happening to individual agents in society (be they individuals, households or businesses). Increasingly, policy agencies and researchers are seeking access to the microdata that underlie aggregate statistics to draw out additional insights regarding these impacts.

The choice of policy levers to achieve policy goals ideally requires knowledge about how communities of interest function as a system. The link between policy interventions, adjustment costs and the benefits of intervention will be complex. The sort of information infrastructure needed to make these assessments will give decision makers the capacity to trace links between interventions and impacts i.e. cause and effect.

These changes are driving and enabling a broader role for NSOs in undertaking or facilitating microdata based work.

3 Legislative Environment

The legislative environment in any country defines the 'feasible space' within which data integration exercises are able to take place. There are a number of relevant pieces of legislation in New Zealand that help to define this 'space'. These include:

Statistics Act (1975)

The Statistics Act (1975) provides the legislative framework under which Statistics New Zealand operates. It defines the 'feasible space' within which Statistics New Zealand is able to operate with regard to data integration, and associated access to the resulting datasets. One of the key limitations of the existing Act is the constraint it imposes on the ability of non-governmental researchers to access the detailed microdata in Statistics New Zealand's Data Laboratory. Proposed amendments to the Statistics Act (1975) are currently being consulted upon that will facilitate access by non-governmental researchers to this microdata. These changes are intended to enable equitable access to bonafide users.

Privacy Act (1993)

The Privacy Act (1993) has as one of its main purposes the promotion and protection of individual privacy in general accordance with the 1980 Organisation for Economic Co-operation and Development (OECD) Guidelines on the Protection of Privacy and Trans-border Flows of

Personal Data. With few exceptions it applies across the public and private sectors. The Act is primarily concerned with assuring good personal information handling practices are applied.

The Act contains twelve information privacy principles dealing with collecting, holding, use, and disclosure of personal information and assigning unique identifiers. The principles also give individuals the right to access personal information and to request correction of it.

They do not override other laws which govern the collection, use, or disclosure of personal information.

Tax Administration Act (1994)

The Tax Administration Act (1994) is mentioned in this context as an example of a piece of legislation that governs the operation of another government department in New Zealand (Inland Revenue) and contains explicit provision for the sharing of both business and personal taxpayer information with Statistics New Zealand (in s81.4 [d]).

Public Records Act (2005)

The Public Records Act (2005) sets a framework for recordkeeping in public offices and local authorities. It requires public agencies to create and maintain adequate records of their business. The definition of a public record is wide, and encompasses data and datasets. The focus of the Public Records Act is on the creation and retention of full and accurate records of the affairs of government. By contrast the Statistics Act 1975, which governs the collection and dissemination of official statistics, emphasises data security and confidentiality and focuses on ensuring that respondents cannot be identified through published statistical data. In regards to integrated data, this means that we will only integrate data for a clear purpose and once that purpose has been met there is no further need to retain the data.

4 Integration work done in Statistics New Zealand

Against the backdrop of this environment Statistics New Zealand has looked for opportunities to integrate data from different sources to meet specific information needs. The main person-level integration exercises Statistics New Zealand has undertaken to date are:

Inter-censal Ethnic Mobility Study 2001 - 2006

This study links individual records from the 2001 and 2006 Censuses of Population and Dwellings so that individual responses made at the two censuses can be compared. Individual records for the study population are linked probabilistically using date of birth, sex, and geographic location

variables. The purpose of the study is to produce statistical estimates of the inter-ethnic mobility between the 2001 and 2006 Censuses, as well as to update parameters used to produce periodic ethnic population projections.

Student Loans Project

This study links individual records from a variety of data sources including Ministry of Education (MoE) data on study; Ministry of Social Development (MSD) data on student loans and allowances; and Inland Revenue (IR) data on student loans and income. The first stage of the integration involved matching the MSD and IR data using tax file numbers. The second stage involved adding MoE education records through probabilistic matching, using student identification number, tertiary institution identifier, and demographic variables. The purpose of the study is to assist the understanding of the costs of the student loans scheme to the Crown, forecasting, reporting of the asset in the Crown Accounts, costing policy changes and assessing the socio-economic impact of, and the return on, education and outcomes.

New Zealand Census Mortality Study

This study anonymously and probabilistically linked each of the four censuses from 1981, 1986, 1991, and 1996 with three years of subsequent mortality data, to create cohort datasets as the basis for ongoing analysis. Linkage to census data via probabilistic matching on key variables (area of residence, date of birth, sex, and ethnicity) was necessary because detailed socio-economic data is included on census records but not on mortality records. The purpose of the study is to measure the association between socio-economic factors and mortality, and to monitor changes over time in socio-economic mortality gradients.

Matching 2001 vitals to Census 2001

This study probabilistically matched birth and death registration records from 2001 with the Census 2001 records using sex, date of birth and usually resident geographic location as matching variables. The purpose of the study was to assess the level of consistency in the data to provide insight into the accuracy of ethnic demographic projections and population estimates.

Cancer Trends

This study matched cancer registration data from the New Zealand Health Information Service (NZHIS) to the 1981, 1986, 1991, 1996, 2001 Census data using probabilistic methods. The purpose of the study is to produce robust cancer statistics that can be used to examine trends in cancer incidence (and survival) in New Zealand by ethnic group and socio-economic status.

Linked Employer - Employee Database (LEED)

This project integrates company level tax data, with data on payments made to employees and information about individual tax payers to create histories of people moving between firms, improved estimates (via personal address information) of the regional distribution of employment and histories of job creation and destruction in individual firms. Developing an overview of the operation of the tax system was essential to understanding what consistency checks are possible. This work has required transforming transaction level information into histories of both businesses and individuals and has succeeded in providing new measures which have illuminated the dynamics of the labour market.

Injury Statistics Manager

Information from New Zealand's accident insurer (known as the ACC) has been merged with hospital discharge data to construct records of the treatment of injuries. The end result has been a database that has a more complete coverage of injuries and more information about the total costs of treatment, rehabilitation, and compensation for these injuries.

5 Lessons learnt

Developing an organisational strategy for integrating data

Statistics NZ has been seeking to realise statistical value from integrated data for many years. The organisation has followed a conservative, but consistent strategy to steadily develop and increase the use of integrated data to meet information needs.

This work began at a time when there was an emphasis in New Zealand on separate agency accountability under the State Sector Act (1998). Over the last decade while this work has progressed there has been a rediscovery of the value of public sector collaboration that has been supportive of cross agency data integration initiatives. Early initiatives faced considerable wariness from other agencies that saw little direct benefit to them from this work, but potentially considerable risk. This was compounded by awareness of events in Canada where a data integration exercise had to be abandoned after adverse comment by their privacy commissioner.

There was a general appreciation that data integration was potentially 'high gain' but also 'high risk'. To counter this risk a deliberate policy of openness and high engagement has been followed regarding integration proposals and practices. Published privacy impact assessments were adopted as a precondition for this work, as was full disclosure of project content and practices.

Statistics NZ's lead role as custodian of cross government integrated datasets for statistical purposes was strengthened by a formal cabinet minute in 1998 that provided a solid mandate for the organisation. The organisation has progressively built broad stakeholder support for this role leveraged off a series of specific integration projects.

Data integration has however, proven to be neither easy nor cheap. Using data in integration exercises requires extensive learning about individual data sources and substantial methodological and systems development.

Given this investment, the challenge is to ensure that this information is available for a wide range of uses to maximise value.

Statistics New Zealand has developed a comprehensive set of Data Integration Protocols published in a Data Integration Manual (Statistics NZ 2006). These protocols were designed to enable consistent application of policies and procedures across data integration projects.

The challenge was to facilitate use while managing multi-stakeholder interests including:

- policy agency information needs
- public expectations regarding safe data management and use (often articulated through agencies such as the Office of the Privacy Commissioner)
- legal and operational interests of the original administrative data custodians such as our taxation, social development, injury, health and education agencies.

These protocols are soon to be reviewed so that they better reflect the current environment as the demand for integrated datasets and our experience in undertaking this work increases. The current challenge is to provide guidelines for safe practice that support the development and retention of integrated databases that not only meet current but also future information needs.

Legislation defines 'feasible space' – policies determine practice

Legislation defines the 'feasible space' within which a NSO (or indeed any other organisation) can conduct data integration exercises. Often, however, it will be the policy decisions of the agency whose data it is that act as the binding constraint on possible data integration exercises.

The different viewpoints behind New Zealand's Public Records Act (2005) and the Statistics Act (1975) provide an example of the tension that this can create. For example should data be kept for its value as a public record or destroyed to preserve confidentiality? In this instance Statistics NZ approached Archives NZ to get agreement that if and when any integrated data is created, we would destroy it once it has been used for its intended statistical purpose. Initial discussions were

difficult as Archives NZ deal in records that exist and were reluctant to make an unlimited commitment to the future disposal of records that did not yet exist.

However, the establishment of a record class (type of record) based on existing integrated data provided a mechanism for reaching agreement. Statistics NZ proposed that already created integrated datasets (such as the census-mortality study) were examples of a wider class of 'integrated data', and then further proposed that this class of data could be destroyed once its statistical purpose had been met.

This proposal was approved by the Chief Archivist. It is interesting to note that, as yet, Statistics NZ has not destroyed any of the integrated datasets we have created as all are still in use. This raises an important question around the long-term value of integrated data, and whether they are better seen as public records of enduring value or the temporary working files of a statistical office.

Another example of such tensions can be seen in Statistics NZ's approach to retention of name and address details from the 5-yearly Census of Population and Dwellings. While there is no legislative requirement to discard individual name and address information after this information had been used in editing and coding processes, Statistics NZ has for many years undertaken in previous censuses to not retain name and address information once it had been used for processing purposes. For the 2011 Census[1], there had been no undertaking to discard the name and address information once processing of the data has been completed. This will enhance the feasibility of future data integration projects that use the Census, resulting in more robust matching.

Both examples illustrate how policy can evolve within a given legislative framework.

Managing confidentiality and privacy concerns

Prior to undertaking any data integration exercise, it is essential that the legislative and policy drivers of the agency supplying the data are understood. In some cases the obligations of the source agency have to continue to be preserved in the statistical office and this may require additional security procedures. Statistics NZ enjoys a productive working relationship with the

[1] Note: Following an earthquake that struck Christchurch (New Zealand's second-largest city) on February 22 2011, the decision was made to postpone the running of the 2011 Census of Population and Dwellings, which was scheduled for 8 March 2011. At the time of writing, no decision has been made regarding the re-scheduling of the 2011 Census.

New Zealand taxation authority (Inland Revenue) and this has been essential to designing procedures that ensure that issues are addressed appropriately.

Privacy agency

Data integration projects also need to ensure the support of the agency responsible for overseeing privacy regulations. In practice this means identifying threshold issues that are of particular concern and attempting to address these as generically as possible so that consistent practical methods are applied. In practice this is a process of constructing protocols that cover use, access, and storage of integrated data. These protocols have to meet legal constraints and demonstrate an awareness of stakeholders' main concerns.

As an example, one of the requirements of the Privacy Act is that where another agency has assigned a unique identifier to an individual (for example an individual's tax number), that identifier cannot be used by another agency for another purpose, even where this purpose is legally allowed.

Perception can be as important as reality

As well as ensuring that legislature, privacy, and confidentiality concerns are addressed in practice, it is also important to manage the perceptions surrounding these concerns. An example where Statistics NZ incurred inefficiencies and cost explicitly to manage perceptions occurred in the development of the information technology (IT) systems for the LEED data integration project.

The IT system was designed as a set of databases on a server that was physically separate from the rest of the servers at Statistics NZ.

This physical separation of the datasets did not actually make the data any more secure – robust database security protocols would have achieved the same thing. What it did provide was the ability to go out externally with a simple message – the LEED data is stored in a physically separate environment from the rest of our data.

Another example from another agency overseas illustrates the same point. There, an administrative dataset was not only housed on a physically separate computer system, but this computer was actually chained (and padlocked) to a structural column in the building – notwithstanding the fact that the building had armed guards at the main doors and swipe-card access controls to the floors of the building.

What is important here was not the actual security, but the perception it provided. In many instances management of perceptions regarding data integration programmes is at least as important as the actual practices undertaken.

Access to the data

Once a data integration exercise has been undertaken, it is important that the resulting dataset be used to the fullest extent possible to get the best return from the (often significant) investment. External researchers form a key community of interest, including academics, and government and non-government researchers.

In many cases, legislation will determine who is able to access the resulting data. In order to comply with the requirements of the current Statistics Act, Statistics NZ has in many cases had to second external government researchers into its organisation to enable them to access the data. In other instances, academic researchers have collaborated with internal Statistics NZ staff, with the academic researchers specifying the research topic and writing the programs, and the staff member running these against the raw data and feeding back aggregated results.

Provision of microdata within a suitably secure data laboratory is also a mechanism that has been used to enable access to integrated datasets to appropriate people.

These are all mechanisms to enable access within a set legislative framework that have been established to manage the interests and concerns of a wide range of stakeholders.

Managing relationships with external agencies

Statistics NZ has a mandated leadership role in New Zealand's Official Statistics System (OSS). Leading initiatives that fill significant information gaps is a key feature of this role but this can at times generate interesting tensions.

For example, part of the reason for creating integrated files is to examine the validity of previously unverifiable assumptions that have been built into social and economic programmes. It is possible that this type of work could show that these assumptions are questionable, or at least are becoming more questionable. This has the potential to cause tension where the agency supplying the data perceives that the possible use of the data (for valid official statistical purposes) by other agencies could call into question their effectiveness in delivering on government goals. Good external engagement and relationship management processes are necessary to mitigate this risk.

To manage these types of situations it is essential to have representatives of the relevant agencies as part of the governance and analysis process. Often they add a sharp understanding of the policy context and are a conduit for information back to the core agencies. This can assist agencies preparing to incorporate the new information into policy design and mitigate the risk of unproductive inter agency conflict.

It is also important to be aware that as circumstances change, relationships can also change. Attitudes of agencies to a NSO data integration exercise can change as a result of changes in the underlying legislative and policy drivers, or simply as a result of a change in personnel within the agencies.

External agencies will often approach a data integration exercise with a degree of caution at first (and hence be relatively restrictive). If the process is managed collaboratively, trust develops and they may become more open to extended use of the data. Successful precedents are also a key element in this evolving relationship. Success does build the foundations for future success.

6 Future challenges

Statistics NZ is working on a number of new data integration projects aimed at improving the use and accessibility of Statistics NZ's (and other agencies') data for research purposes. These are requiring us to re-examine our policy and legislative environment. These challenges include:

Review of our data integration protocols

Our current data integration protocols were developed over five years ago, and reflect both the policy drivers at that time as well as the legislative environment and linking techniques. Recent developments in legislation and the underlying policy drivers have occasioned a review of these protocols to ensure that they adequately reflect the current environment.

Changes to the Statistics Act (1975)

A change to the Statistics Act (1975) is currently under review. The proposed change will allow a wider range of researchers to access microdata under controlled environments for valid statistical research purposes. This change is designed to meet changing demands for access to a wide range of data. Rather than restricting access to microdata to government researchers, the intention is to open up access to 'approved researchers' who are conducting valid/approved statistical research. This amendment is scheduled to be followed by a full review of the Statistics Act within the next five years.

Managing the tension between statistical and operational purposes

Periodically, Statistics NZ has faced contention between the statistical and operational uses of integrated datasets. While the drive for cross agency collaboration and data sharing offers many opportunities for a NSO it also creates expectations that the information resource created will be made available for broad use. In this situation boundary issues between statistical and operational functions may be tested.

For example, for an NSO where does the boundary sit between the legitimate use of data for statistical evaluation of a programme, which includes innovation subsidies to promote growth, and the policy outcomes that may then impact on individual business participants? In particular, if they are affected by the subsequent cancellation of the programme as a result of the statistical evaluation.

The drive to maximise the use of these rich data sources can and does raise challenging issues regarding legitimate research and statistical purposes and subsequent operational outcomes. We have found that this is not an issue that can simply be resolved but is rather a tension that requires on-going and careful management based on active cross agency engagement.

LEED extensions (iLEED)

Since the initial development of the LEED dataset, a number of additional sources of administrative data have been integrated with the core LEED data.

Examples of data integrated with LEED include:

- social assistance benefits data
- tertiary education enrolment and attainment data
- Household Labour Force Survey data.

All of these integration projects have been undertaken as separate, stand-alone integration projects, with no sharing of data between them. A project is currently underway which is seeking to coordinate these diverse projects within a managed environment that supports the integration of multiple data streams in a safe and consistent way.

This initiative is one of a range looking at establishing core social and economic research databases as elements of New Zealand's official statistical system.

7 Concluding remarks

Over the last 10 years, while Statistics NZ has been undertaking data integration projects, our legal and political context has remained largely unchanged. This framework has provided the 'feasible space' within which data integration can be undertaken.

Within this relatively stable legal framework successive Government Statisticians have been required to interpret how to exercise their statutory statistical discretion to meet statistical user's information needs. This has been a dynamic and evolving process that has required the balancing

of a range of factors including legal bounds but also public expectations regarding the protection and legitimate use of their data, government and broader user expectations for new and more detailed statistical products (including microdata) and a strengthening drive for cross government efficiency.

Statistics NZ has established and developed key relationships over this period to create and built on successive precedents to steadily develop broad support for data integration as a mainstream and now core activity of our NSO.

The key to the success of this process has not been legislative or policy change but a clear vision and consistent and persistent leadership.

Integration projects have significant potential to provide answers to complex policy questions. The drive for extensions of this work to meet challenging user expectations is stronger than ever.

The risk however of a loss of faith by the public or partner agencies also remains. This is an enduring tension that must be monitored and managed, it is not a problem that can simply be resolved.

The experience of Statistics NZ and many NSO's has shown that statistical agencies are well placed to successfully manage these tensions. Our agencies are trusted data custodians and can provide the 'safe pair of hands' that are necessary to assure the public, stakeholder agencies and government that complex datasets can be used safely and effectively to meet the ever growing need for information to better inform decision making.

8 References

Blakely, T (2002). The New Zealand Census–Mortality Study: Socioeconomic inequalities and adult mortality 1991–94. Wellington: Ministry of Health. Available from http://www.moh.govt.nz/moh.nsf/pagesmh/1741?Open

Blakely, T., Shaw, C., Atkinson, J., Tobias, M., Bastiampillai, N., Sloane, K., Sarfati, D., & Cunningham, R. 2010. Cancer Trends: Trends in Incidence by Ethnic and Socioeconomic Group, New Zealand 1981-2004. Wellington: University of Otago, and Ministry of Health. Available from http://www.moh.govt.nz/moh.nsf/indexmh/cancer-trends-incidence-by-ethnic-socioeconomic-nz-1981-2004

The Injury Information Manager. (nd). Retrieved 17 May 2011, from http://www.stats.govt.nz/browse_for_stats/health/injuries/injury-information-manager.aspx

Linked Employer-Employee Data (LEED). (nd). Retrieved 17 May 2011, from http://www.stats.govt.nz/browse_for_stats/work_income_and_spending/employment_and_unemployment/leed.aspx

OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data. (nd). Retrieved 17 May 2011, from http://www.oecd.org/document/20/0,3746,en_2649_34255_15589524_1_1_1_1,00.html

Privacy Act 1993. (1 April 2011). Retrieved 17 May 2011, from http://www.legislation.govt.nz/act/public/1993/0028/latest/DLM296639.html

Public Records Act 2005. (1 February 2011). Retrieved 17 May 2011, from http://www.legislation.govt.nz/act/public/2005/0040/latest/DLM345529.html

State Sector Act 1988. (29 March 2011). Retrieved 17 May 2011, from http://www.legislation.govt.nz/act/public/1988/0020/latest/DLM129110.html

Statistics Act 1975. (1 February 2011). Retrieved 17 May 2011, from http://www.legislation.govt.nz/act/public/1975/0001/latest/DLM430705.html

Statistics New Zealand. (2005). Matching and Results of the Student Loans Data Integration Project 2000–2002. Available from http://www.stats.govt.nz/~/media/Statistics/Publications/Research-reports/Student-loan-2000-2002.ashx

Statistics New Zealand (2006). Data Integration Manual. Available from http://www.stats.govt.nz/~/media/Statistics/about-us/policies-protocols-guidelines/Data-integration-Further-technical-info/DataIntegrationManual.ashx

Statistics New Zealand. (2009). Inter-censal Ethnic Mobility Study 2001-2006. In *Final report of a review of the official ethnicity statistical standard 2009.* Available from http://www.stats.govt.nz/~/media/Statistics/Publications/Census/2011-Census/final-report-review-official-ethnicity-statistical-standard-2009.ashx

Tax Administration Act 1994. (1 April 2011). Retrieved 17 May 2011, from http://www.legislation.govt.nz/act/public/1994/0166/latest/DLM348343.html

A Whole of Government Approach. (nd). Retrieved 17 May 2011, from http://www.e.govt.nz/plone/archive/about-egovt/programme/e-gov-strategy-dec-01/chapter6.html#Toc536000742