Developing the Prototype Longitudinal Business Database: New Zealand's experience

Paper presented at the International Association for Official Statistics Conference, Shanghai, People's Republic of China

14-16 October 2008

Julia Gretton[1]

Abstract

This paper describes Statistics New Zealand's development of the prototype Longitudinal Business Database (LBD). The LBD is the result of a two-year feasibility study to integrate longitudinal administrative and survey data, at the enterprise level, that will meet users' needs to better understand the dynamics of enterprise performance without increasing respondent load. The LBD has been used to produce a selection of prototype official statistics. These outputs demonstrate the benefits of integrating unit-record data from survey and administrative sources. The paper covers current and future uses of the LBD, its construction, the legislative environment, the challenges faced and the solutions adopted.

---

1 The author wishes to thank Richard Fabling and Claire Powell for their significant contributions and Gary Dunnet and Hamish Hill for their helpful comments. Corresponding Author: Julia Gretton (Julia.Gretton@stats.govt.nz), Business Performance and Agriculture, Statistics New Zealand, P O Box 2922, Wellington, New Zealand.

1. Motivation

The need for a common set of financial variables linked to non-financial survey data became clear when Statistics New Zealand began developing a suite of new business performance surveys in 2003. It was thought that the impact on respondent load could be reduced if the common financial variables could be collected once and used repeatedly.

With this in mind, Statistics NZ undertook a two-year project known as IBULDD (Improved Business Understanding via Longitudinal Database Development), to test the feasibility of producing such data. The project was funded by the Cross-Departmental Research Pool with a significant additional contribution from the Ministry of Economic Development (MED).

The project had four main objectives:

- produce new and improved official statistics
- improve the access and usability of microdata for researchers
- reduce respondent burden
- improve the efficiency of Statistics NZ's data handling (statistical architecture).

Policymakers and a range of interest groups need official statistics that allow them to make evidence-based decisions. IBULDD developed and tested the methodology to produce new official statistics that measure the dynamics of business growth and performance. The ability to produce longitudinal outputs from the prototype Longitudinal Business Database (LBD) represents a 'quantum leap' in information for users of business performance data. The project successfully delivered the LBD, which will be a key element in Statistics NZ's new statistical architecture.

This paper describes the LBD; the project that developed the LBD, including the challenges faced and solutions adopted; New Zealand's legislative environment; and the current and future uses of the LBD.[2]

2. Description of the LBD

The LBD is a longitudinal dataset of integrated business-related data. It is enterprise-based and annual. Its length, breadth and depth make it a powerful tool.

- Length – annual data for 2000–06 (some data is available for earlier and later years)
- Breadth – population coverage and imputation ensures that key financial data is available for a census of enterprises
- Depth – range of data includes information on business demographics, financial data, employment, goods exports, government assistance, and management practices.

The backbone of the LBD is the Longitudinal Business Frame (LBF). This is a longitudinal register of businesses and includes demographic data. Administrative data from other government agencies, including Inland Revenue and the New Zealand Customs Service, is linked to the LBF, along with a number of Statistics NZ sample surveys that measure business practices and performance. A key part of the IBULDD work involved establishing the common linkages, common units, and common timeframes for all the source data. Table 1 presents the integrated components of the LBD.

2 For interested readers more information can be obtained from the extended version of this paper, Fabling, Gretton and Powell (2008) and the detailed prototype official statistics Statistics NZ (2007).

Table 1

Prototype Longitudinal Business Database (LBD)
Integrated components

Component	Years	Description
The backbone of the LBD		
Longitudinal Business Frame (LBF)	2000–06	Contains longitudinally linked data for most enterprises operating in New Zealand. It includes information on employment, location, industrial activity, and ownership relationships. The LBF enables individual business units to be tracked over time.
Administrative data linked to the LBF		
Business Activity Indicator (BAI)	1992–2006	The BAI is a monthly series based on the supply of administrative data from Inland Revenue. The main source of this data is Inland Revenue's GST (goods and services tax) 101 form. GST is a tax based on the sale of goods and services.
Financial accounts (IR10)	1999–2006	The Accounts Information Form (IR10) collects a summary of information relating to the business and its operations (profit and loss statement, and balance sheet). The Inland Revenue-supplied data is transformed by Statistics NZ and then linked to IBULDD.
Company tax returns (IR4)	1999–2006	The IR4 income tax return is compulsory for businesses that are registered as companies. It includes income, tax calculation, refunds and/or transfers, provisional tax, and disclosures. IR4 data is supplied to Statistics NZ by Inland Revenue and is then linked to IBULDD.
Linked Employer-Employee Database (LEED)	2000–06	A Statistics NZ integrated database that provides an insight into the operation of New Zealand's labour market, such as job and worker flows. Created by linking a longitudinal employer series from the Business Frame to a longitudinal series of employer monthly schedule (EMS) payroll data from Inland Revenue.
Overseas Merchandise Trade data	1988–2007	A daily shipment-level series based on administrative data supplied by the New Zealand Customs Service. In the LBD, this daily data is aggregated to monthly, and provides information on the importing and exporting of merchandise goods between New Zealand and other countries.
Government assistance data	2000–06	Information on the assistance provided directly to businesses by the Foundation for Research, Science and Technology, New Zealand Trade and Enterprise, and Te Puni Kōkiri.
Sample survey data linked to the LBF		
Annual Enterprise Survey (AES)	1997–2006	Provides annual financial performance and financial position information about industry groups operating within New Zealand. AES is the basis of the national accounts produced by Statistics NZ.
Business Operations Survey (BOS)	2005–06	Collects measures of business performance and a range of practices and behaviours which may have some impact on that performance, including innovation and business use of information and communication technology.
Innovation Survey	2003	Collected information on the characteristics of innovation in New Zealand private-sector businesses.
Research & Development Survey (R&D)	Biennially 1996–2006	Collects information on business, government and higher education (university) spending on R&D.
Business Practices Survey (BPS)	2001	Collected information on business and management practices.
Business Finance Survey (BFS)	2004	Collected information on the capital structure of businesses in New Zealand, the sources of finance they use, and their recent financing experiences.

Population

The population of the LBD consists of all enterprises that have at any time been included on the Statistics NZ Business Frame (and consequently have been allocated an industry classification). Enterprises are included on the Business Frame when they are identified (from administrative information) as having met the threshold criteria. Eligible enterprises are further defined as being economically active (and therefore in scope for the prototype statistics presented later in this paper) in a particular year if they have filed administrative data with selected financial or employment values greater than zero.[3]

3. The IBULDD project

The IBULDD project began at a time when Statistics NZ was working on a number of data integration projects that used administrative data, including:

- the student loan and allowance integrated dataset – currently used to produce annual statistics on the borrowing, qualifications and income of students who have participated in the student loans scheme or have received an allowance
- LEED – currently used to produce quarterly and annual releases on the labour market dynamics of New Zealand.

This meant that some of the legal aspects of integrating administrative data had already been investigated. In addition, the LBF, used by the LBD, had already been developed by the LEED project. However, there still remained challenges specific to the development of the LBD and these are examined in a later section.

At the same time, Statistics NZ was developing its vision for statistical architecture, which included the development of a production LBD. The IBULDD project fitted within the architecture by developing the prototype model.

The project followed traditional project management principles. This section describes some of the key aspects of the project.

Governance

Governance was critical to the project's success. The strong commitment of the sponsor, business owner, project manager and the project team, along with the extensive support across government, enabled challenges and issues to be resolved effectively. The broad range of people involved in the project, both within Statistics NZ and externally, is illustrated by the governance groups below:

- a sponsors' group (made up of internal and external members)
- an internal steering committee
- a project board (internal members)
- an external technical advisory group
- a working group (the core project team)
- other internal advisors (from related projects or infrastructure).

Needs analysis

One of the first stages of the project was to analyse the needs of users, in particular researchers, in order to meet the second objective of the project (improve the access and usability of microdata for researchers). This involved working with a number of

3 Consequently, there is some under-coverage of micro-enterprises because they never meet the criteria for being on the Business Frame. For more information, see Statistics New Zealand (2007).

government agencies to identify research questions that could potentially be answered by the LBD. These needs, combined with Statistics NZ's objectives (see section 1, Motivation), were used to determine the structure and content of the LBD.

Data providers

As the LBD uses significant amounts of administrative and survey data, the cooperation of each data provider was critical. Most of the administrative data (excluding the government assistance data) is processed and used elsewhere within Statistics NZ. The IBULDD project team worked with the relevant areas of Statistics NZ to:

- gain approval from the providing agencies to link the data to the LBD
- access and understand the data and metadata (including understanding the existing processes applied to the data).

The government assistance data was collected and linked to the LBD as part of an interdepartmental policy evaluation project.

International peer group review

Because the project was so novel, an international peer group review was held 10 months into the project, in October 2006. The review team was made up of representatives from the Australian Bureau of Statistics, the US Bureau of the Census and the UK Office for National Statistics. The team expressed enthusiasm for the aims and methods expounded, remarked positively on the way Statistics NZ had addressed many difficult problems, and provided recommendations for the continued development of the LBD (Blanchette, Jarmin and Ritchie 2006). The peer group was informed of progress throughout the project and continues to be notified of statistical and research outputs.

4. Challenges and solutions

The challenges faced during the development of the LBD included: designing the structure of the database, developing imputation methodologies, establishing microdata access, and ensuring confidentiality was protected. This section examines each of these.

Database structure

The work on the structure of the database began with deciding on and applying the standard units and periodicity, followed by structuring the data and metadata. The approach was to keep it simple to start with and enhance as necessary. This work is detailed below.

Periodicity

Most of the source data is provided on an annual financial-year basis. Because most businesses in New Zealand use a March balance date, the decision was made to create an annual database using March years. When needed, data was aggregated to an annual basis using the enterprise's own balance date. Each balance date was then allocated to the March year with the greatest overlap.

All core financial information and business characteristics are available for 2000–06. While this is good, it can be improved by adding the data for current years as it becomes available. The IBULDD project has prepared for this by establishing systems and processes for linking additional years of data. The costs of an annual update are comparatively small, but the benefits are significant. This will be particularly true as the LBD extends to encompass complete business cycles.

Unit of observation

Most of the data used in the LBD is enterprise or tax-unit based. Tax data is provided by IRD number, which, for almost all economically significant businesses on the Business Frame, has a one-to-one relationship with enterprise number. The decision was made to use the enterprise as the unit of observation.

The IBULDD project developed and tested the methodologies to aggregate data to the enterprise level. For example, overseas merchandise trade data records consignments against client numbers. This data is mapped to a unique enterprise number and aggregated to the relevant March year. The Annual Enterprise Survey is based on kind of activity units, and is aggregated to enterprise level before it is linked into the LBD. Government assistance data records grants, advice, or other services provided. This data is probabilistically matched to the LBF and then aggregated to enterprise level within IBULDD.

An enterprise is defined as a business or service entity operating in New Zealand, and is based on a legal entity. From an economic perspective, enterprises are subject to 'false' births and deaths. For example, if a business reregisters with the Companies Office, it may be assigned a new enterprise number. The original enterprise number will appear to have ceased operating even though the same economic activity may be continuing in the same location with the same owners. The proposed solution to this issue is to develop the concept of a longitudinal enterprise, which will refer to the business itself, not the legal entity. The LBF tracks individual plants (geographic units), using several methods including employment matching. In simple cases of false exit and entry, these unit-level links can be used to repair enterprise-level links. However, further work is needed for groups that are more complex.

Database structure

The LBD is stored on a Microsoft SQL server, which can be queried using SQL or SAS. The requirements were that the storage be secure, be able to store large amounts of data, and be easy and quick to query.

Raw data from each source is stored as a ‘load’ table. This data is then aggregated to an annual, enterprise-based ‘fact’ table. The unique identifiers in each fact table are the enterprise and year. An in-principle decision was made to make both raw and transformed data available to users so that any change to the data can be unwound. This approach recognises that generic rules are not best for all applications.

Metadata

As data from each source was collected and linked to the LBD, the metadata was also collected and collated into a central repository. Metadata already existed for survey data, but some additional metadata was required for the administrative data. An ongoing challenge is to develop better methods to distil and harmonise this metadata.

Imputation

Survey data linked to the LBD, as well as data from the Business Activity Indicator (BAI) series, have already had some imputation applied during the survey processing stage. IBULDD developed the imputation methodology to apply to missing BAI and Inland Revenue IR10 data. Cases that lead to an enterprise receiving an imputed value for sales or income include:

- enterprises that fall below mandatory size thresholds (currently $40,000) for filing GST
- enterprises that file an incomplete or internally inconsistent IR10
- enterprises that are GST-exempt

- misidentification of start-up and ceasing dates for enterprises (eg due to administrative lags in filing).

Three imputation methods were used to impute missing BAI and IR10 values in the LBD. Interpolation was the first method of choice. If this was not applicable historical imputation was used, and if this was not applicable then donor imputation was used. The nearest neighbour for use as a donor is determined using financial and employment data from LEED, IR10 and BAI.

Confidentiality

Linking data from different sources raises issues about privacy, confidentiality and security. The LBD has the following protection methods in place.

- Access to the entire LBD is restricted to the IBULDD project team.
- Government departments and researchers may only access an anonymised version of the LBD, and only through Statistics NZ's Data Lab. A rigorous application process and strict eligibility criteria apply, based on the requirements of the Statistics Act 1975 and Inland Revenue.
- The anonymised version of the LBD has had all variables that might identify an enterprise removed.
- The LBD is protected by Statistics NZ's strict security policy that applies to both the IT system and the physical building.
- All outputs from the Data Lab are checked to ensure they do not breach confidentiality.

5. Legislative environment

The LBD was developed within the legislative requirements of the Statistics Act 1975, the Tax Administration Act 1994 and the Privacy Act 1993.

The Statistics Act covers how Statistics NZ collects, uses, securely stores, and manages access to data. Data obtained under the Act is to be used only for statistical purposes. The Act provides for discretion to approve data access by other government departments for bona fide statistical and research purposes under tightly controlled conditions. Statistics NZ is also able to contract independent contractors (which can include individuals from non-government or academic institutions) to carry out work that contributes towards enabling Statistics NZ to deliver quality official statistics.

The Tax Administration Act allows Inland Revenue to provide data to Statistics NZ under the discretion of the Commissioner, who needs to be satisfied that the integrity of the tax system isn't going to be compromised by Statistics NZ's use of the data. The data that can be provided is that which is "not undesirable to disclose", and which is "reasonably necessary" for Statistics NZ to carry out its official duties under the Statistics Act.

The Tax Administration Act also covers Inland Revenue's declaration of secrecy that must be signed before accessing the data.

The Privacy Act protects the privacy of individuals. While there is no specific legislative requirement around data matching for statistical purposes, Statistics NZ is concerned that the privacy of individual people is safeguarded, and is seen to be protected. The Data Integration Manual describes Statistics NZ's Data Integration Policy. As the LBD includes information on enterprises only, and not on individuals,

some of the requirements, such as preparation of a privacy impact assessment, do not apply.[4]

Microdata access

An anonymised version of the LBD can be accessed through Statistics NZ's Data Lab or through secondment to Statistics NZ. A rigorous application process and strict eligibility criteria, based on the requirements of the Statistics Act 1975, apply. Government departments and other researchers (with backing from a recognised institution) undertaking research for public benefit can be granted access to microdata.

The induction process for new users includes a description of the structure of the database and the imputation applied. All outputs from the LBD are confidentialised to protect individual businesses from identification.

6. Current use

The IBULDD project confirmed that the creation of a database such as the LBD is feasible and valuable. The value of the LBD is illustrated by its current uses.

- A set of prototype new and improved official statistics have been produced.
- It has enabled government to evaluate policy and understand business dynamics without commissioning new surveys and adding to respondent load.
- Lessons learnt during the IBULDD project have been used to develop the road map for the implementation of the statistical architecture, which will improve the efficiency of Statistics NZ's data handling (McKenzie 2008).
- Statistics NZ's understanding of the difference between the data and metadata requirements for aggregate, cross-sectional statistics, and the requirements for microdata research and longitudinal analysis, has increased.
- A dozen researchers have access to and are using the LBD for microdata research related to topical public policy issues.

The LBD became available in April 2007. Within only three months, four research papers were produced by the Ministry of Economic Development (MED), and presented at the New Zealand Association of Economists Conference (NZAE) 2007 on these topics:

- firm level patterns in merchandise trade
- firm dynamics, market structure and performance
- currency hedging behaviours of exporters
- comparison of quantitative and qualitative performance measures.[5]

Papers on the following topics were presented at the 2008 conference ESAM/NZAE (Econometric Society Australasian Meeting):

- estimating aggregate R&D expenditure using a micro model (the Reserve Bank of New Zealand (RBNZ) and Motu Economic and Public Policy Research (Motu))

4 The LBD does contain aggregated worker information from LEED.

5 See Fabling & Sanderson (2008); Fabling et al. (2008); Fabling & Grimes (2008a); and Fabling, Grimes & Stevens (2008), respectively.

- exporters' optimal and selective hedging choices (RBNZ, Motu and University of Waikato)
- complementary personnel practices and firm performance choices (RBNZ, Motu and University of Waikato)
- export market choice for New Zealand firms (RBNZ, Motu and University of Waikato)
- multi-product exporters and product switching behaviour of New Zealand firms (the New Zealand Treasury)
- relationship banking and access to finance in New Zealand firms (MED).[6]

Improvements underway

The 2008 annual update of the LBD is due to be completed in December 2008. This update will add an additional year's worth of data to the database (data for the year ending March 2007), as well as updating any historical data that has been edited by the data providers. During this update, imputation will be rerun over all the years in the LBD, including the new 2007 data.

In addition to maintaining and updating the current data in the LBD, Statistics NZ is continually exploring and evaluating options to add new data from existing, and new, survey and administrative sources.

7. Future use

The IBULDD project developed the framework necessary to integrate data using common linkages, units and timeframes. This framework allows additional years of data and additional datasets to be efficiently integrated into the LBD, enabling its length and depth of information to increase over time. Beyond this updating process, it is intended that a future production model of the LBD would be used to:

- reduce respondent load by removing the need for financial questions in surveys
- produce new official statistics and enable existing official statistics to be produced more efficiently
- contribute to dissemination tools (eg benchmarking, Table Builder)
- support the production of new financial performance metrics and continue to be a valuable resource for microdata research.

Benefits of using the LBD for microdata research

The LBD has a unique ability to "turn data into relevant knowledge, efficiently" (Statistics NZ mission), by:

- providing answers to more specific questions on the economy, business behaviour, and performance distributions, than could be answered using unlinked raw data
- meeting the expectations of a national statistical office to adapt to a changing world and provide information as and when it is needed
- providing the information needed to meet the increasing demand for evidence-based policy

6 See Fabling (2008); Fabling and Grimes (2008b, c); Fabling, Grimes and Sanderson (2008); Adalet (2008); and Stevens (2008), respectively.

- promoting new insights on the impact of policy and business practices through the linking of different datasets
- giving information on the economy that can be used to inform government, businesses, communities, and citizens.

The LBD is an example of intelligent data usage because it:

- increases the amount of information extracted from data already collected
- enables longitudinal analysis
- increases and maximises the use of administrative data
- provides a valuable tool for government that does not increase respondent load
- increases the quality of information by giving alternatives to using imputed data and improving imputation methods.

8. References

Adalet M (2008). *Multi-product Exporters and Product Switching Behaviour of New Zealand Firms.* Paper presented at ESAM/NZAE 2008.

Blanchette J, Jarmin R and Ritchie F (2006). *IBULDD Project Peer Group Review: Findings and Recommendations,* Statistics NZ, Wellington.

Fabling R (2007*).* "How innovative are New Zealand firms? Quantifying & relating organisational and marketing innovation to traditional science & technology indicators". In *Science, Technology and Innovation Indicators in a Changing World: Responding to Policy Needs*, OECD, Paris.

Fabling R (2008*). Will a BERD model fly? Estimating aggregate R&D expenditure using a micro model*. Paper presented at ESAM/NZAE 2008.

Fabling R, Gretton J and Powell C (2008). *Developing the Prototype Longitudinal Business Database,* Statistics NZ, Wellington.

Fabling R and Grimes A (2008a). *Do Exporters Cut the Hedge? Who Hedges, When and Why,* MED Occasional Paper 08/02, Wellington.

Fabling R and Grimes A (2008b). *Over the Hedge or Under It? Exporters' Optimal and Selective Hedging Choices.* Paper presented at ESAM/NZAE 2008.

Fabling R and Grimes A (2008c). *The "suite" smell of success: Complementary personnel practices and firm performance.* Paper presented at ESAM/NZAE 2008.

Fabling R, Grimes A and Sanderson L (2008). *Export Market Choice for New Zealand Firms.* Paper presented at ESAM/NZAE 2008.

Fabling R, Grimes A, Sanderson L and Stevens P (2008). *Some Rise by Sin, and Some by Virtue Fall: Firm Dynamics, Market Structure and Performance,* MED Occasional Paper 08/01, Wellington.

Fabling R, Grimes A and Stevens P (2008). *Comparison of Qualitative and Quantitative Firm Performance Measures,* MED Occasional Paper 08/04, Wellington.

Fabling R and Sanderson L (2008). *Firm-Level Patterns in Merchandise Trade*, MED Occasional Paper 08/03, Wellington.

McKenzie R (2008). *A Statistical Architecture for a New Century*. Paper presented at this conference.

Statistics New Zealand (2007). *Potential Outputs from the Longitudinal Business Database*, Statistics NZ, Wellington.

Stevens P (2008). *Just Good Friends? Relationship Banking and Access to Finance in New Zealand Firms.* Paper presented at ESAM/NZAE 2008.

Related links

LBD: http://www.stats.govt.nz/economy/business/longitudinal-business-database.htm

LEED: http://www.stats.govt.nz/leed/default.htm

Peer Group Review: http://www.stats.govt.nz/developments/ibuldd-peer-review-report.htm

Data Integration Manual: http://www.stats.govt.nz/NR/rdonlyres/35662748-4DBC-41DA-A519-E6D9D7748C20/0/DataIntegrationManual.pdf

Data Lab: http://www.stats.govt.nz/products-and-services/microdata-access/data-lab/default.htm

NZAE: http://www.nzae.org.nz/conferences/ (for papers presented at ESAM/NZAE 2008.)

Table Builder: http://www.stats.govt.nz/products-and-services/table-builder/default.htm