

# A Statistical Architecture for a New Century

Ron McKenzie

19 May 2008

Paper Presented at International Seminar on the Use of Administrative Data  
Daejeon, Korea

## Abstract

Statistics New Zealand has recently developed a new Statistical Architecture for Economic Statistics. The basic principle of the architecture is that administrative data will be used wherever possible with surveys filling the gaps. Survey data will be integrated with administrative data by a comprehensive tax-based business register. A database approach will be used for large complex businesses.

The objective is to bring core information about every business in the economy into a Longitudinal Business Database that will support the full range of needs being faced by a national statistics office, including aggregate statistics and microdata analysis. New surveys and administrative data will be linked to this core information to support more detailed statistical analysis and relevant official statistics as required.

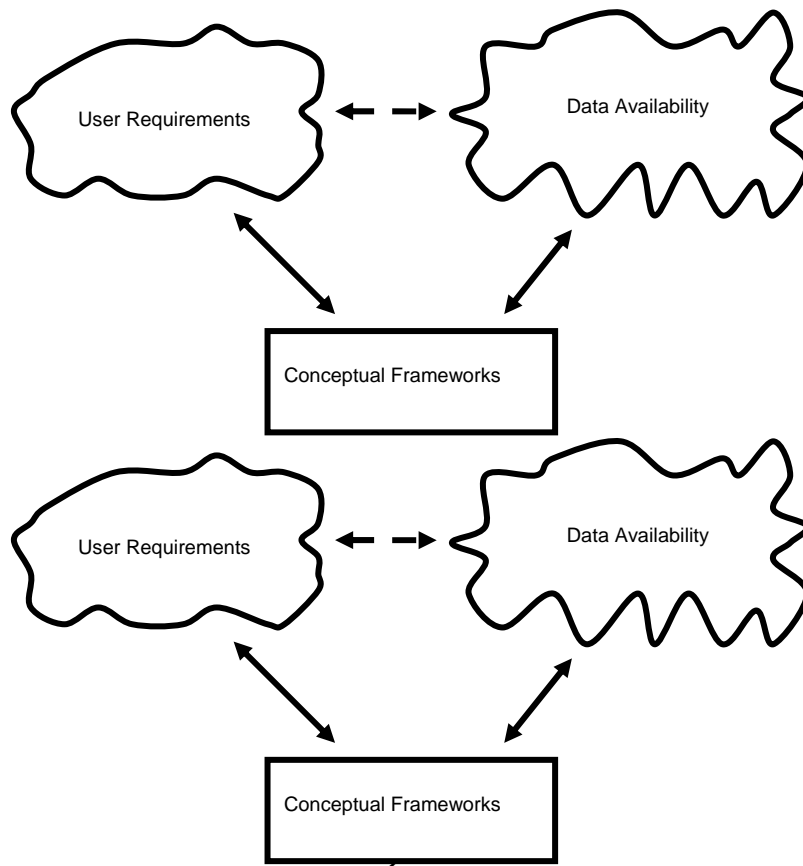
### Architect or Engineer

A core part of an architect's role is to present the design of a new building in a way that will enable the owners to understand what it will look like and how it will function, before any building work starts. Once the owners have accepted the design, the engineers and draughters can then prepare the more detailed plans that are needed to allow the new building to be completed. They will also ensure that all aspects of the new building will comply with the building code.

In much of our work, Statistics New Zealand has taken an engineering approach, using the best available techniques to build an existing design. An architectural approach goes a step further and thinks about how emerging user needs will be met. This approach uses best practice techniques and materials in the best possible way to produce a design that will meet all the new features that users require. The purpose of a statistical architecture is to describe an integrated and systematic approach to data collection that will support future information needs.

Statistics NZ already has a well integrated system of economic statistics. The first part of the paper will describe this existing system. The second part of the paper outlines a vision that will shape the ongoing development of our economic collections to support a broader range of user needs. It will also show how the various parts of the system will fit together.

Architects A team of architects operate under several constraints. They must design a new building that will meet the requirements of the owners, using proven materials and building techniques that comply with the building code. Their work is shaped by the tensions between the dreams of the owner and the limitations of their budget. They also have to deal with the limitations of existing building materials and techniques and the requirements of the building code.



Statisticians face a similar tension between user needs and the availability of information. We attempt to resolve this tension by expressing user needs within standard conceptual frameworks and then fitting the information that can be collected into those frameworks

## **1. INTEGRATED ECONOMIC STATISTICS**

Over the last decade Statistics NZ has developed an efficient and effective system for collecting a broad range of economic information using a combination of administrative data and sample surveys.

### **1.1 Comprehensive Business Register**

The foundation of this collection system is a comprehensive business register called the Business Frame (BF) that provides an unduplicated list of businesses and organisations of interest to Statistics NZ.

Business units (including non-profit institutions and government units) are the basic building block for economic statistics. They make decisions, employ the people and control the assets that make up the economy. Therefore, a business register that records every business active in New Zealand is an essential foundation for economic statistics.

- The Business Frame is a database of all known individual private and public sector businesses and organisations active in New Zealand.
- Every economically significant business (turnover greater than \$30,000 or employing staff) that operates in the New Zealand economy is included on the core BF.
- Limited information about businesses below the “economic significance” threshold is available on a separate part of the register.
- New businesses are identified immediately through registration for Goods and Services Tax (GST).
- Information from the tax system is supplemented wherever possible with information from other administrative sources, such as business directories, land registries, government actuary, etc.
- Information from the companies office is used to update name changes and group structures.
- Corporate group structures have precise ownership hierarchy and ownership levels fully mapped.
- All units are coded to a range of classifications, including institutional sector, industry (5-digit ANZSIC), Business Type, meshblock (regional dimension). All classifications are stored on a linked classifications database (CARS). Tests for coherence are applied as units are being classified on the BF.
- A three-level unit structure based on ISIC ensures that different statistical needs can be supported.
  1. Enterprise (sector view)
  2. Kind of Activity Unit (industry dimension)
  3. Geographic Unit (employment statistics)

- Larger or more complex businesses may have a number of statistical units. Each of the statistical units is given an industry classification based on its predominant activity.
- Agricultural units have been matched to various land-based registers to improve the quality of information.
- Additional non-profit institutions (that have not registered for GST) have been identified using information from registers of incorporated societies and charitable trusts. All non-profit institutions have been classified to the New Zealand Standard Classification of Non-Profit Organisations (NZSCNPO).
- All government units are recorded and classified on the BF. Government trading units and state-owned enterprises are set up as distinct units and classified to the industry appropriate for their activity. Government ownership is identified by their institutional sector code. Local authority trading units are recorded and classified in the same way.
- The Business Frame records business demographic information that can be used for selecting and stratifying sample surveys. Several size indicators, including annual turnover and number of employees, are available for selecting survey populations, stratifying samples, and defining imputation groups. These indicators are updated monthly from administrative sources.
- The updating strategy uses a mix of administrative data and survey information to optimise the quality of information held while minimising respondent load. Large complex businesses (making up more than 85% of the economy by value) are surveyed each year to update their detail. Medium sized business are surveyed every third year or when an administrative trigger indicates that the nature of business may have changed. Small businesses (less than \$200,000 turnover and only 2.5% of the economy) are updated entirely from administrative data.
- The BF records the history of all changes to both individual units and group structures. When the ownership of a local unit changes from one enterprise to another, the BF tracks the local unit provided it carries the same activity with the same staff and plant.
- The BF has a graphical interface that displays the structures of groups. Users can click through to tax data and survey data for any unit at any level in the structure of a group. This is invaluable to an analyst trying to understand changes to financial performance. The impacts of takeovers and restructuring of groups can be quickly identified.
- The BF has been operating for nearly two decades and during that time significant improvements have been made to its efficiency, coverage and quality.

### **1.2 Benefits of a Comprehensive Business Register**

A comprehensive business register has several benefits for the production of economic statistics.

- The Business Frame provides a common reference point for standard classifications for all units. This facilitates the integration of statistical outputs by ensuring that classifications are applied consistently across all surveys and statistical outputs.
- The Business Frame links all economic and financial survey data to the tax system, allowing more effective use of tax data to reduce respondent load.
- All administrative datasets are incorporated to our statistics by first being matched to the Business Frame. This eliminates problems with duplications and inconsistent coverage of administrative datasets.
- The populations for all economic surveys are selected from the BF. This ensures coherence of information between different surveys and administrative data sources. Coverage adjustments are unnecessary, because we always know which units are covered by each data source. Where a unit is included in two different data sources, it can be excluded from whichever is appropriate to ensure that coverage is coherent.
- Administrative data and survey data can be combined in a statistical output with the Business Frame ensuring coherence between data sources. For example, the frame can be partitioned with tax data being used for one partition and survey data being used for the rest.
- When additional information about a specialised population is developed for a new statistical output, it is loaded onto the BF. For example, the population for a recent study of non-profit institutions was built up by identifying those already on the BF and adding missing units identified on lists of incorporated societies and charitable trusts. This ensures that the new information feeds through to other surveys and eventually on to the National Accounts.
- When undertaking supply use balancing, national accountants can have confidence that information from different sides of the accounts are derived from data sources with consistent population coverage and classifications.
- When calculating productivity statistics, the BF is used to ensure coherence between the measures of output and measures of labour inputs and capital services. This ensures that the numerator and denominator cover the same population and are consistently classified.
- The information on the BF history files has been used to create a Longitudinal Business Frame (LBF). This database provides a longitudinal view of BF data that allows the development of businesses to be analysed. This ensures that longitudinal microdata analysis is consistent with other statistics.
- The Linked Employer Employee Database (LEED) joins the LBF with information about their employees derived from a monthly tax form in which employers report the names and addresses and remuneration of all their employees. The LEED provides a link between businesses and individuals, which will be the key to integrating business and social

statistics in the future. It will be the foundation on which all employment statistics are built.

### 1.3 Economy Wide Economic Survey

If the BF is the foundation for economic statistics, the Annual Enterprise Survey (AES) is the superstructure. This survey, which was introduced in 1986, covers almost the entire economy.

The target population for AES is all economically significant businesses operating within New Zealand. The following industries are excluded on pragmatic grounds.

- Residential property operators (L771100-90)
- Foreign government representation (M813000)
- Religious organisations (Q961000)
- Private households employing staff (Q970000).

Residential property owners are considered to be too difficult to measure. The other three exclusions do not make a significant contribution to the NZ economy. When the Charities Commission database comes online in the next few years, coverage will be extended to include religious organisations.

The name of the survey is a little misleading, as the collection unit for the AES is the KAU. By definition, a KAU is engaged in predominantly one activity. The AES is designed to produce estimates by institutional sector and for 107 industries (approximately the four-digit ANZSIC level).

The AES collects detailed measures of financial performance and financial position by industry and institutional sector. Key output variables include income, expenditure, profit, purchases of fixed assets and equity. A full range of business statistics, including economic ratios such as the return on assets and profit margin on sales can be derived from this data.

Since 1999, the AES has used a mix of postal survey and administrative data. The aim is to maximise the quality of estimates while minimising respondent load. The administrative data source is the IR10 a form administered by the Inland Revenue Department (IRD). The IR10 is mandatory for most businesses and collects about thirty key variables from their Statement of Financial Performance and the Statement of Financial Position. It is used for sole traders and partnerships and for all businesses in the agriculture industries.

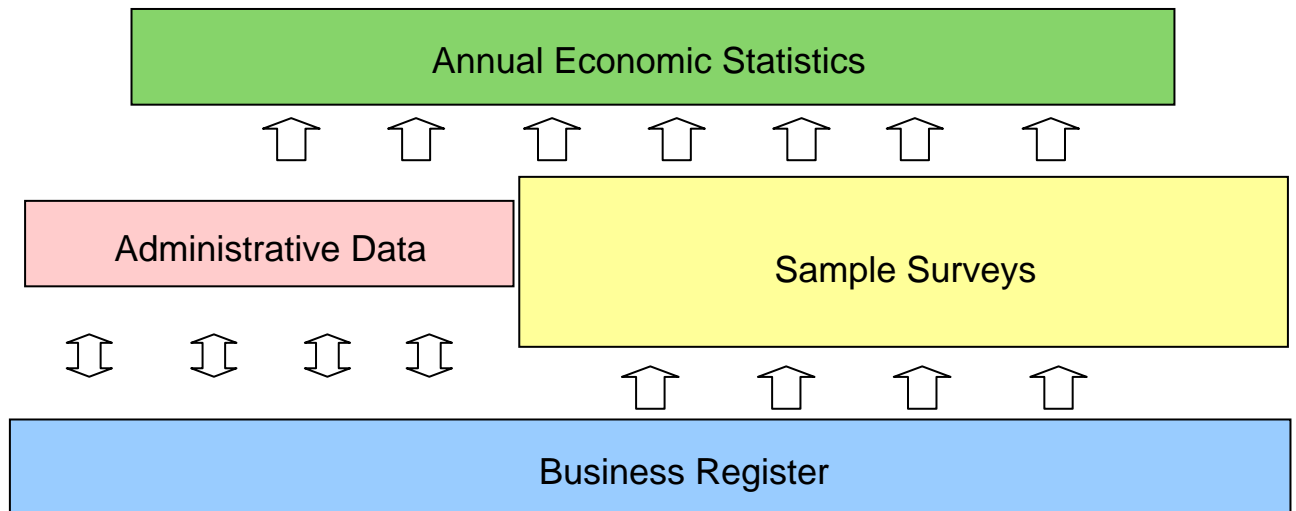
The AES has a three-strata design with stratum being defined by size of turnover and numbers of employees (RME). The three strata are

**Full coverage** - Each industry has a full coverage stratum made up of large or complex units with significant economic activity within their industry group.

**Tax Strata** - Most industries also have a tax stratum where IR10 information is used for self-employed individuals and partnerships with turnover less than \$10 million. More than half of all businesses are in the tax strata.

**Sample strata** - The remaining strata contain a postal sample of medium-sized units drawn from the BF. Respondent load is managed by placing a limit on the number of units sampled each year.

The wide range of activities undertaken by New Zealand businesses makes different types of postal questionnaires essential. Although standard variables are collected in all questionnaires, some variables are tailored to specific types of business or particular industries. Demographic information on the BF is used to ensure that each business gets the right type of questionnaire.



#### 1.4 Benefits of an Economy Wide Survey

A comprehensive economic survey has several benefits.

- Consistent data is collected from across the economy. All the AES questionnaires ask the same core questions and collect a core set of variables. The only differences in the questionnaires are to capture items that are specific to a particular industry or to obtain relevant commodity breakdowns.
- The AES was designed as the principal collection vehicle for data used in the compilation of New Zealand's National Accounts. The survey questionnaires are designed to collect both standard accounting variables and the additional information needed to calculate all core national accounting variables. For example, the questionnaires collect wages and salaries paid. They also collect the other components of compensation of employees as defined in the SNA. This means that compensation of employees can be derived by adding the appropriate line codes for each unit, rather than “adjusting” wages and salaries” at the aggregate level to take account of the estimated difference. This approach ensures coherence between the national accounts and other business statistics.
- As much as possible of the information needed for all the annual accounts is collected in one survey. The AES collects information about production, but also measures interest and dividend flows and capital formation. This contributes to coherence between different accounts.

- The AES linecodes are aggregated to National Accounts variables prior to data editing. Analysts working on the AES are able to check and explain unusual movements in both SNA variables and accounting aggregates as a normal part of data checking. This minimises the need for the compilers of national accounts to check unit record data.
- The AES data feeds into the calculation of the National Accounts through the current price annual industry accounts, which are compiled within an input-output framework. When compilers are confronting data from supply and demand sides of the economy, differences should be the result of different survey designs, rather than inconsistencies between sample frames or classification of units.
- The combination of a comprehensive business register and economy-wide economic survey eliminates the need for benchmarking to five-yearly economic censuses. The national aggregates from AES are of sufficient quality to be incorporated directly into the National Accounts.

### **1.5 Supporting Information**

The Annual Enterprise Survey is supported by a range of other economic surveys that meet specialised needs for information that is not available from administrative sources. The population for each of these surveys is identified from the BF. Some of the topics covered are:

- Innovation
- R&D
- Biotechnology
- International Engagement
- Business Performance
- Balance of Payments

Detailed commodity breakdowns are collected in a survey that covers all industries over a rolling, four-year cycle. This survey asks respondents to provide detailed breakdowns of sales and expenses they recorded in the AES. This approach ensures that the commodity breakdowns that feed into the National Accounts and the Producers Price Index are consistent with other information on the AES.

The surveys that feed into Quarterly GDP are also selected from the BF. This ensures that survey coverage and classification of units is consistent with the Annual Enterprise Survey. Differences still exist due to sampling and information coming from management accounts, but coverage differences are eliminated. Where appropriate, variables of interest are defined in the same way.

## 2. Future Directions

### 2.1 Emerging Needs

The system described above was designed to produce national estimates of the aggregate measures needed by the national accounts and other economic observers. The Annual Enterprise Survey has fulfilled this need with sample surveys supplemented by administrative data for the smaller units. However this strategy does not support the growing demand for statistical analysis to support policy development and to monitor the effect of policies as they are implemented.

- Longitudinal studies of business development will often be focused on small businesses, where sample coverage is limited.
- The non-profit satellite account needs information about units that have relatively small financial flows and numbers of employees. These are under-represented in the AES.
- The AES sample has poor coverage of Maori business activity.
- The existing AES sample does not fully support measures of regional GDP.
- Statistics NZ faces increasing pressure to minimise respondent load.

### 2.2 Microdata Analysis

As policy analysts started to make greater use of statistical analysis, they soon discovered that the answers to many questions of interest are hidden in the details, so the emphasis shifted from analysis of broad aggregates to an increased use of microdata analysis. For example, productivity statistics are produced from aggregate statistics at the industry level, but to fully understand the drivers of economic development, analysis of firm-level productivity has been essential.

Microdata analysis often has a longitudinal basis. For example, understanding how productive firms grow over time is an important aspect of productivity analysis. The shift to microdata analysis has been accompanied by increased integration of datasets from different sources.

These changes in the approach to policy development and statistical analysis have led to a demand for a broader range of information, including:

- unit record datasets for microdata analysis;
- integration of administrative datasets and survey data;
- longitudinal links that track businesses over time;
- detailed regional information;
- links between businesses and households;
- movements between market and non-market activities.

Our challenge is to respond to these needs while minimising respondent load and maintaining the quality of existing outputs.

To deal with these increasing and diverse demands, Statistics NZ has embarked on a strategy that places a greater emphasis on the use of administrative data with sample surveys being used to fill the information gaps.

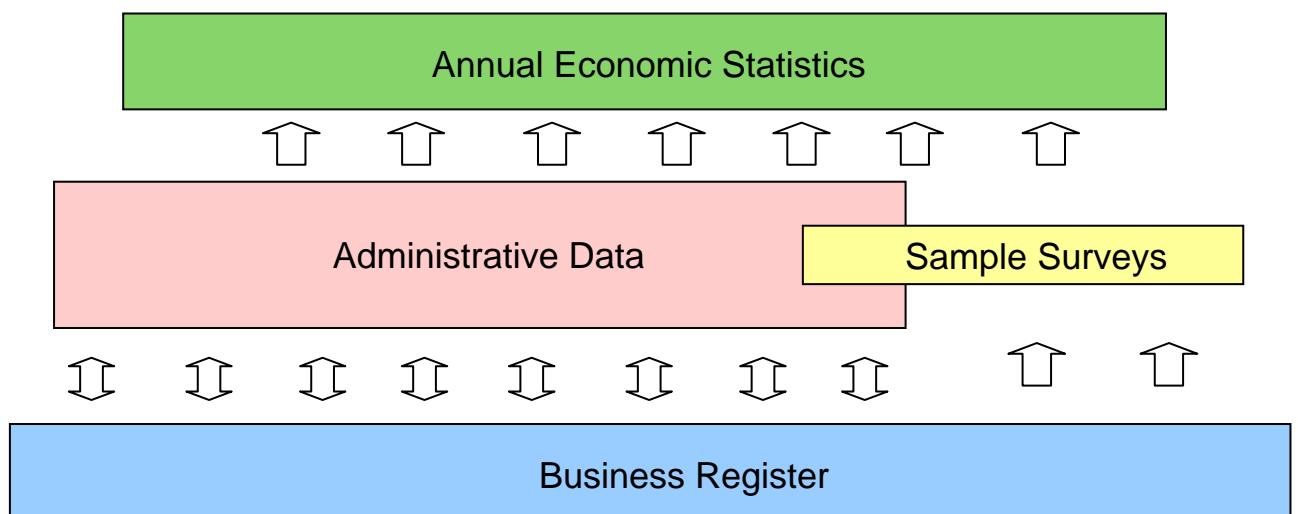
The foundation of the statistical system will be an integrated set of core information about people and businesses. New and existing collections will add information to this core infrastructure in a way that increases the power of statistical analysis. An optimal mix of survey and administrative data should reduce the volume of information that has to be collected from people and businesses, while increasing the range and usefulness of the information produced.

### 2.3 General Principles

The statistical architecture outlined in the remainder of this paper will be shaped by the following general design rules.

1. Information can only be collected if a clear user need has been established.
2. Information should only be collected once.
3. Administrative data will eventually be used as the primary source of data.
4. Surveys will only be used to fill the gaps that cannot be met from administrative sources.
5. Survey and administrative data will be integrated using a comprehensive business register.
6. Large complex business units will be closely managed to facilitate the collection of all the data that is needed from them.
7. Information quality will continue to be fit for purpose.
8. Reliance on administrative data will increase in incremental steps, beginning with the parts of the population for which tax data is robust and then expanding into more difficult areas as data issues are resolved.

Using administrative data first with surveys filling the gaps is a reversal of the current strategy of using surveys for important information with administrative data being used where the contribution is insignificant.



### 2.4 Business Frame to Business Register

A couple of changes will have to be made to the Business Frame to support the increasing emphasis on longitudinal analysis of integrated administrative and

survey data. The BF was originally designed as a sampling frame for economic surveys and for this purpose, coverage of smaller economic insignificant ~~businesses~~ users was less important. Now that the business register is becoming more important for data integration the quality of information about smaller units becomes more important. Longitudinal analysis of the dynamics of business development may need to track businesses right from their first beginnings. The next redevelopment of the BF will resolve this problem by more fully integrating the units that are currently defined as not economically significant.

Administrative churn on the frame creates difficulties for both matching with other datasets and for longitudinal analysis. This problem has been largely resolved by the development of a Longitudinal Business Frame (LBF). This is a different view of the information on the Business Frame that uses information about employees to identify continuous businesses.

Our statistical units model is currently being reviewed. As dependence on administrative data increases, statistical units will need to be much more closely aligned to the units reporting on administrative databases. Statistical units that differ from ~~that~~ administrative units are will only be justified if the benefit outweighs the cost.

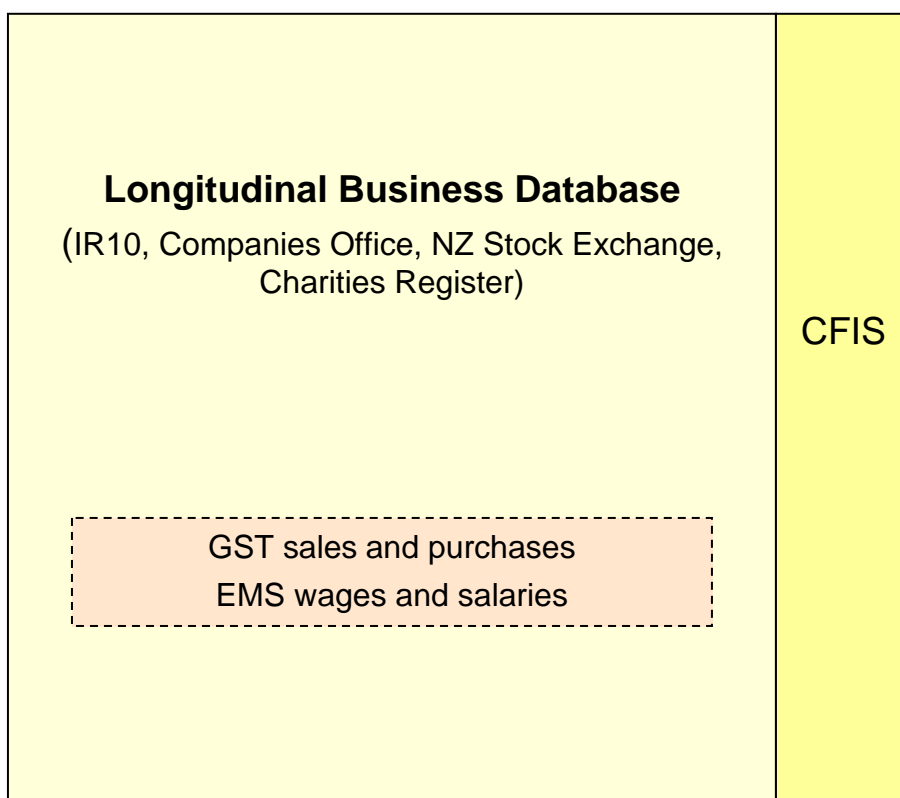
## 2.5 Database Strategy

Increasing demands for more detailed financial information from a broader range of businesses cannot be supported by a sample-based survey strategy. The need will be met by shifting from a sample-based strategy to a database strategy. The heart of this new strategy will be a fully-populated Longitudinal Business Database containing core financial information for every business. Less detailed information will be needed for each firm, but ideally some information should be available for each one. The availability of administrative data makes this a possibility.

The two approaches are very different. The sample-based strategy collects a large number of variables from a limited group of significant units. The database strategy will provide only a few variables, but for all businesses active in the economy. It will be supported by a supplementary database to produce the aggregates of the broad range of variables needed for the National Accounts.

## 2.6 Longitudinal Business Database

This Longitudinal Business Database (LBD) will record key accounting variables from the statements of financial performance and ~~statement of~~ financial position for every business on the business register.



The core variables still need to be defined, but the following might be sufficient for most purposes.

- Sales of Goods and Services
- Total Income
- Exports
- Total Expenses
- Wages and Salaries
- Depreciation
- Net Profit
- Dividends.
- Fixed Assets
- Intangible Assets
- Total Assets
- Total liabilities
- Owners Equity

The long-term aim is to obtain this information for every business in the economy. More variables may be added as data sources improve. The analytical dimension will be provided by classificatory variables from the Business Frame

- Industry
- Institutional Sector
- Region
- Size

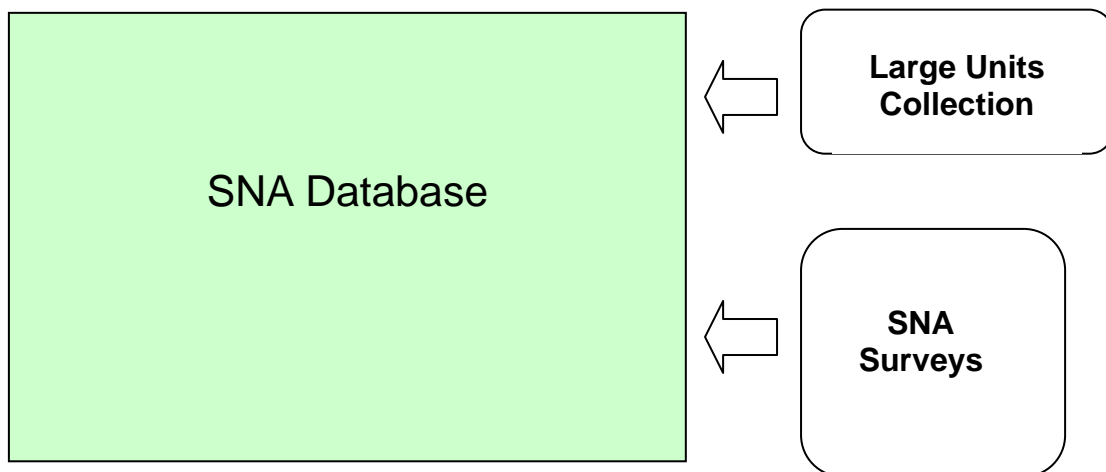
The main data source for the LBD will be the IR10. This form is administered by the Inland Revenue Department and is completed by most businesses. It collects about thirty summary variables from the Statement of Financial Performance and the Statement of Financial Position.

The IR10 will be supplemented with information from the Companies Office, the Crown Financial Information System (CFIS) and the Charities Register. Information for large complex business groups might come from the Large Units Collection described below, rather than IR10s.

### 2.7 SNA Database

The variables held on the Annual Finance Database would not provide all the information needed for deriving all SNA variables, so a separate database will be needed for this purpose. The SNA financial database will hold the additional data needed to derive the variables needed for the national accounts, such as compensation of employees, goods sold on margin and change in inventories.

This additional detail will generally only be material for larger more complex units, so the SNA database will only hold records for units for which we have this additional detail. Therefore, whereas the LBD holds limited information for most units on the register, the SNA database will hold a larger number of variables for a limited number of units. For smaller units, where the additional variables are material at the aggregate level input into the National Accounts, the information on the LBD will be sufficient.



The information held on the SNA database will come from the Large Unit Strategy and the supplementary surveys described below. Most of the information that is required for the central government sector will be available from the Crown Financial Information System. CFIS data would feed into the SNA database and the LBD.

## 2.8 Large Units Collection

The existing processes for large-units will be developed further. About 500 units account for nearly half of New Zealand's economic activity, so a significant share of our efforts should be devoted towards collecting information from these units. 500 units might be too many to manage initially, so we will start with a small number and expand out as the processes are bedded in.

The collection strategy will be tailored to each enterprise using a database approach and drawing on the potential of [standard business reporting technologies, such as](#) XBRL. Units would supply Statistics NZ with a couple of quarterly reports, one from the payroll unit and one from the accounting unit, which will supply most of the ongoing information that we need. A further report could be sent once the annual accounts have been finalised.

The features of large units are.

- Complex structures that limit the usefulness of administrative data, as taxation is often paid at a group level.
- Consolidation issues are important for understanding Balance Sheets.
- Group returns are often used for filing GST and EMS. These filing groups can change over time.

The Large Units Collection will provide all the variables needed for the SNA database. Information from the Large Units Collection will replace administrative data on the Longitudinal Business Database, provided the units are consistent.

## 2.9 Supplementary Surveys

A survey strategy will be developed that focuses on collecting the information that is not available on the LBD or from administrative data sources.

**SNA Survey.** In the immediate future some of the additional detail needed for the national accounts will not be available from administrative data sources or the Large Units Collection. An SNA Survey will target units with additional detail that is material at the aggregate level relevant to the national accounts.

**Regional Breakdowns.** For some units a regional breakdown of assets or depreciation might be needed for an accurate apportionment of regional GDP. For other units, a regional breakdown of sales might be sufficient.

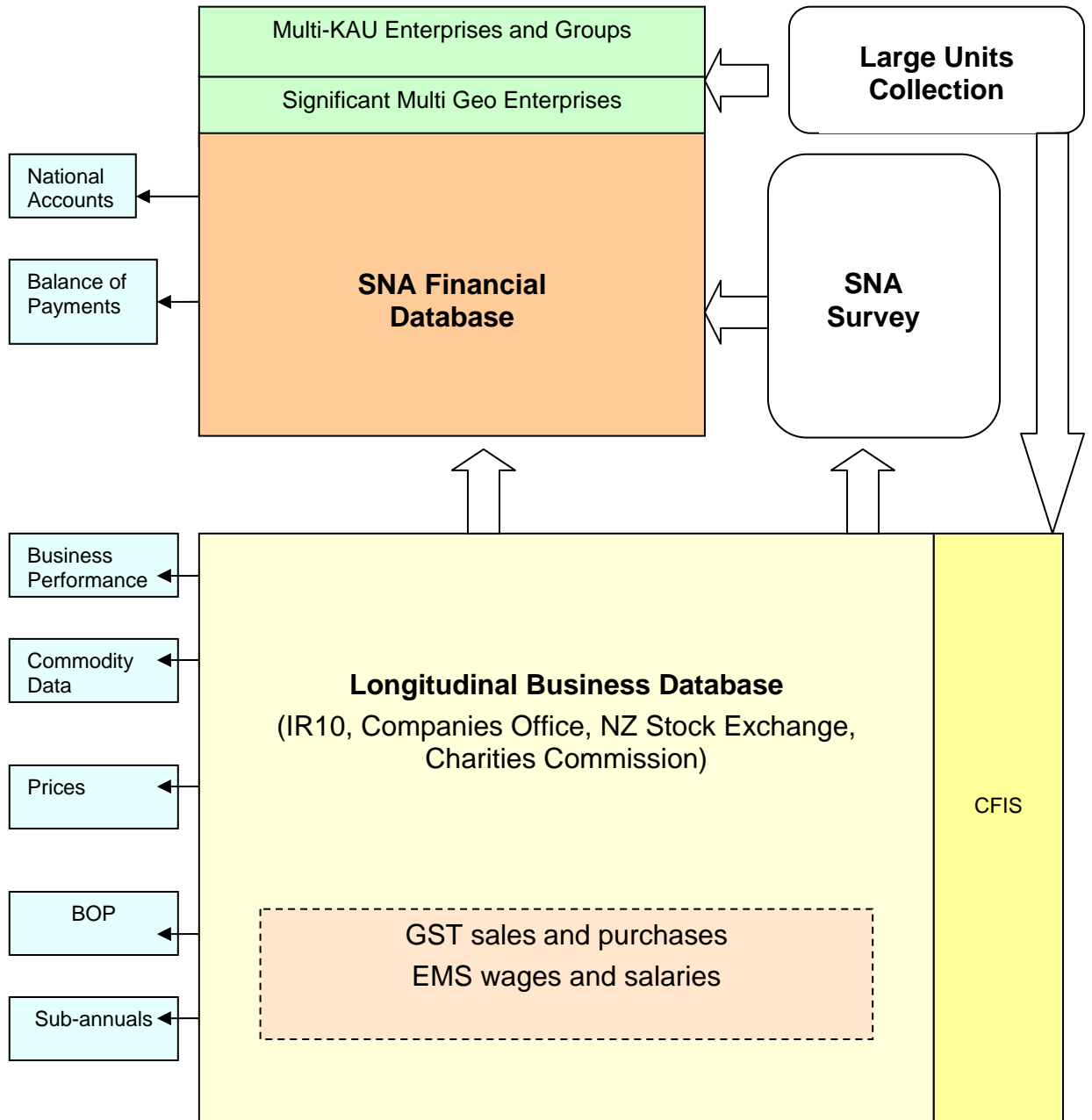
**Business Performance Survey.** Separate surveys will still be needed because this information is not available from administrative sources. They will not collect finance data but draw it from the Longitudinal Business Database.

**Balance of Payment Surveys.** Additional information will need to be collected for input to the Balance of Payments.

**Commodity Breakdowns.** The current approach involves a five yearly cycle for I-O balancing and business price index re-weighting. More frequent surveys may be needed where a market is changing more rapidly.

### 2.10 SNA Estimation

The SNA estimation process will use all the information from both the LBD and the SNA database to produce the best possible estimates. The Longitudinal Business Database will provide control totals for core variables with the SNA database providing the additional detail needed.



### 3 SUB ANNUAL FINANCIAL INFORMATION

#### Needs

Two basic types of sub-annual information are needed.

- Short-term measures
- Indicators

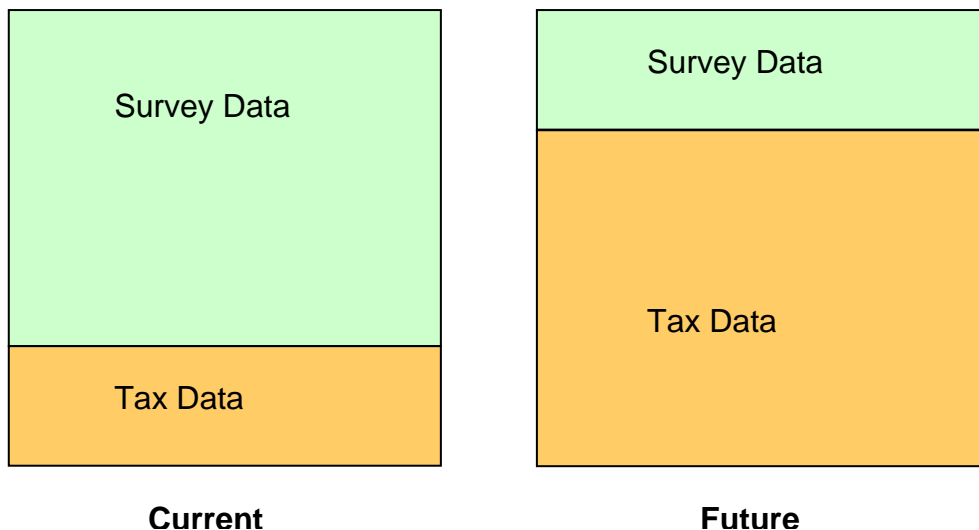
These meet quite different needs.

The requirements for short-term economic measures are largely driven by the design of the quarterly national accounts. Measures are needed for each major component of the various accounts. These measures are also of interest in their own right, provided they are published in advance and are consistent with the quarterly accounts.

Indicators series are designed to give an early warning of turning points. They do not need to measure absolute levels of activity, but should focus on providing decision makers with early warning of changes in the economy before quarterly GDP is published. They must be frequent and timely.

#### Strategy

The current sub-annual method is a mix of postal surveys and Goods and Services Tax (GST) data. This very effective strategy will not be replaced in the foreseeable future. The proportion of sales estimated from GST data is currently limited to 15% of the total for each industry. The immediate challenge is to shift the boundary of the tax strata as high as possible without compromising quality.



The longer-term strategy will be to produce a Quarterly Financial Statistics Database for the entire economy. The core variables would be:

- Income/sales
- Expenditure
- Wages and salaries
- Stock change

The current method of using GST data to estimate sales and expenditure will be expanded to as many units as is practical.

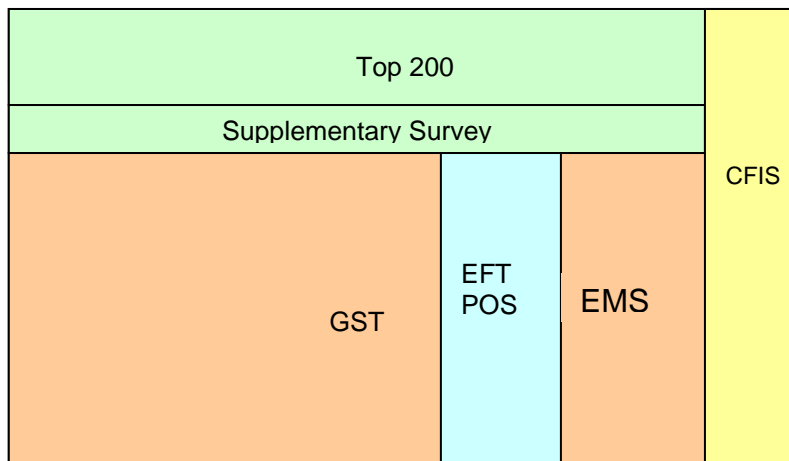
Wages and salaries information will come from the Employer Monthly Schedule (EMS) which collects taxes on wages and salaries

The Large 200 will provide detailed quarterly reports for the more significant and complex units. This information should have concepts and coverage consistent with their annual information. The implementation of this strategy will be incremental, beginning with a few large cooperative businesses and rolled out to others, if it proves to be successful.

Coherence between quarterly and annual measures is really important, so in the long-term, the Quarterly Financial Statistics Database will be built off the Longitudinal Business Database, so that annual and sub-annual measures are comparable and differences can be explained.

EFTPOS data might provide good measures for relevant service industries.

CFIS will provide a reliable measure of Government consumption expenditure.



A supplementary survey program will focus on measuring:

- Sales for units that cannot be estimated from GST or EFTPOS data
- Regional break-downs for large businesses trading in several regions.
- Stock change for units with sufficient stocks to be material to our estimates that can provide accurate information. Respondents will only be asked to provide estimates of stock change, if the information is available from their accounting system.
- Capital purchases for units with large or variable purchases of capital equipment.
- Profit estimates for units whose profit cannot be modelled from GST data.

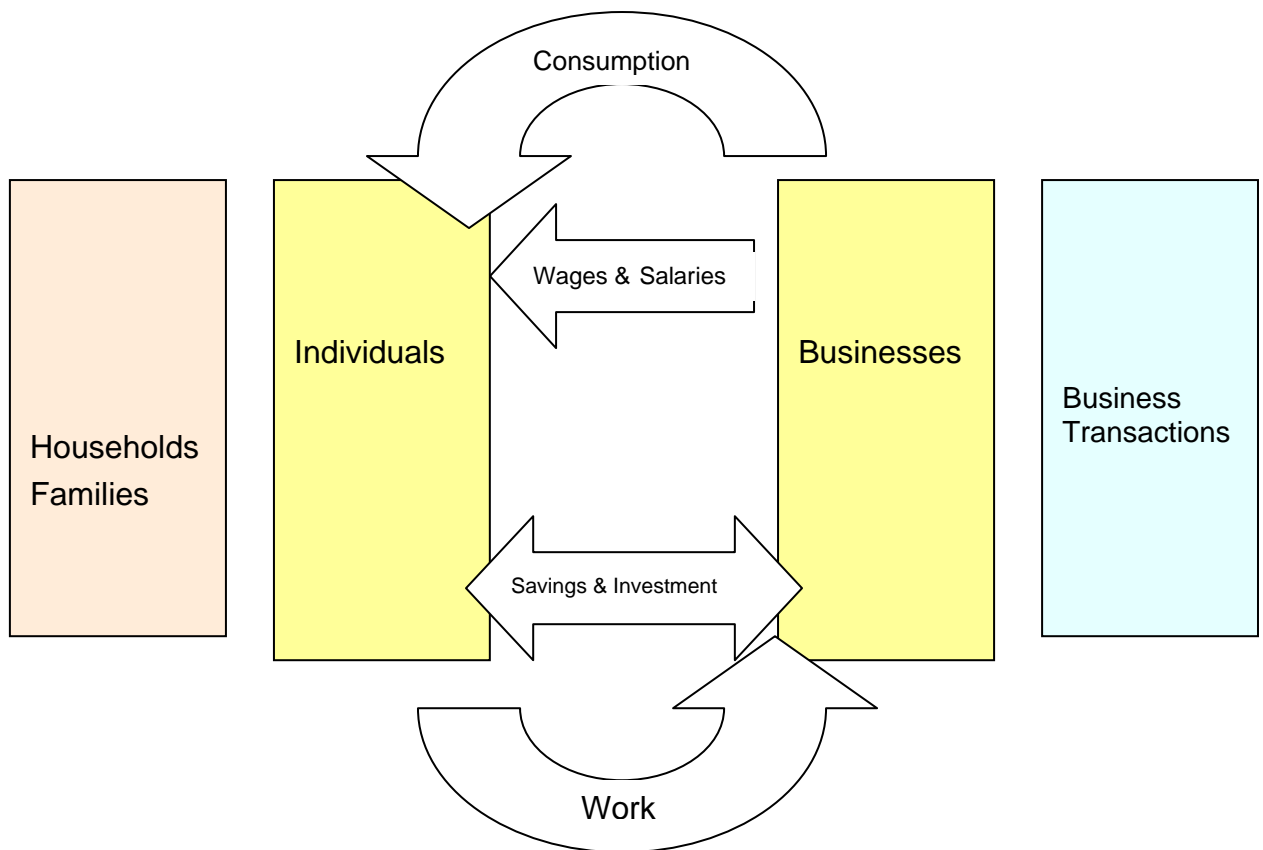
#### 4. STATISTICS ABOUT PEOPLE AND BUSINESSES

Our long term objective is to build up a picture of the entire economy that will provide robust aggregate measures as well as supporting microdata research. Many important transactions are personal so the complete picture must include information about the links between people and businesses. These interactions occur in a variety of ways, including:

- employment in productive activities
- receipt of wages and salaries
- purchase of goods and services
- savings
- business investment

Understanding the impact on individual and household well-being of business activity is an important objective of economic statistics. A secondary goal will be to understand the impact of these decisions on business performance.

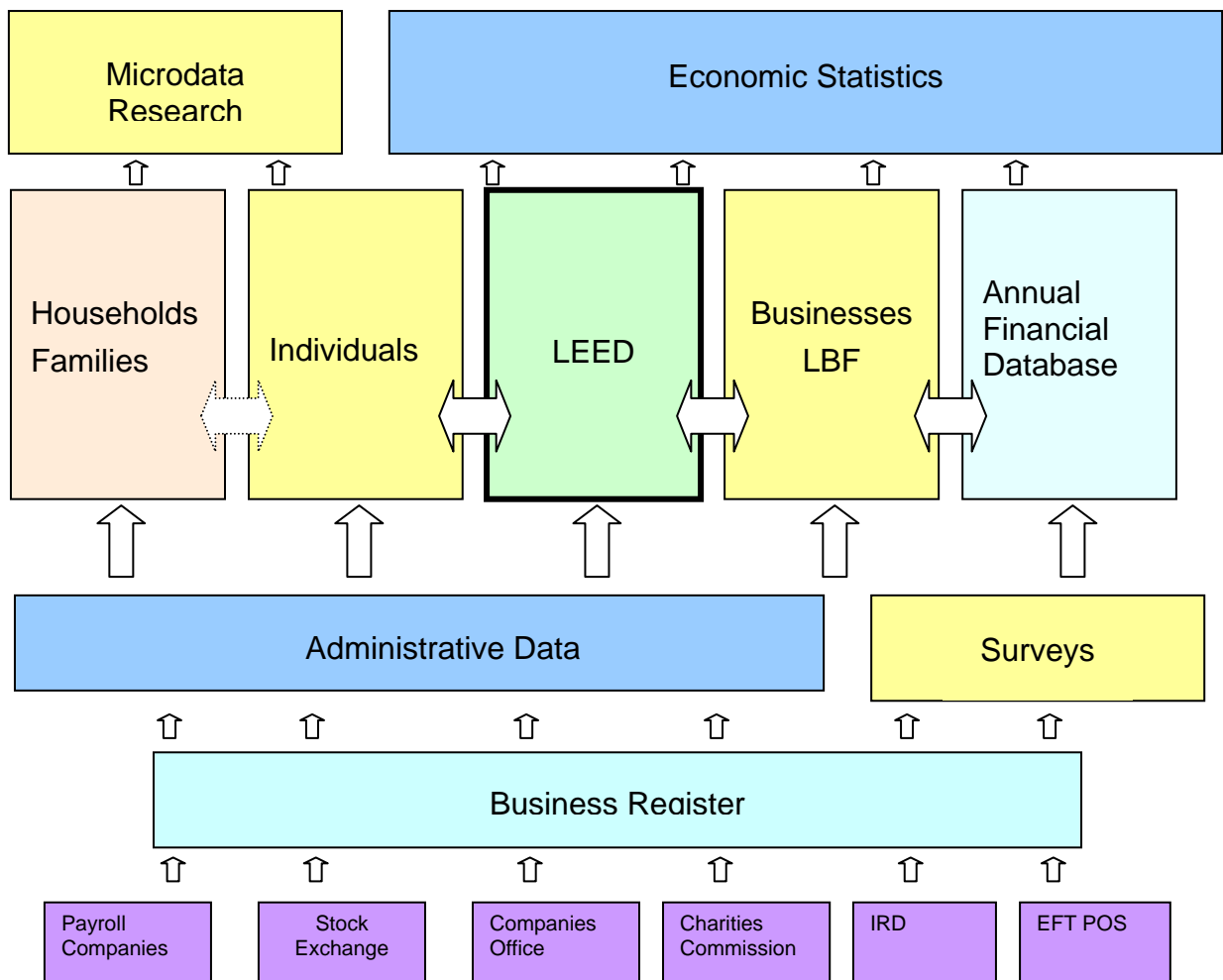
The following chart shows how all the statistics described above will fit together to provide an integrated set of information about people and businesses.



- The core of the planned statistical architecture is the Linked Employee Employer Database (LEED). It uses tax data to establish a link between businesses and individuals. The LEED is the key to integrating business and social measures.

- The Longitudinal Business Database will provide basic financial information about every business in the economy.
- Both the LEED and the LBD link to the LBF providing a link between individuals and businesses.
- The LEED currently only contains information about individuals. In the future, links to their families and perhaps their household may be possible.

This statistical architecture will change the shape of business and financial statistics. New and existing collections will be combined with administrative data to increase the power of statistical analysis. An optimal mix of survey and administrative data should reduce the volume of information that has to be collected from people and businesses, while increasing the range and usefulness of the information produced.



## **5. ROAD MAP TOWARDS NEW ARCHITECTURE**

Implementing the statistical architecture is a long-term exercise. A key feature of our strategy is that decisions about the use of administrative data will be quality driven. The continued quality of core statistical outputs will be ensured by expanding the use on administrative data in incremental steps. Administrative data will be used first in industries and sectors for which the data is known to be robust. Dependence on administrative data will be expanded into other parts of the population as data issues are resolved.

Several projects are currently underway to test whether this approach is practical. The IBULDD project (described in a separate paper) has produced a prototype LBD that has demonstrated that administrative and survey data can be linked together longitudinally to produce coherent statistics.

Five steps are essential for making administrative data usable.

- Linking to the Business Frame
- Quality assessment
- Imputation/modelling of missing records
- Establishing standard units and periodicity
- Combining data sources to model the best record for every unit.

Repeating these processes every time data is used would be very inefficient. Where possible, these processes will be done once, so that usable data can feed into a variety of statistical outputs without further checking of the data.

### **1. Links to Business Frame**

All business-related administrative data received by Statistics NZ is matched to the Business Frame to identify missing records to be identified and ensure units are classified consistently. Linking to the BF allows some parts of the population to be covered by the tax data and others with survey data without fear of double counting or undercounting.

### **2. Assessing Tax Data Quality**

The quality of tax data varies according the degree of checking by IRD. Experience so far indicates that data from tax forms that have revenue implications (GST) are of better quality than those provided as supporting information (IR10).

The aim is to establish standard editing processes for all relevant tax forms. The design of these processes will endeavour to take into account both the longitudinal and cross-sectional dimensions. Understanding the economy is not enhanced, if cross-sectional analysis produces different results from longitudinal analysis.

All new processes for editing tax forms will be implemented on the LBD to ensure that the consistent data is available for other uses.

Editing processes for administrative data will generally be automated, as the volume of data precludes the possibility of manual data edits. Manual edits may be applied to some important units, if resources are available.

Repairs of problem records will have to be automated too, as the volume of data is too large for manual correction. Contacting the taxpayer is not generally an option with administrative data. A quality indicator will be created for each record to indicate its quality status. Sometimes a separate indicator may be needed for each variable or group of variables.

Repaired records will be flagged to indicate the quality of the repair. Records that fail quality checks and which cannot be repaired will be labelled as unusable.

Quality checking processes will be regularly reviewed to determine if they can be further optimised. When opportunities arise, we will work with data suppliers to improve their form design and data checking.

### **3. Imputation of Missing Records**

Missing forms are a problem with the use of administrative data. The rate of missing forms tends to be lower for form types where there are financial penalties for non-compliance. However delays in filing these forms can still cause problems for the use of tax data in the production of statistics.

All imputation and modelling will aim to preserve the longitudinal dimension of the data. In some situations the imputation may be done after the data is reshaped to standard units and periodicity and after data combination and modelling is complete. Where better quality information is available from another source, imputation will not be unnecessary. For example, missing tax data will not be imputed for units which have responded to a postal survey.

### **4. Establishing Standard Units and Periodicity**

Before data from different sources can be used coherently, data from each source will be reshaped to match the standard units and periodicity required for the LBD.

Monthly and quarterly data will be aggregated to the appropriate financial year. Data from administrative units that do not match the SNZ statistical units will be either aggregated or allocated to the appropriate statistical units.

### **5. Combining Data and Modelling**

Most users of financial information from businesses, whether they are producing national statistics or doing micro data research, want the best set of financial accounts that can be obtained from each business unit. In the future, this will not just involve a choice of a data source, but combining data from various sources together to produce the optimal set of financial records for each unit.

Two methods of combining data will be used.

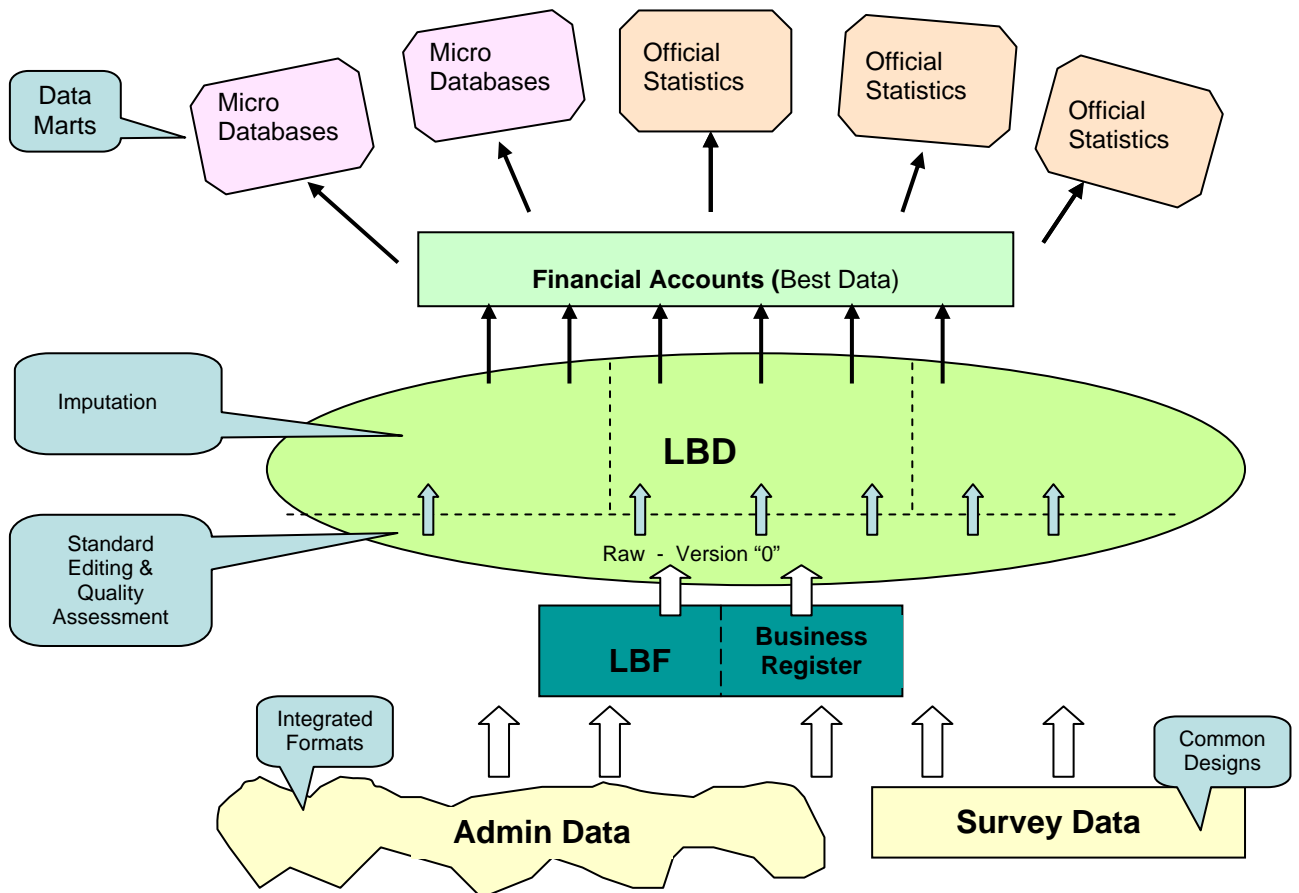
1. In some cases, gaps in the core data source for a particular unit will be filled with information from a different data source. For example, the IR10 accounting summary form from some units record zero wages and salaries, yet their EMS wage and salary form indicates that they have paid wages and salaries. The accounting summary usually balances, so it seems that the wages and salaries are being mixed with other expenditure. The information from the EMS might be used to separate their wages and salaries out from the other expenditure on their IR10.

2. If no IR10 accounting summary is available for a unit, it may be possible to model an accounting summary using the information from other tax records to place bounds on the core variables, such as income, expenditure and wage and salary payments. The extent to which missing variables can be modelled has still to be determined.

The objective will be to produce the best possible set of accounts for the unit, while maintaining the coherence of the accounts within each period and over time.

### The Shape of Longitudinal Business Database

Administrative and survey data will be loaded in a way that allows information from different sources to be easily linked. This would allow the same source data can be used for a variety of purposes by different users. The following diagram describes the shape of the proposed solution.



As the data is loaded, it will be cleaned and edited. Corrected and imputed data will be stored as subsequent versions. A quality indicator will tag the best data source for each group of units and variables. Data may be drawn off into datamarts to support the production of statistics or microdata research.

### Conclusion

This statistical architecture provides a clear direction for future development of economic statistics. The use of administrative data will be expanded incrementally over time until the objectives of this strategy have been achieved.