Using the idea of Statistical Architecture at Statistics New Zealand

Paper presented at the 12th East Asian Statistical Conference,
Tokyo, Japan,
13-15 November 2008

Vince Galvin, Deputy Government Statistician, Standards and Methods;

Natalie Rhodes, Subject Matter Project Manager, Information Management,

Statistics New Zealand

P O Box 2922
Wellington, New Zealand
info@stats.govt.nz

www.stats.govt.nz

Introduction

This paper is in two separate parts. The first describes how the idea of architecture has been used in economic statistics to design a development pathway. The second part describes how analogous work is progressing with our social statistics. These two parts of this paper highlight the difference between these two areas:

- on the economic side the existence of the System of National Accounts framework, our relatively long experience in using tax data and the centrality of the business frame to economic measurement mean that specific plans have been developed and it is possible to articulate the role that administrative data will play in those plans;

- in our social statistics the plans are more in outline and there are more fundamental issues that need to be evaluated before we can describe our pathway forward at the same level of detail. Social statistics also have more diverse requirements and this requires a comparatively wider range of measurement approaches.

This paper looks at the two areas. The economic part of the paper outlines Statistics New Zealand's plans, while in the social area the paper will give more emphasis to the way we have been weighing issues in developing an approach.

In this context the word architecture is being used to describe a set of data collections seen from an overall design point of view. It describes an integrated and systematic approach to the collection and organisation of data that will support current and future information needs. We define the key elements of a statistical architecture as: the target populations and statistical units that define the scope of who and what we wish to measure; the frames used to select samples; the set of individual data sources; certain features of how these data sources are held; the ways in which data sources are combined to enhance the original source or to produce new outputs, and finally the means of access to unit record data and dissemination of outputs to the public. Underlying our ability to produce statistics are legal obligations, Statistics NZ policies and wider political and cultural acceptance of the ways in which data is collected, output and used.

Economic Statistical Architecture

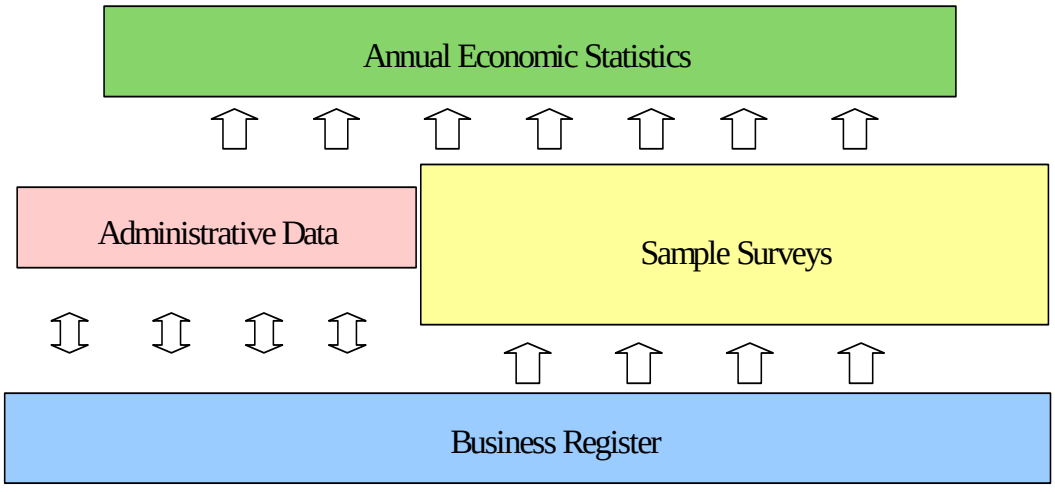
Background of New Zealand's Economic Statistics

The central role of the Business Frame and the Annual Enterprise Survey

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. The foundation of the current 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. 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 into 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.

If the Business Frame is the foundation for economic statistics, an economy-wide economic survey is the superstructure. The Annual Enterprise Survey (AES), which was introduced in 1986, fulfils this role by providing financial information for the entire economy using a mix of survey data and administrative data.



This combination of a comprehensive business register and an economy-wide economic survey provides Statistics NZ with a system of integrated economic statistics.

Emerging Needs and Constraints

Taking New Views of Core Economic Data

Statistics NZ has had to reconsider the range of statistical information needed to support policy development and monitor policy implementation, as the global situation has changed and as Governments are now more prepared to consider intervening in the economy. These policy interests can have two broad impacts on the demands for data.

First is that a sub population that is not identified on the Business Frame may need to be identified. We are used to having to identify firms who trade internationally for Balance of Payments purposes but there is increasing interest in Maori owned businesses, firms who have undertaken research or even firms who have received government support being identified as separate study populations.

Second is on the nature of the data that needs to be collected. There is increasing interest in looking at how firm behaviour has impacted on financial performance. Investment activities, Research and Development, Innovation and the use of strategic management tools have all been the subject of interest. The challenge is that this information needs to be collected from a range of senior sources inside the organisation, and needs to be put alongside detailed financial performance information. This has led us to explore solutions that establish central stores of financial data that can be used with a range of survey data depending on the context.

An example of an issue that needs to be accommodated in our wider strategy is the need to 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, and business investment

Microdata Analysis

The forms of analysis have been diversifying. As policy analysts started to ask more demanding questions about cause and effect relationships, they soon discovered that the answers to many questions are hidden in the details, so the emphasis shifted from analysis of broad aggregates to an increased use of microdata analysis.

Microdata analysis often has a longitudinal basis. For example, understanding how productive firms grow over time is an important aspect of productivity analysis. This has led us to develop solutions that have have a significant component at being able to link actions in one good point in time with financial performance in another point in time.

It is also in the nature of microdata analysis that it is most useful when it can draw on as full a set of explanatory variables as possible. The shift to microdata analysis has consequently been supported by a range of dataset integration projects. One project in particular (we have called it IBULDD - this will be covered later) has brought together a wide range of survey and administrative data and has supported a series of econometric analyses of public policy issues.

Minimising Respondent Load

The New Zealand government has undertaken a series of reviews of a variety of aspects of the way that the Government activity imposes burdens on business. The collection of statistics has always been in the

scope of these reviews and consequently there has been sustained interest over many years in us reducing the burden from business surveys.

There have been two broad aspects to this concern. One has been to minimise the total load of surveying business, the other has been to monitor the distribution of load, especially on small and medium sized business, to ensure that it is not disproportionate.

These concerns have led Statistics NZ embark on a strategy that places a greater dependence on administrative data, with sample surveys being used to fill the information gaps. We have also had to keep monitoring how well our survey overlap control mechanisms are succeeding in spreading the survey load.

Future Directions

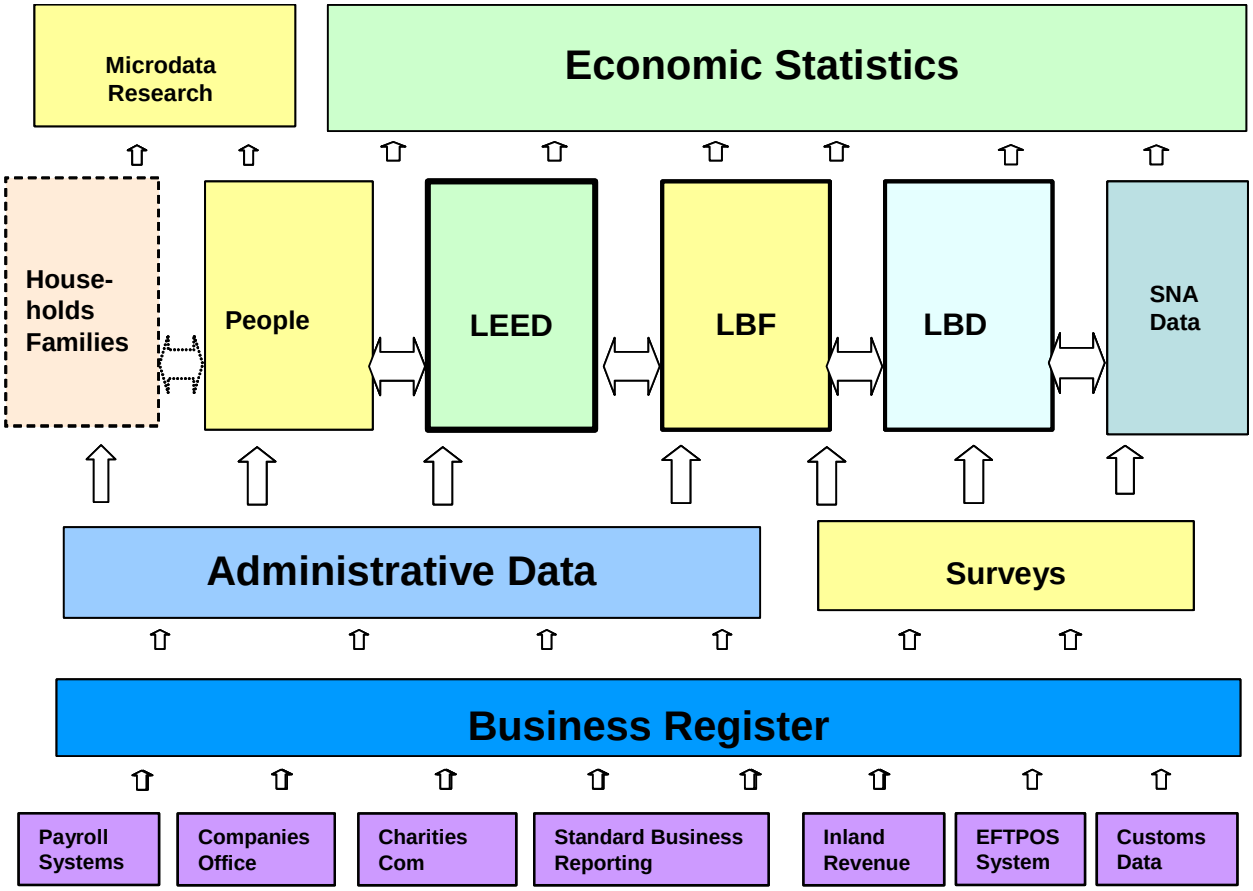
General Design Principles

With this history and these emerging needs in mind the architecture outlined for our economic statistics is shaped by the following general design rules:

- Information can only be collected if a clear user need has been established.
- Information should only be collected once.
- Administrative data will eventually be used as the primary source of data.
- Surveys will only be used to fill the gaps that cannot be met from administrative sources.
- Survey and administrative data will be integrated using a comprehensive business register.
- Information should only be collected from units that can readily provide reliable and meaningful estimates.
- Large complex business units will be closely managed to facilitate the collection of all the data that is needed from them.
- Information quality will continue to be fit for purpose.
- 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.

The foundation of the future 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 supports a wider range of statistical analysis. Statistical outputs will not be based on a single survey, but will be compiled by combining different data from several different sources.

The roadmap for this strategy is defined in the diagram on the next page.



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.

In very broad terms this picture consists of a set of databases and plans to populate them with administrative data and sample surveys. The next section describes the nature of the key elements of statistical architecture and then the following section looks at how the use of this infrastructure will modify our strategies to acquire data.

Key Infrastructure Elements

- 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 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 administrative units are only justified if the benefit outweighs the cost.

- The Longitudinal Business Data Base

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 database will provide fewer variables than a sample-based strategy, 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.

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.

This Longitudinal Business Database (LBD) will record key accounting variables from the statements of financial performance and financial position for every business on the business register. The long term aim is to obtain this key information for every business in the economy. If Standard Business Reporting is implemented the range of variables may expanded significantly. The analytical dimension will be provided by classificatory variables from the Business Frame. The longitudinal database being developed by one of Statistics NZ's recent projects, Improved Business Understanding via Longitudinal Database Development (IBULDD), is a prototype of the LBD. The IBULDD project was a feasibility study with four main objectives:

- to test the feasibility of creating a complete business data collection with longitudinal depth based on linking existing survey and business administrative data;
- to establish where imputation can be used and what methods will be deployed to fill the gaps;
- to identify new official statistics and potential improvements for current official statistics; and
- to conduct research that provides relevant information on business dynamics specifically relating to firm productivity growth, financial constraints, economic development and research and development investment decisions.

The IBULDD project has successfully proven that we can gain more value from our data by maximising the potential of existing administrative data sources. Increasingly, the database is being utilised to produce a variety of research papers, however to retain this value, more years of data must be added.

It is clear from this study that the best way to provide the key accounting variables needed for a full coverage unit record database is to use administrative data sources. The main data source for the LBD will initially be from a form administered by the Inland Revenue Department (the IR10) which is completed by most businesses. It collects about thirty summary variables from the Statement of Financial

Performance and the Statement of Financial Position. This 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. We have done considerable work on the LBD idea and further details are given in appendix A.

- SNA database

The variables held on the Longitudinal Business 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.

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.

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.

- Linking data on people and businesses

A core element of the future 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 scope of possible statistical analyses. The core of the planned statistical architecture is Statistics NZ's Linked Employee Employer Database (LEED). This database uses tax data to establish a link between businesses and individuals. The LEED is the key to integrating business and social measures.

Redesigning the way we acquire data

Key information requirements

Our key information requirements fall into four areas:

- **Annual information** - to provide a basis for understanding the structure of the economy.
- **Sub-annual information** - to provide early indicators of economic turning points and to feed into quarterly national accounts measures
- **Relationships between people and businesses** – to provide a complete picture of interactions between people and businesses
- **Interactions with the rest of the world** – to provide information on the flows of goods, services, revenue, capital and ownership.

Strategy

Some specific aspects of our collection strategy that we are developing are;

- 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. The best approach would be to start with a few cooperative businesses and develop an efficient process that could be rolled out gradually to other larger 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.

Large units are complex structures that limit the usefulness of administrative data, as taxation is often paid at a group level. The collection strategy will be tailored to each enterprise using a database approach. Statistics NZ will work with the enterprise to develop the required business reports. 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.

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

- Standard Business Reporting

A project to develop Standard Business Reporting (SBR) is currently underway. This project will build tools into commercial accounting packages that can send accounting data to government agencies. If this project is successful, Statistics NZ will receive detailed accounting information from all participating businesses. This data will be timelier and more detailed than the data currently received from Inland Revenue. When businesses switch to reporting with SBR, their Inland Revenue information on the LBD database will be replaced with SBR data.

The introduction of SBR may allow the number of variables recorded on the LBD to be expanded. The marginal cost of obtaining an extra variable using SBR is quite low.

- Supplementary surveys

The annual statistics are the main collections that will draw on the infrastructure outlined above. The idea is that we will use a range of sources to build a comprehensive view of the financial performance of data and then we will run a range of supplementary surveys to collect information that doesn't reside in accounting systems. The sort of information will include;

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. (When SBR is operational the scope of these surveys may be reduced).

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. (An SBR process might eventually be developed for collecting commodity information).

- A sub-annual collection 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 increase this proportion without compromising quality.

The longer-term strategy will be to produce a Quarterly Financial Statistics Database for the entire economy, with the core variables being: income/sales; expenditure; wages and salaries; and stock change. The current method of using GST data to estimate sales and expenditure should be expanded to as many units as is practical. SBR should gradually replace GST as a source of quarterly data.

The other elements of our subannual data acquisition strategy that are being evaluated are;

- Wages and salaries could be estimated for all other units from other tax data
- The Large Units database will provide detailed quarterly reports for the more significant and complex units.
- EFTPOS data might provide good measures of a range of industries as well as providing estimates of household consumption expenditure. EFTPOS data on trading day and holiday effects will also improve seasonal adjustment.
- A supplementary survey program will focus on measuring other areas such as:
    - Sales for units that cannot be estimated from GST or EFTPOS.
    - 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.
    - Capital purchases for units with large or variable purchases of capital equipment.
    - Profit estimates for units whose profit cannot be modelled from GST data.

- Interactions with the rest of the world

The challenge is to integrate information that already exits about interactions with the rest of the world into the economic statistics measurement system. There are a range of challenges in this regard;

- The experience of migrants needs to be integrated from a range of sources.Short term travel is captured in migration and tourism statistics while Longer term migration is captured in migration statistics. Labour force surveys will collect information about the skills and employment experience of migrants.
- Merchandise trade information is captured from the customs database. This information would be more useful if a link to the Business Register could be maintained. This would enable a link to the LBD to support studies that measure the impact of trade and overseas ownership on profitability, productivity and growth. There is growing interest among economic researchers in the relationship between exporting and business development.

Some international financial flows are captured in administrative datasets. However, most of the information needed for the Balance of Payments will be captured in special surveys of units that engage in international activity. Once SBR is operational, the need for special surveys should decline.

Understanding the impact of overseas ownership and globalisation of business decisions and activities will become more important in the future. The Business Frame records information about overseas ownership for the relevant businesses

- Quality assurance of Administrative data

The use of administrative data is a key element of the strategy and we are working to try and ensure that we generalise the lessons we have learnt about evaluating data sources and employing them effectively. Our framework can be briefly outlined as:

- 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. These comments are expanded on in Appendix B.

We have made extensive use of tax data in our measurement of the economy over a period of about 15 years. As we have understood the particular features of the various tax data sources, we have used it variously as a source of data, a source of size measures in stratification, weighting information and the basis for modelling of different types. The increased range of data available to us has enabled us to progressively improve our business sampling processes.

Key Issues Being Evaluated

While we have a great deal of clarity about what we would like to build we do have some feasibility evaluations to do around establishing the databases that have been described. The Longitudinal data base, is the central feature of this system and will be the main focus of our development work.

Outside of this we will be working hard at ensuring that we develop an optimal mix of making full use of administrative data in the design and operation of surveys, alongside using the administrative data as the data source. This process is one we began some time ago. We have been making a careful transition towards making greater use of administraive data, mindful that we have to learn progressively, both about the features of the administrative sources and our capacity to deal with changes in those sources that arise over time.

The prospect of our Standard Business Reporting developing provides another source to weigh in the balance. Potentially, this offers a richer data source with less burden on respondents but a reasonable expectation seems to be that there will be variable take up of this option, and that there will be a set of complexities around getting to understand the diversity of this data source.

Economic Statistics Concluding Observations

This statistical architecture sets out an integrated and systematic approach to data collection that will support current and future information needs. It 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.

The main challenge in front of us is to get the full richness of data held in businesses reflected in economic statistics. This has started with the work done on using administrative data and will continue when SBR provides more options.

Social Statistical Architecture

Background of New Zealand's Social Statistics

New Zealand's social statistical system has historically been relatively decentralised. Large surveys were funded as individual projects separately in a wide range of agencies. This created fragmentation and users had no certainty, other than in the short term, about the data that was going to be available.

The social statistics data collections at Statistics NZ are based on a traditional model of household surveys supported by a range of administrative data sources. The five-yearly census of population and dwellings is a cornerstone of official social statistics in New Zealand, providing the source of information about NZ's population for small areas and small population groups over a range of variables.

Following a review of the co-ordination of statistical work in New Zealand, Statistics New Zealand has been funded to introduce the Programme of Official Social Statistics (POSS), enabling us to work towards a more coordinated approach to social statistics across government. Launched in 2005/06 the POSS was established to provide a coherent system of official social statistics across the government sector. It mapped out a ten-year plan of work involving a number of government agencies, led by Statistics NZ. The key elements of the POSS work programme are:

- the consolidation of existing surveys into a managed programme
- the introduction of new surveys to fill information gaps
- exploitation of other sources of data such as administrative databases
- the improvement of analytical capability, dissemination of information and access to data.

POSS is made up of four Programme Development Plans containing statistical development work on:

- statistical infrastructure
- administrative data and data integration
- census and surveys
- analysis and dissemination.

Improved coherence will be assisted by the development of 12 domain plans, which will link information needs to the priorities for statistical development work. These domains comprise of:

- population
- housing
- safety and security
- economic standard of living
- knowledge and skills
- health
- paid work
- culture and identity
- social connectedness
- human rights
- physical environment
- leisure and recreation

A ten year programme of POSS surveys to fill the information gaps that were identified in conjunction with a range of agencies are currently being developed.

Accumulating and Evaluating Key Emerging Requirements

In the current New Zealand context the key emerging needs can be characterised by

- The requirement for detailed information about sub-groups, especially regional data and the Maori. This data is necessary to enable understanding of the social differences across regions, communities and various other population groups. It will inform the creation, implementation and evaluation of policies to effectively reduce disparities. However as well as issues around overburdening respondents, information can be difficult to collect efficiently from these groups.

- The growing need for information about life cycle transitions and cause and effect relationships, means that longitudinal data is essential. We had a long history in New Zealand of trying to understand social change by making inferences from successive cross sectional surveys, and it produced the expected frustration. When the key policy interest is in how people are making transitions, for example, from education into the labour market doing anything other than directly observing the transitions can be very limiting.

- A desire to examine people's circumstances across policy domains. Increasing effort is being made to trace the total picture of people's needs and assistance from government and understand where connections between policy areas matter. An example has been the concern in policy circles about understanding all the factors that impact on Youth Potential. From a statistical perspective this involves taking estimates from general surveys (such as unemployment) and using them alongside very specific administrative data sources (on issues like taking children into the care of the state). The practical difficulties in keeping these estimates referring to a consistently defined population are considerable, and trying to understand the links across data sources even more challenging.

- Needs arising from newer domains (social connectedness, culture and identity and human rights) might prove to be the most demanding. They are characterised by both nascent concepts and complex units of analysis. These topics push us towards new methodologies such as, for example, constructing families (or other across networks) across households.

Looking at the range of requirements that have been identified the hardest question to answer is deciding what is a "must have" requirement and what is a "desirable direction" that we can make progress towards as opportunities emerge. As a matter of pure practicality any system needs to have a set of minimum conditions it has to meet, and it is difficult enough to trade these sort of objectives off within an individual survey let alone over a whole collection of statistical work.

Constraints

As we develop our approach there are a number of constraints on our approach;

- The Impact of a small population

It is worth noting that, although not unique in this respect, New Zealand has a particular set of challenges due to its population size. These challenges present themselves in the form of achieving representative samples from small-population areas or social groups while ensuring that respondents are not overburdened, and the costs are balanced with the value of the survey. The relative proportion of operational costs to population size is much higher in New Zealand than in a lot of other countries, so this makes the case for finding ways to make efficient and effective use of administrative data all the more

imperative. Confidentiality concerns are also more acute - some of the communities requesting planning information are so small that it is difficult to meet normal confidentiality protection guidelines.

- The difficultly of identifying threshold quality issues.

In the process of publishing the 2006 Census results, the data confrontation work done, highlighted that New Zealand seemed to be losing a relatively large proportion of young men. The subsequent more detailed analysis highlighted that the relative error in the data the Census was being confronted with was underestimated, creating some doubt about validity of the estimate of the gap. The general point is that we need to ensure that we identify whether any of these incoherencies in our estimates justifies more investigation and possibly new collection approaches.

- Minimising respondent load across all surveys funded by the NZ Government.

Potentially this is a significant issue for some regions and the Maori population where we estimate that sampling fractions could be very high. Even if we are not able to negotiate ways of spreading the load it is desirable to be able to evaluate how the total burden of government funded surveys compared with the burden of private surveying, phone charities and other types of approaches to households that are seen as burdensome.

- Privacy related constraints

The scope of data linking that will prove to be publicly acceptable is yet to be tested. In New Zealand there is no common personal identification number across government. This restricts the range of direct linking that can be done between administrative data sources.

The privacy legislation in New Zealand provides for some exemptions for statistical purposes but the most important judgement remains around how to draw a line about what use of information about individuals should be deemed acceptable. The most obvious " threshold" issue for us is around the role of the Census. We have not had any adverse reaction to the work we have done so far (explained below) but it has yet to be the discussion of any high profile public debate.

- Capability and Resourcing

An indirect consequence of New Zealand's small size is that the pool of potential researchers is relatively small, and the financial pressures of running surveys that are large enough to be useful are considerable. The greatest challenge that arises out of this is to ensure that the value of available data is maximised. It also provides healthy challenges around trying to ensure that methods and systems are designed to be "fit for purpose".

What have our main achievements been?

- Using our existing methods to ensure we produce consistent estimates from separate surveys

The core questions initiative has provided us with a useful launch pad for managing common content across surveys. The identification of a set of core demographic questions to be used in all our surveys will provide a useful tool for ensuring the coherency of comparisons across surveys and there is enthusiasm from other agencies to adopt this approach.

We also have a well developed system of maintaining statistical classifications and our collection and statistical methods are well standardised across our surveys. This has left us in a good position to develop coherent statistics from a range of separate Surveys.

- Integrating administrative data sources has given us a good understanding about what can be achieved.

The LEED work (mentioned in the economics statistics section – page 6) has been very successful. It has taken in data from existing sources, has provided a platform for integrating further information and provided a source from which to publish regular outputs. In addition to this, there have been a range of integration projects that have worked to varying degrees that have all provided useful insights. We have gained a considerable amount of expertise in the technical aspects of this work, and have been able to identify some general principles that can be applied to new files in new areas. We feel we are getting a better sense of what to expect as we look at the potential role of administrative data in our overall system. Generally this is very positive but we have learned that there are no short cuts in learning all the idiosyncrasies of a new data source.

The LEED data set will be the main focus of our detailed work looking at how we might expand our knowledge of the labour market, and look at the scope for linking this data to our survey and administrative sources. While this will help establish the conditions under which we might further our wider picture, it will not give any sense of where there might be significant need or opportunity to conduct other linking work. As we construct our domain plans we need to gather together information about needs and assess unrealised potential in other administrative data sets.

Our integration work has successfully linked data across sectors: LEED employment outcomes to Education; and LEED employment outcomes to income assistance receipt details. We have a concordance between student ID number and personal tax number, so we have the potential to see all three sector outcomes together.

- The use of the Census data in Integration projects has been very successful

The New Zealand Census-Mortality Study was the first major data integration exercise carried out at Statistics NZ. Its aim was to measure mortality differences by socio-economic status in New Zealand. Linkages between the deaths register and census have been established for censuses from 1981 through to 2001. Similarly the Cancer Trends project creates links between census and the cancer register. New cancer diagnoses in the years following census day and up to the next census are linked back to the census record of the individual. A smaller data integration exercise matching birth registrations to census, and deaths registrations to census was conducted to assess the consistency of ethnic responses and Maori Descent responses between census and the two registers.

We have not had any adverse reaction to matching our Census data to administrative sources, and the resulting files have produced widely reported results. The demographic detail provided by the Census both corrected the personal characteristic information on the administrative file and enabled a more credible set of analysis to be undertaken. The result of all this work was that a data source that had been seen as inconsistent with other information is now regarded as a valuable source of knowledge. This project has confirmed both that the Census can be matched in some circumstances and that doing so can add valuable context to administrative information.

- The Official Statistics System (OSS) structure provides a focus for working with other agencies

The agreements around the establishment of New Zealand's OSS provides a framework for discussing cooperation with other agencies. Most importantly it has provided us with the opportunity to work with other agencies in developing the statement of future statistical priorities (the domain plans mentioned in the introduction). This work has gone well. Individual plans are time consuming to produce and discuss but they are a useful approach to sifting big issues from small ones and they do provide early indications of some of the more demanding statistical needs that are emerging.

This has also “trickled down” to more specific technical issues. We have had some encouraging messages from other government agencies regarding the sharing of sampling frames in order to reduce

respondent burden. It has also provided the opportunity to work with other agencies on topics such as how to sample Maori efficiently.

What could the response to our requirements look like?

In order to meet ongoing and increasing demand, we need to look towards a solution which would bring survey data together with administrative data, using more sophisticated statistical methodology. Working towards this picture will have legal, policy, IT infrastructure and processing systems implications.The picture being experimented with starts from the perspective that, in broad terms, surveys provide rich content for a representative subset of the population, while administrative data provides full population coverage and some detailed information about some specific lifecycle transitions. The Census has a very special role in this picture; it has rich demographic detail on the full population, little content coverage and some significant existing constraints on our capacity to use it in linked unit record files.

Our vision for the future of our social statistics system is not as clear-cut as for economic statistics. The questions we have been asking ourselves have been around managing the units of collection and analysis, and around managing content across data sources. To expand on this:

- In countries with population registers, the idea of a Social Statistics Database (SSD) as described in Statistics Netherlands paper "Towards an integrated Statistical System at Statistics Netherlands, N. Heerschap & L. Willenborg" is attractive as an approach to exploiting overlap between admin files and sample surveys. In New Zealand there is no population register but there are a series of identifiers that are used across administrative systems within a sector. Many of these sectors run their own surveys as well so we could look at our data collection system as a series of SSDs. In this system we would be very active in promoting common standards across these SSDs to enable them to be used usefully together at more aggregated levels. Would there be any unit record linkage across SSDs? It is possible that the Census could play a role in adding detail to the administrative data within each SSD. Statistics New Zealand would run surveys that collected content that crossed sectorial domains, and some of these surveys might ask permission to use identifiers from within each SSD so that these surveys could be matched to administrative sources. The Census and the various administrative sources might be used to add extra variable richness to sample surveys.
- On the content side, the question we have been asking ourselves is whether there are significant overarching concepts that might be useful to try and create out of a range of data sources? For example, could we look at creating information on human capital by combining education, labour force participation and health data in one enhanced administrative data file that could be potentially be linked to sample surveys as a set of core information. This might be the social statistics equivalent of the Longitudinal Business Database in the economic surveys. From a design point of view it would be valuable in determining what data needed to be brought together in one file to meet a clearly defined purpose.

Should it be possible to construct such a system, it would provide the mechanisms to answer some of the needs we have identified above;

- The combination of Census and administrative files would be the basis for examining the needs of specific sub-populations
- The need to obtain more holistic views of New Zealanders circumstances would be met out of adding various administrative sources onto sample surveys
- The whole system would facilitate the confrontation of Census, sample surveys and administrative sources, giving us the best possible view of coverage issues, and the possible ways ahead in terms of identifying new approaches to acquiring data on difficult to reach sub-populations.

- From an efficiency point of view this sort of system would offer considerable scope to rationalise which blocks of questions we asked in which surveys.

Future directions and issues

Maximising clarity around information needs

- Accepting a sectorial approach to identifying information need

While the above sections outlined potential complexities in balancing requirements, it also clearly points in the direction of getting as much information about these requirements identified as clearly as possible. In doing this we will progress domain plans sector by sector. We have been through a process apparently similar to that recorded by Statistics Canada. Essentially we have also found that there is so much complexity and diversity in each separate topic in social statistics that operating within sectorial areas of concern is a practical compromise that enables progress to be made.

- Identifying the forms of analysis that we are aiming to support

The essential information in this process is the clear identification of how analysts will use data to answer policy questions. Establishing where there is a clear understanding of what types of analysis is being planned is the first step towards thinking about the sorts of benefits that might arise from creating new “rich” unit record files. Our experience suggests that existing data sources remain under-utilised, so it would seem important to ensure that work does not begin on building data resources that are at significant risk of not being used.

- Monitoring current data use.

It seems almost too obvious to point out, but our experience is that it can be surprisingly easy not to capture information about exactly what analysis is currently done and trace the use of this analysis in subsequent decision processes. Similarly, it is important to understand where public policy is made without reference to available data and establish whether this was a conscious choice. Understanding all the non-technical barriers to data usage is essential to ensuring that the potential of data to inform debate is realised.

Assessing potential efficiency issues

- Making the best use of our areal approach to household survey sampling

We will examine the possibilities of what has been called an integrated system of survey designs. The idea we will be exploring is how to regard all the households we visit as one "master sample" and look at the best way to allocate "contact time" with respondents over the full range of topics we want to ask. The possible gain being that we may be able to ensure different sample sizes for different blocks of questions because we would be effectively running them over two collections. Our initial thoughts are that this approach may provide some useful flexibility but it does not look likely to provide any order of magnitude improvements to what we do.

This assumes we keep the same approach to defining our Primary Sampling Units. We are looking at how we can use Geospatial information sources and operational research techniques to optimise the way we define enumerator work loads in the Census. The initial signs from this work are promising

and it has highlighted that we could usefully do more work looking at the way we define and use our finest level of geography and how this affects the cost of our household collections.

- Evaluating whether we can construct a frame that improves the information available at selection time

A recent initiative within New Zealand to attempt to construct a National Address Register (NAR) that we could use as a basis of sampling has ended without a facility being created. We operate our areal frame by selecting areas and using an administrative source (House Valuations) to get a preliminary list of addresses that we update when we go to enumerate. This is relatively efficient in that enumeration costs are well contained but it does not give us the option of using a mode of data collection other than face-to-face visits. The question of whether we could create a NAR that facilitates "multi modal drop off (or initial approach)" in our household surveys is one we will look to evaluate now. There is considerable impetus to pursue this work. As things stand, we have a cost structure for Census and Household surveys that we might struggle to keep funded. It may even be that we could get better coverage if we could use different modes. This appears to be challenging, as there is a wide variety of potential sources but it is not at all clear how they could be evaluated against each other.

Evaluating our key "directional" issues

- Determine the role of the Census

The picture identified above implies shifting the current use of the Census as a source of demographic data to enhance individual administrative files to making it the centre of a system enriching a wide range of administrative sources. The notion that we use Census files to enrich sensitive administrative sources contrasts with the current messages used in our Census publicity about names and addresses not being retained in an electronic form. This can be changed in the future, but the question that lies before us is how we explore whether there is a notion of "acceptable" use in the general public's mind.

A large part of the perceived benefit of using the Census in this way will come from being able to decide whether the Census can play the role of the person register in the type of system described above. Without a source like this it is hard to see how the benefits of bringing data from different sectors together can be achieved. Certainly this remains a point of contrast with the economic collections where the Longitudinal Business Frame can play this role. This provides sufficient incentive to encourage us to undertake further exploration of the role of the Census.

- Evaluate the feasibility of linking Survey and administrative data sources

We have not had the experience of other countries in asking for permission to use unique identifiers from different administrative systems in our survey questionnaires. We are aware that this has been successful in other countries, and we are hopeful that the lack of one unique personal identifier in New Zealand will mean people will be less anxious about individual requests. Our initial experiences in our main longitudinal survey have been very encouraging. We have been looking at how we will rationalise the collection of income data, so this exercise looks like it will provide us with a specific project in which we can trial these activities.

Social Statistics Concluding Observations

In the context of social statistics, the idea of an architecture is helpful but it has to be simple, highlight some really significant changes, their implications for our work and their potential benefits. It can then be used to highlight the key areas of potential change, and what investigations need to be undertaken.The

interesting question is how firm a view of the possible shape of this architecture you need to have, in order to start investigating. It will be interesting as we go forward to see the extent to which issues turn out to be generic across collections and to what extent they are very individual to each administrative source. At first consideration, many of the issues look common but the experience of using administrative data is that it is individual issues that prove challenging.

The main challenge we face at the moment is making good choices at the margin. The cumulative assessment of these experiences is that a more ambitious programme is possible but that there is a sequence of work that needs to be done to test potential barriers. We are not expecting this to be a quick process, so we will be organising the work so that we can get significant benefits out of every major piece of work. In a similar vein we need to keep demonstrating the value of what we are delivering. Which raises the interesting question of whether we are proving able to provide insights into change that are helping decision makers - what data usage do we actively monitor?

In the mean time we will have to keep looking at how we can improve the information available to us so that we can achieve efficiencies in the data collections we carry out. We must keep working with other NZ agencies to support common approaches to ensure that the main drivers of coherence are supported. We must also maintain close enough links with our specialist user community that we understand the value they are getting out of current data and where there might be new opportunities to bring administraive data sources within the scope of the wider measurement system.

The contrast with economic statistics has been instructive. The relative ease of using the idea of an architecture to develop an approach and identify the architecture needed to support it in the economic area initially raised our hopes about the being able to do the similar work in Social Statistics. However, the greater richness in the data frameworks, the greater complexity of the data requirements, the wider range of administrative sources and the greater sensitivities to the risk of this work being seen as programme evaluation quickly led us to temper these expectations. However it is still important to try to get a picture of how social data will fit together. A complete grand plan may yet turn out to be unfeasible but there are key requirements to use data from multiple sources together in policy development and these relationships between sources need to be designed as far as possible.

Final Observations

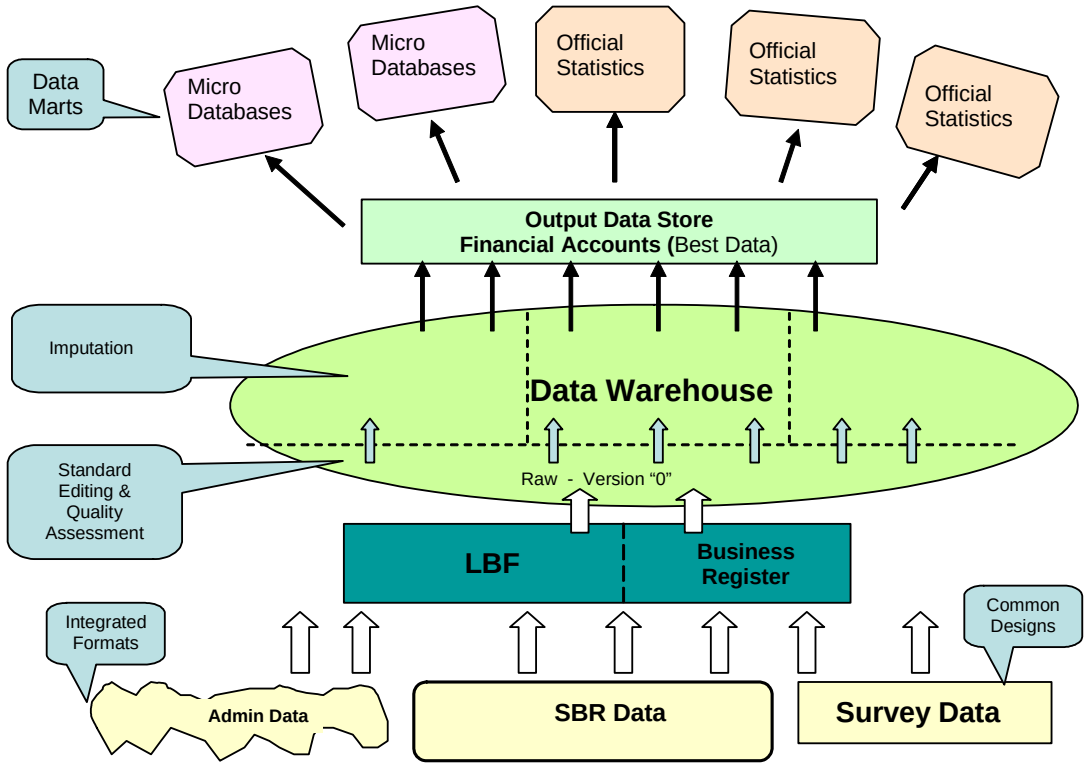
In doing this work we have been looking at the benefits of taking an overview of our survey work, and looking at how a broader view enables us to introduce data from other sources and make a wider range of uses possible. However there are still balances to be struck. Working towards infrastructure that is not going to be built immediately can take resources away from continuous improvements and there is always the risk of starting to count on infrastructure that might not be feasible.

The main point is that the ongoing challenge is to keep working on the range of data that is available to the official statistical system. Once it is available, there will be good ways of using it, so the main challenge is to focus effort in areas where there are priority needs. The challenge is to find the points of influence with agencies outside the National Statistical Office that gets them enthused about the idea of using standard populations, classifications, questions and methods. If there are clear and simple sources of coherence built into data sources, then this creates possibilities that can potentially be exploited.

So despite this paper exploring more technical matters in some detail, the challenge remains, as usual to manage relationships in a way that ensures that progress towards a wider set of possibilities is made continuously.

Appendix A - 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 to 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

Appendix B - Making administrative data usable

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.

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.

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.

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.

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 combing data will be used.

- 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.

- 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.

Bibliography

[1] McKenzie, R (2008) “Statistical Architecture for a New Century”, paper presented at the Korea National Statistical Office Conference, 19-20 May 2008

[2] Heerschap, N. and Willenborg, L. (2006) “Towards an integrated statistical system at Statistics Netherlands”, International Statistical Review, VOL 74; NUMB 3, pages 357-378

[3] Various internal papers and other resources from Statistics New Zeland