WP. 17
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Oslo, Norway, 24–26 September 2012)

Topic (iii): Editing and Imputation in the context of data integration from multiple sources and mixed modes

All the answers? Statistics New Zealand's Integrated Data Infrastructure

Prepared by Felibel Zabala, Rodney Jer, Jamas Enright and Allyson Seyb, Statistics New Zealand

I. Introduction

1. Since 2002, Statistics New Zealand has been working on integrating data supplied by various government agencies, including its own. Several independent projects successfully linked these datasets where each linked dataset answered specific questions. These linked datasets can provide more information other than what they currently deliver if they can all be linked together. Linking these datasets in a single environment will allow us to use and analyse data across the different datasets as part of research, official statistics, or special queries. However, when pulling the individual data together, difficulties can arise in the environments where these linked datasets are created and stored.

2. Statistics NZ's Integrated Data Infrastructure (IDI) is being developed to bring together data from these linked datasets into a single environment. The future IDI will be an integrated data environment with longitudinal microdata about individuals, households, and businesses. The IDI will allow the linking of individual and business-level data. It will also be responsive to new data sources and changes to existing datasets. Researchers will be able to select the variables they need to answer research, policy, and evaluation questions that will inform decision-making.

3. Data in the IDI environment comes from different suppliers, including Statistics NZ. The quality of the different datasets is variable, both within a dataset and between datasets. Linking clean datasets is not easy. Linking datasets of various qualities is much more difficult, so an effective and efficient editing and imputation strategy is essential.

4. This paper will present some of the issues on and solutions to any linked administrative dataset, with a focus on one of Statistics NZ's first integrated dataset, the Linked Employer-Employee Data (LEED). Section II introduces the LEED dataset and discusses its editing and imputation methods. An overview of the prototype of the IDI is presented in Section III. Future work on LEED and the IDI follows in Section IV.

II. Linked Employer-Employee Data

5. One of Statistics NZ's first linked datasets is LEED. Information in LEED provides the backbone of the IDI prototype described in Section III. LEED is created by linking a longitudinal employer and non-employer series from Statistics NZ's Business Frame to a longitudinal series of payroll tax data from Inland Revenue (IRD). The Business Frame is a regularly maintained list of all businesses and organisations with a turnover greater than NZ$30,000 that are engaged in the production of goods and services in New Zealand.

6. LEED is used to produce quarterly statistics that measure labour market dynamics at various levels – including industry, region, territorial authority, business size, sector, sex, and age. These statistics provide an insight into the operation of New Zealand's labour market. Some statistics it provides are filled jobs, worker flows, and total earnings. These statistics, together with other official statistics from Statistics NZ on aggregated movements on employment, help explain the causes of aggregate movements in jobs and worker flows and are therefore useful for explaining changes in the labour market. In LEED, a job is defined as a unique employer-employee pair present on an employer monthly schedule (EMS) on the 15th of the middle month of the reference quarter. Counts of filled jobs are restricted to people aged 15 years and over.

A. LEED payroll data

7. Inland Revenue collects payroll data from employers for New Zealand's taxation system through the EMS. EMS data includes the following information for an individual employee:

- name and IRD number
- taxable earnings (plus earnings not liable and lump-sum indicator) for work performed including social security payments taxed at source of income
- tax deductions (pay-as-you-earn or PAYE, withholding tax, child support payment, student loan indicator amount)
- start and finish dates of employment.

The EMS data also includes payments to beneficiaries by government, for example, retirement pensions, unemployment benefits, and student loans.

8. Data from a subset of the self-employed are captured in the EMS. These are taxpayers who are deducted withholding tax. A majority of them, however, pay their taxes annually using a different return. Also, a small number of the self-employed are deducted PAYE.

9. LEED covers every taxpayer who pays tax at source on income. These payments are made by employers registered with the IRD. Each employer and employee carries a unique identifier – their IRD number. These numbers are of high quality and are generally stable over time, enabling LEED to contain longitudinal data of employment. Date of birth, address, and a resident indicator (New Zealand or overseas) are also available from IRD tax data.

10. In LEED, the collection unit is the legal entity that files the EMS return. The statistical unit, referred to as 'employer' in LEED, is the geographical or physical location of the business. This means that each branch of a nationwide retail chain is a distinct employer in LEED regardless of the chain filing only one EMS return. This is needed to ensure that statistics can be aggregated by region and industry.

B. Methods of integration in LEED

11. Two different entities are involved in the linking – the businesses and individuals. For each time period and over time, records are linked within each data source. There is also a link between each data source. The links are indicated by the arrows in figure 1 (Bycroft, 2003).

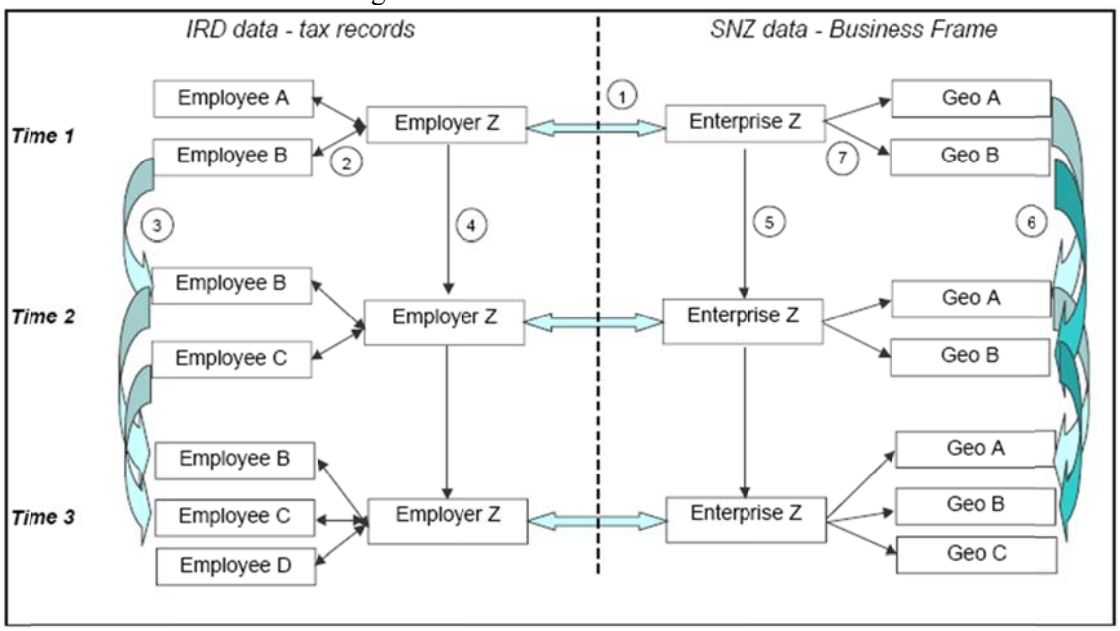
Linking employer to enterprise

12. The key link for businesses exists between the tax data and the Business Frame. Since all employees and employers have unique IRD numbers, these numbers can be used to link unit records in the tax data. In particular, IRD numbers can be used to match businesses from tax data to employers in the Business Frame since these are transferred from one to the other. The Business Frame provides the link of enterprises to their associated geographic units (or geos).

13. In a month, 90 to 95 percent of all employers on the EMS have an exact match on the Business Frame. Some of these are false matches that occur when an employer is linked to the wrong enterprise in

the Business Frame. This type of error has serious implications for LEED outputs since incorrect information was obtained from the wrong enterprise. Linking aims to minimise this type of error.

Figure 1. Unit record links in LEED



14. Different forms of reporting units are a major concern. This occurs when several businesses, possibly involved in different business activities, report their tax as one unit, known as an EMS group filer. Only one enterprise from the Business Frame will be matched to the IRD number so that the whole group's employees will be assigned to the enterprise. This will usually mean that the wrong industry was assigned in the Business Frame, resulting in an underestimate of the job and worker flows for an industry at the enterprise level (Bycroft, 2003). Because of the significance of EMS group filers, the LEED system built in methods to clearly identify these entities. The process of allocating jobs to a business is used to help identify affected employers in these situations. This process is dependent on the imputation of the workplace of an employee. A discussion of the imputation method used is presented in the next section.

15. A clerical review process also picks up the following situations: a change in the structure of an existing group filer, cessation of an existing group filer, and the birth of a new group filer.

Linking employer longitudinally

16. Linking an employer over time is simply a matter of matching that employer's IRD number each month. A false positive match occurs if there was an error in the IRD number. However, errors of this type are rare because of the high quality of tax data. A false negative match occurs when an employer continues its business but reports its tax using a different IRD number. The Business Frame discontinues the old IRD number indicating the death of the business, while the new IRD number indicates the birth of a new business.

17. A continuing business, regardless of a change in IRD number, usually retains most of its characteristics over time. Therefore, it is important to recognise when employers in different periods are in fact the same business. Failure to identify this link will result in some businesses that are continuing enterprises being incorrectly classified as business births or deaths. This will upwardly bias both job and worker flow statistics, and will also adversely affect the ability to produce statistics on the dynamics of new businesses as opposed to continuing businesses. Therefore, false negative matches should be minimised.

18. LEED uses probabilistic matching to repair employer links. There are two kinds used. The first one matches geo births to existing geos on the Business Frame. The other matches IRD reporting unit births to existing IRD units.

19. Linking of employers longitudinally is straightforward if a one-to-one match exists between a ceased unit and the birthed unit. A more complex process is used for an EMS group filer whose IRD number changes when a merger, acquisition, or a disaggregation activity is involved.

Linking enterprise and geographic unit (geo) longitudinally

20. The Business Frame, which constantly updates the values of a business, was used to create the Longitudinal Business Frame. Linking a geo is usually simple even for continuing units that change IRD numbers. Probabilistic matching is employed using the name, address, and phone number of a geo. The Annual Frame Update Survey provides information about the death of a business. Knowing the continuity of geos that comprise an enterprise improves the linking of an enterprise over time.

21. Greater emphasis is placed on minimising false negative matches. This will ensure that the performance and employment dynamics of new and ceased businesses and continuing businesses are reported correctly. At the same time, false positive matches should also be kept at a minimum.

Linking employee longitudinally

22. Good quality outputs for LEED are dependent on the successful linking of an employee over time. Although an employee reports their tax using the same IRD number each month and this is guaranteed by IRD, recording errors in the EMS are inevitable. Two percent of records have a missing employee IRD number, 1 percent have IRD numbers registered for employers, and other records will have incorrectly recorded IRD numbers (Bycroft, 2003).

23. False negative matches lead to an increase in worker flow statistics, so care should be taken to minimise these errors when linking employees longitudinally. Two methods of probabilistic matching are used to minimise false negative matches. The plug-hole method assumes that errors create one-period gaps in a job history, which may be caused by a transcription error. Linking requires finding the correct employee IRD number for all those that are missing or incorrect. Where errors occur at random, they will largely be one-period job histories, known as 'plugs'. Using probabilistic techniques, we attempt to match one-period job histories, or plugs, to the holes. The matching variables include an employee's name, gross earnings, tax deductions, and employer's IRD number. This method usually finds the correct IRD numbers for 4 percent of employees with zero IRD numbers and affects 1.3 percent of eligible IRD numbers.

24. The second method, zeros-valid processing, aims to find the correct values for employees with zero IRD numbers. This is done by looking at the other job records for the same employer, called 'valids', and doing probabilistic matching on employee name. The plug-hole method is performed first before the zeros-valids method.

C. Editing in LEED

Editing IRD numbers

25. Even if the quality of the IRD data is high, a significant amount of editing and transforming is needed before the production of official statistics. The following editing activities are done as soon as the IRD data are received and loaded. First, to ensure that each taxpayer is identified by a single, stable, and unique identifier, their IRD numbers need to be standardised. An example for this is when an individual is declared bankrupt. At the time of declaration, IRD issues a new IRD number. The LEED system should be able to cross-reference the old and new IRD numbers to ensure the longitudinal linking of the individual across time. This is one process that updates employee IRD numbers.

Editing gross earnings

26. The next step is to edit gross earnings. Gross earnings are among the key outputs of LEED but IRD does not edit an individual's gross earnings. Since IRD edits an employee's PAYE, this variable is used to detect errors in gross earnings. Systematic errors are the most common type of error in gross earnings. One example is when an IRD number with eight digits or more is mistakenly recorded as gross earnings giving an amount greater than 10 million. Another example of systematic error is misplaced decimals. To detect this type of error, a ratio edit – PAYE/gross earnings – is used. Ratios that are neither between 0.1 and 0.5 nor equal to 1 are flagged as suspicious values. There are several valid reasons why PAYE rates may be outside this range. For example, if an individual is claiming a loss against previous earnings, the tax rate can be as low as zero. If an individual has a large amount of unpaid tax owing to the IRD, the rate can be as high as 1.0. Therefore, where the rate is outside the expected range we need to check if the entry is unusual based on the history of the entries for that individual. Employee histories with a flagged suspicious record are then used to verify these records are atypical. The LEED editing strategy is to not replace any IRD data unless there is strong evidence that it is an error.

27. For a failed record for each employer-employee return period, a typical PAYE/gross earnings rate from entries that passed the edit is calculated from the history of the employee. The rate is used from the employer associated with the edit failure, where possible, or from all the records for that employee otherwise. This rate is then applied to the PAYE deductions in the records that have failed the edit check to derive an imputed value for gross earnings. This imputed value is then compared with the original gross earnings to check that the error in the original value is worth correcting. A check is also made to see if the raw value is incorrect because of a misplaced decimal point. The gross earnings time series is then examined to see if the imputed value has reduced the variability of the series significantly. If not, then the original raw value is retained (Graham, 2005).

Editing date of birth

28. The age of an employee is a key variable when analysing employment and earnings patterns. Ninety-six percent of the records supplied by IRD have dates of birth. However, a few of these have either the wrong century recorded or do not match an employee's past data. The editing rules for date of birth are based on checks on an employee's age against some events. Some examples of edit rules follow. An individual collecting parental leave should be aged between 15 and 60. An individual collecting superannuation should be at least 25 years old when they first started collecting the pension. An individual should not be older than the oldest person currently alive in New Zealand (113 years).

29. The latest supplied birth date of an employee is retained regardless of whether this is inconsistent with the employee's historical data. Entries before 1900 are adjusted to the 20th century. Nearest-neighbour donor imputation is used to impute for missing and invalid dates of birth. The date of birth is copied from the individual with the closest mean gross earnings within the same income source. Care is taken to use a donor age for only a few recipients to avoid clustering of imputed dates of birth.

Imputing sex

30. Sex is another key variable used in analysing employment and earnings patterns. This variable is not available from the supplied IRD data. In the first instance, sex is imputed from sex-specific titles, for example, Mr, Mrs, Miss, for each individual where such a title exists. Otherwise, sex is derived from the first name field, that is, the first name and middle name supplied by IRD. Stochastic imputation is used to impute for sex. A frequency table that provides the number of times a given first name has been associated with males or the number of times it has been associated with a female is used in the imputation.

Imputing workplace of an employee

31. Assigning an employee to a workplace is carried out for complex enterprises that are either multi-geo enterprises or enterprises in an EMS group filer. Accurately assigning an employee to a workplace guarantees aggregated outputs, that is, the regional and industry totals are of good quality. The job history of an employee is also affected by the imputation of the workplace of an employee. In this paper, 'workplace' refers to a geo the employee is assigned to.

32. The difficulty with employees from a complex enterprise is the lack of information on their exact workplace. Unless employee counts are available in the Business Frame, another problem is the lack of information on the number of employees to expect from a geo of a complex enterprise.

33. A linear programming method, the transportation method, is used to impute the workplace of an employee. The objective of the transportation method is to minimise the cost of transporting goods from each of several points of origin (or supply points) to each of several points of destination (or demand points). The constraints are the output available from each supply point and the demand required from each of the destination points. As a transportation problem, the supply point for the imputation of the workplace of an employee corresponds to an employee, that is, a supply point is an employee and supply equals 1 for all employees (or supply points). The demand point is an employer or a workplace where an employer has a predetermined number of employees. This number is available from the Business Frame and is sourced from either a survey or other tax data. The cost is the distance (in kilometres) from an employee's home address to each possible workplace.

34. As a transportation problem, the imputed workplace of an employee is the geo that minimises the distance between an employee's home address to the geo, subject to the constraints that:

- each employee is assigned to a geo
- the total number of employees allocated to a geo should equal the number of employees expected from the geo.

35. A process, called rebalancing, is performed after imputing the workplace of employees from a complex enterprise. This is done to ensure that the employee proportions associated with each geo in a continuing complex enterprise do not constantly change over time and should reflect what is reported in the Business Frame. Thus, after employees from complex enterprises are allocated to a geo, the proportion of the total number of employees allocated to a geo is validated against its target proportion on the Business Frame. If the number varies significantly from the target, a number of jobs are reallocated between geos and this may require jobs that were set to be continuations to be reallocated to a different geo.

Imputing start and end dates of employment

36. The start and end dates of a job are only recorded in the month the job started or finished. Jobs with consistent start and end dates are not imputed. A job has a consistent start date if the start date in the job history falls within the month corresponding to the start of employment. A similar condition holds for a consistent end date. Jobs with missing start/end dates are imputed for. Jobs with inconsistent start/end dates have these values nullified and imputed for. The imputed values ensure that the start and end dates are consistent. Logic rules are used in the imputation. Some examples follow.

37. Examples of logic rules used to impute the end date are:

- if the start date is not in the current month, impute the end date as a random date within the month
- if the start date is in the current period and no other job starts in this period for the employee, impute the end date as a random date in the month after the start date unless start date is the last day of the month, in which case, impute the end date as the job start date
- if the start date is not in the current month and another job starts in this month for the employee, impute the end date as the day before the new job starts unless the new job starts on the first of the month. In this case, impute the end date as the new start date.

38. Examples of logic rules used to impute the start date are:

- if no other job finished in the month, impute start date as a random day in the month unless the end date is the 1st of the month, in which case, impute start date as the end date, that is, the 1st of the month.
- if the employee had another job finished in the month, and if the job being processed finished in the month and the other job finished after this job, impute start date as a random day in the month before its finish date, unless the end date is the 1st of the month, in which case, impute the start date as the end date, that is, the 1st of the month.

III. The Integrated Data Infrastructure prototype

39. The IDI prototype was developed to enable Statistics NZ to test, analyse, and develop the IDI, and its processes and outputs. Researchers can use data in the prototype while a full productionised IDI is being completed (Statistics NZ, 2012).

A. Data sources

40. The IDI prototype integrates five independently linked datasets. A considerable amount of information is duplicated in these datasets. A common dataset for four of the linked datasets is LEED. Data consolidated from these datasets form the integrated Longitudinal Employment and Education Data (iLEED) component of the IDI. The following lists the integrated datasets linked with LEED:

- benefit data from the Ministry of Social Development (MSD) to form the LEED-MSD dataset
- tertiary education data from the Ministry of Education (MoE) to create the Employment Outcomes of Tertiary Education (EOTE) dataset
- administrative tertiary education data from MoE and student loans and allowances data sourced from IRD, MSD, and Student Loans and Allowances Manager (SLAM).
- survey data from Statistics NZ's Household Labour Force Survey (HLFS) and its supplementary surveys.

B. Linked datasets in the IDI

41. The benefit data linked to LEED provides information on the benefit type received and the demographics of beneficiaries. Types of benefit received include unemployment, domestic purposes benefit, and invalids. Personal details of beneficiaries include sex, date of birth, and ethnicity. Thus, the LEED-MSD dataset produces statistics on transitions onto and off main benefits by benefit type received at the point of transition (Statistics NZ and Ministry of Social Development, 2008). The personal details available from the MSD data are used to impute missing personal details of employees in LEED.

42. The EOTE dataset contains information on institution-based tertiary education and workplace-based tertiary education. Statistics on employment and earnings of recent participants in tertiary education are produced from this dataset. The personal details available from the MoE data are also used to impute missing personal details of employees in LEED. The SLA dataset provides information on the different characteristics of student loan borrowers, in terms of personal variables (such as sex and age), and characteristics of their loans (amount borrowed, total debt, and voluntary repayments) (Statistics NZ, 2009).

43. The HLFS data consists of variables related to individuals and their work and labour force status. The following information is available from the HLFS supplementary surveys:

- a comprehensive range of income statistics from the New Zealand Income Survey
- changes in the employment conditions, working arrangements, and job quality of employed people from the Survey of Working Life
- factors behind internal migration in New Zealand from the Survey of Dynamics and Motivation for Migration in New Zealand (Statistics NZ, 2011).

44. The LEED-HLFS integration project is new. Some of the aims of the LEED-HLFS integration project are to test the feasibility of linking HLFS data to LEED and to identify the benefits and risks of integrating administrative sourced data to Statistics NZ's survey data. No outputs were produced so far from the dataset. Further work on the integration project is now being undertaken under the building of the IDI.

45. The Longitudinal Business Database (LBD) prototype integrates longitudinal administrative and survey data at the enterprise level. It was developed to provide information on the dynamics of business performance. Data from the LBD includes information on business demographics, financial data, employment, goods exports, government assistance, and management practices (Gretton, 2008).

46. The usefulness of new administrative data sources and files to be linked in the IDI is assessed using Statistics NZ's metadata template for administrative data. The metadata template is based around a quality framework for statistical outputs, where quality is made up of six dimensions: relevance, accuracy, timeliness, accessibility, interpretability, and coherence. Use of the template provides:

- a focus and direction for those needing to understand an administrative data source
- a structured format for the documentation of the administrative data
- the information needed to assess the statistical integrity of an administrative data source
- the information needed to determine the usefulness of the administrative data source in any given context (Statistics NZ, 2006).

B. Integrating and editing data sources

47. There are two methods used for record linking in the IDI, depending on the existence of a unique identifier in the records of the files being linked. Exact linking is used when an identifier of good quality exists. Using this method creates a dataset with records that will either exactly match or not match. This type of record linkage is easy to do and doesn't require specialised linking software.

48. When a unique identifier is not available, or is not of sufficient quality, probabilistic linking is used. This method requires a combination of partial identifiers to compute scores or 'weights' that measure the probability that two records match. A cut-off weight is set to decide whether a pair of records is to be linked or not. Partial identifiers used in the IDI include names, addresses, date of birth, and sex. Probabilistic linking is more complex and requires sophisticated linking software to obtain high-quality results.

49. Figure 2 represents the linking of the five linked datasets and the integration of new administrative migration data from the Department of Labour in the IDI prototype (Statistics NZ, 2012). A key part of the iLEED component is the link between the different data sources. The lack of a common identifier across these datasets complicates the linking process, so probability linking is used. The lack of a common identifier is offset by the presence of common demographic variables across the data sources. These are names that are usually split into first and last names, sex, and date of birth, which are the main variables in the Central Linking Concordance (CLC). In some cases, administrative data sources have extra identifiers that allow for a near exact match, such as IRD numbers, passport numbers, and provider student ID number/MoE national student number. These identifiers will be of high quality and are also part of the CLC. Links based on just the demographic variables are of lower quality depending on how common the names are.

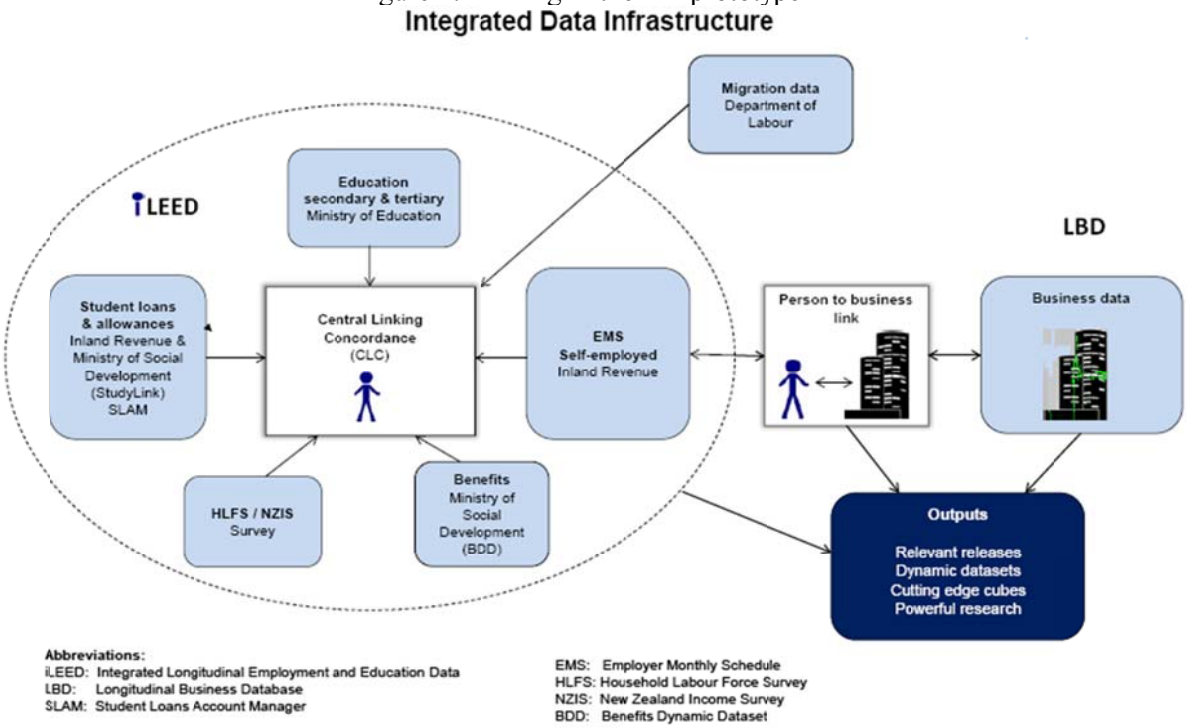
50. Due to the number of data sources in the IDI, the linking process was split into four distinct stages. Stage 1, DoL-IR link, involved linking migration data with the tax data. Stage 2, MoE-SLA link, linked Inland Revenue, Ministry of Social Development, and Student Loans Account Manager student loans and allowances data with Ministry of Education data. Stage 3, MoE-IR link, integrated the tax and tertiary education data. Stage 4, IR-HLFS link, integrated tax data with HLFS data.

51. Integrating data requires high-quality linking variables to obtain high-quality results in record linking. Editing data received from the source agencies will focus on ensuring high-quality linking variables are used in linking. Validity rules were used to edit names across data sources. Some

examples of these rules were the detection and deletion of non-alpha characters and replacing these with a space. If a name belongs to an identified list of words that were not used (for example, agency, associate), the word is deleted from a name. If a person is supplied with multiple names that are recorded across rows, these names are concatenated to provide only one row per person.

52. The variables sex and date of birth are only reformatted to ensure common coding is used across data sources. For data from the Inland Revenue, imputation of a person's sex and the editing of the date of birth follow the methods used in LEED. An IRD number check is used to validate an IRD number. Invalid IRD numbers are set to null. The provider student ID number/MoE national student number are standardised to ensure the quality of information derived for education-sourced data.

Figure 2. Linking in the IDI prototype



53. Once linking variables are edited, an externally assigned identifier must be replaced by a new identifier when linking needs to be done on an ongoing basis. This new identifier is used for integration in the IDI prototype. It must not be possible to derive the externally assigned identifier from the new identifier (Statistics NZ, 2006). Transformation of externally assigned identifiers ensures the information used for analysis and research complies with principle 12 of New Zealand's Privacy Act 1993. Imputation is not carried out on IDI data used for outputs and research.

54. Two types of errors are inevitable with the volume of records being linked in the IDI. A 'false positive error' occurs when two records are linked together, when in reality they are not the same person or unit. A 'false negative error' occurs when two records are not linked together, when they do in fact belong to the same person or unit. Generally there is a trade-off between the two types of errors since, for example, reducing the rate of false positives may increase the rate of false negatives. Thus, it is important to consider the consequences of each type of error and to determine whether one is more critical than the other. Results on the four stages of linking are given in table 1 (Statistics NZ, 2012).

55. Overall, the link rates for the four stages of linking indicate integrated data is of high quality to enable researchers to use data from the IDI prototype. Developing the IDI will include work to further improve its linking process especially around the IR-HLFS stage where the false-positive rate is much higher than other stages.

Table 1. Summary of linking results

Link	Percentage linked	False positive rate (% of total number of links)
DoL-IR	77% of New Zealand citizen-residents linked to the Inland Revenue population	0.3
MoE-SLA	94% of the merged population of Inland Revenue, MSD, and SLAM data linked to the Ministry of Education population	0.1
MoE-IR	73% of the Ministry of Education population linked to the Inland Revenue population	1.1
IR-HLFS	78% of data with name information in both IR and HLFS data linked to the Inland Revenue population	9.4

56. Statistics NZ does not have the ability to validate the accuracy of most data provided by external agencies. Data quality is the responsibility of the data provider. A critical factor in ensuring the quality of externally sourced data is the management of the relationship between Statistics NZ and the data providers. Processes are in place to ensure this. Examples of these processes include meeting with providers to ask questions and to communicate knowledge critical to data being provided, communicating openly with data providers, and providing feedback on the quality of data provided.

C. Issues in integrating and editing data sources

57. Due to the volume of records linked in the IDI, standardised processes need to be in place. The processes will need to be responsive to administrative changes in the data supplied and to new administrative data available to Statistics NZ. First and foremost, standard software for automated data linkage robust to these changes should be identified. Various evaluations of record linkage software carried out within Statistics NZ led to the choice of QualityStage. To help link various data sources, each dataset is configured into a master table, which is brought into QualityStage. QualityStage is robust in that we would construct a master table for new datasets and then work in QualityStage to link them.

58. Other linking issues include timing, unless receipt of data are common and regular. This will always be an issue for linked datasets and for statistics produced from these datasets. An example is where personal income needs to be provided annually in the June quarter. Personal income is an output from one of the HLFS supplements, the NZIS. If new entrants to the workforce are not picked up quickly in the IRD tax data, potential bias may result in personal income reported. Initial research work indicates this is an issue but needs more investigative work.

59. A common editing issue with linked datasets is inconsistencies in the records linked. Where inconsistencies occur in records linked from two different data sources, it is important to know which of the two data sources is more reliable. Sometimes, even the order in which the datasets are linked is important in determining where an inconsistency arose. A process was developed to resolve inconsistencies in personal details where outputs are based on. First, the most common value present in the datasets is judged to be of higher quality so should be kept. The data sources were also prioritised to determine the order of retaining their values in case of inconsistencies in the datasets. It is expected that as the number of datasets being linked together increases, the potential for efficiencies in detecting and treating inconsistencies in records increase as the number of variables increase. However, this may also increase the amount of editing required for the linked datasets.

60. Issues to be addressed by an editing strategy for linked datasets can be summarised by its ability to:

- edit inconsistencies from the same unit from different sources
- treat erroneous and missing variables in a record
- ensure consistency in variables across a record for a time period and over time.

IV. Next steps

61. So far, there are numerous research projects currently using or scheduled to use the IDI prototype. Many of these projects are only possible because of the unique opportunities brought about by the linking of large-scale data across government agencies. By bringing together many datasets into one infrastructure, the IDI allows for longitudinal analysis across education, employment, migration, welfare, and business, with potential for further datasets to be added and linked in the future.

62. Our next steps include building the IDI, using the experience gained from the prototype with a focus on improving the linking methodology and redeveloping the LEED and SLA systems. These steps are part of the organisation's Statistics 2020 Te Kapehu Whetū (Stats 2020) transformation programme. Building the IDI will maximise the use of available administrative and survey data, and will be more responsive to changes in existing data sources than is currently possible (Statistics NZ, 2012). It will also increase the ability of the IDI to answer a wider range of research questions.

63. Stats 2020 is a long-term programme of change so that Statistics NZ achieves its vision of 'an informed society using official statistics' (Statistics NZ, 2010). The outcome sought from the Official Statistics System's activities is that statistics are increasingly used to inform decisions, and to monitor and understand the state and progress of New Zealand.

64. Another Stats 2020 project will determine the standard quality measures for outputs produced using administrative data and those from mixed sources, and for the modes of collecting data.

65. Building the IDI provided the opportunity for redeveloping and improving LEED. So far, most of the redevelopment work focused on reviewing the employee allocation method and the relationship between LEED and the Business Frame. This work aimed to improve the linking between the employer in the tax system and the enterprise in the Business Frame.

66. The review of the employee allocation method covers the fixes to the job history of employees, that is, the link of employees over time from a methodological perspective. The linking of employees to employers (or business locations) will be improved as geospatial information is used more to improve the quality of the allocation. New technology will enable the association of residential and employer addresses with Global Position System (GPS) coordinates. The current method uses the centroid of the area unit containing the residential address, and the centroid of the territorial authority containing the employer.

67. The editing and imputation of gross earnings is also being reviewed from a methodological point of view. The use of Banff is currently being investigated from the perspective of moving the processing of LEED into one of Statistics NZ's standard processing platform, the microeconomic platform. The development of a platform aims to standardise, where possible, the methods and tools for processing. In the case of the microeconomic platform, Banff is the suitable standard tool.

REFERENCES:

Bycroft, C (2003). Record linkage in LEED: An overview. Available from www.stats.govt.nz.

Graham, P (2003). The IRD base data – Quality and characteristics, and LEED derivation, edit and imputation methods. Available from www.stats.govt.nz.

Graham, P (2005). *Gross earnings edit and imputation specs for LEED (Model no 2.1.4).* Unpublished document. Statistics New Zealand.

Gretton, J (2008). *Developing the prototype longitudinal business database: New Zealand's experience.* Paper presented at the International Association for Official Statistics Conference, 14–16 October 2008, Shanghai, People's Republic of China.

Statistics New Zealand (2005). *LEED system implementation documentation*. Unpublished document, Statistics New Zealand.

Statistics New Zealand (2006). Data integration manual. Available from: www.stats.govt.nz

Statistics New Zealand (2009). Employment outcomes of tertiary education technical report and feasibility assessment. Available from www.stats.govt.nz.

Statistics New Zealand (2010). Statistics New Zealand strategic plan 2010–20. Available from www.stats.govt.nz.

Statistics New Zealand (2011). Privacy impact assessment for the feasibility study for linking Household Labour Force Survey data with Linked Employer-Employee Data. Available from www.stats.govt.nz.

Statistics New Zealand (2012). Integrated Data Infrastructure and prototype. Available from www.stats.govt.nz.

Statistics New Zealand and Ministry of Social Development. (2008). Linked Employer-Employee Data/Ministry of Social Development Feasibility Study: A report on the feasibility of integrating benefit data with Linked Employer-Employee Data, to produce official statistics. Available from www.stats.govt.nz.