Feasibility Report:  
Linked Employer-Employee Data (LEED)  

30 September 2003
Contents

Executive Summary........................................................................................................................ 2
Recommendations.......................................................................................................................... 3
1. Introduction.............................................................................................................................. 5
2. Background ............................................................................................................................. 5
3. Prototype Official Statistical Outputs ................................................................................... 6
   3.1. Summary of anticipated output quality........................................................................... 9
4. Research Potential.................................................................................................................. 11
5. Confidentiality ....................................................................................................................... 13
6. Data Security ......................................................................................................................... 13
7. Access .................................................................................................................................... 14
8. Data Quality Issues ............................................................................................................. 15
   8.1. IRD data ......................................................................................................................... 15
   8.2. SNZ Business Frame data ............................................................................................. 19
9. Unit Record Links ................................................................................................................ 20
   9.1. Description of links ....................................................................................................... 20
   9.2. Linking Issues .............................................................................................................. 21
10. Privacy ................................................................................................................................ 22
    10.1. Privacy assessment ..................................................................................................... 22
    10.2. Potential future privacy issues ..................................................................................... 23
11. Information Technology Issues .......................................................................................... 23
12. Remaining Project Risks ..................................................................................................... 24
13. Conclusion ........................................................................................................................... 24
14. Next Steps ........................................................................................................................... 25
Executive Summary

1. The creation of a linked employer-employee dataset (LEED) in New Zealand is feasible. LEED integrates existing, full coverage administrative data from the IRD PAYE and income tax systems together with business data from SNZ's Business Frame. The conclusion that this initiative is feasible has been reached following an examination of the quality and characteristics of the input data, its ability to support the proposed use and an examination of the legal context and issues surrounding the protection of individual privacy and data confidentiality.

2. LEED has the potential to make the single most significant advance in recent years in statistical information. By using existing information, it will make no demands on respondents for additional information. There is also the potential to reduce or refocus existing surveying for the production of employment statistics.

3. Potential LEED outputs will provide new insights into the dynamics of labour market adjustment and market outcomes from both a business and employee perspective. These statistics will inform policy development, monitoring, debate, and decision-making both within government and by the wider community.

4. Initial work undertaken to test the ability of the data to support policy relevant research has confirmed the potential of the data in this area. Preliminary information has been produced for the first time on the dynamics and earning outcomes for beneficiaries moving into and out of employment. This work provides a first step to producing official indicators of stable and sustainable employment.

5. The potential of this data to support a wider variety of initiatives means that priority is being given to preparing a broad development programme with users to guide and prioritise both the development of additional official statistics and policy-relevant research.

6. The base data is of high quality and will, once fully developed, support the requirements for the production of relevant, high quality statistical outputs.

7. Existing confidentiality practices will provide a high level of protection for individuals’ data in both official statistical outputs and during micro data research.

8. The protection of business data is more problematic. While existing practices, in theory, apply in this situation, in practice they appear unsustainable. New practices are now emerging internationally that confidentialise information through the addition of statistical 'noise' to unit record data. This approach has been approved by the United States Bureau of the Census Disclosure Review Board and has been adopted to successfully protect similar data in the US. SNZ specialists are currently reviewing these practices in a New Zealand context.

9. All data is subject to SNZ's strict security provisions. Access to this information is on a 'need to know' basis with use logged. All identifying information is removed from individuals' records early in processing.

10. Access to micro data for bona fide statistical research is seen as an important component of the programme. Users and data suppliers are both looking for clarification of the boundaries and processes that will govern LEED microdata access. The revision of policies and processes governing microdata access is a corporate issue that SNZ is progressing. This process will include consultation with data suppliers and users before a final strategy is determined.

11. The SNZ Business Frame will support the main requirement of the LEED project for a comprehensive longitudinal frame. Work remains in designing and testing a process to update this series on a production basis.
12. The longitudinal business frame series has the potential to provide enhanced measures of business survival and growth, as well as providing the potential to act as a frame for the linking of business survey data.

13. Much of the power of LEED data is derived from the number of different links that can be made within the integrated data, e.g., between people and their employer and for both businesses and people over time, etc. The IRD number is a high quality numeric identifier that allows these links to be established simply in the majority of cases. Individuals’ IRD numbers will be removed following cleaning and will not be retained in the linked dataset.

14. There are seven major links within the LEED data; four require repair to support the demands of statistical use. Failure to repair these links will significantly distort longitudinal outputs and analysis. This work is new to SNZ and in development requires skilled mathematical statisticians with the ability to apply probabilistic matching techniques. It will take time to complete these repairs and carry out testing. Overseas agencies undertaking this work have faced similar challenges. Their work provides a precedent for development, and gives the project team confidence that the issues are manageable.

15. The linking of data collected for unrelated purposes does create risks to individual privacy and the confidentiality of individual and business information. The risks to individual privacy arising from the LEED project have been identified, and processes have been developed that will ensure that privacy is not jeopardised. Credible responses to public concerns are available if required.

16. IMS specialists have concluded that the existing SNZ IT system will not support LEED requirements in a production environment.

17. IMS and external warehousing expertise has been engaged to assess the IT architecture options available to support these projects for the next three to five years. These findings will form part of the budget business case proposed for 2004/05. This work will be completed by 31 October 2003.

Existing project risks have been identified and are being managed. Remaining risks include the absence of baseline funding, uncertainty among users and data stewards regarding the boundaries and processes governing microdata access, and the development and maintenance of critical expertise.

**Recommendations**

That the Government Statistician, following consultation with the IRD Commissioner and Chief Executive of the Department of Labour:

1. Agrees that the first stage of the project has successfully demonstrated the:

   a) feasibility of producing statistically robust data from linked employer and employee administrative data

   b) manageability of associated data security, confidentiality and privacy issues

   c) potential of these data to produce new and useful employment- and firm-related official statistics and to provide a basis for public policy research.
2. **Directs** the LEED team to complete the current CDRP-funded development programme, including:
   
   a) completion of the data linking, update and confidentiality methodologies  
   b) testing the feasibility of incorporating self-employed persons  
   c) continued development of prototype statistical outputs policy research applications.

3. **Directs** the SNZ team, in consultation with the Sponsors, to progress the development of an ongoing LEED programme, including:
   
   a) preparation of an appropriate case seeking funding to complete the development and support the operation of the programme  
   b) consultation with stakeholders to build a user constituency and maintain the confidence of data providers  
   c) the establishment of a medium-term work programme prioritising the development of new official statistics and policy-research applications  
   d) continued development of appropriate policies and protocols to govern micro data access, protect confidentiality and manage privacy issues.

4. **Approves** the retention of the existing trial datasets for incorporation into a production system.
1. Introduction

This paper reports on the feasibility of developing a linked employer-employee dataset (LEED) for the production of official employment and business statistics. The dataset would also support policy-relevant statistical research. Information is presented to enable the three sponsoring agencies, Statistics New Zealand (SNZ), the Department of Labour (DoL) and the Inland Revenue Department (IRD), to evaluate the potential benefits, costs and risks that would result from developing this new information source.

Key findings are presented on:
- the first set of official statistics proposed for development
- the data’s ability to support policy development and monitoring
- confidentiality
- data security
- data use
- data quality
- the Business Frame’s ability to support LEED requirements
- linking
- privacy
- IT requirements, and
- major project risks.

Following an assessment of this information, the Government Statistician, in consultation with the IRD Commissioner and Chief Executive of the DoL, will determine whether the benefits can be achieved with satisfactory management of privacy and therefore whether the project should proceed beyond the current investigatory phase. If approval is given to establish the proposed outputs as new official statistics, the existing trial datasets would form the basis of an on-going production series. Should approval be declined, development work will cease and all input and trial integrated datasets shall be destroyed, as agreed at the inception of the project.

This report is supported by a compendium of technical papers that provide detailed information on the issues encountered and the proposed methods to transform the base data into a robust statistical dataset.

2. Background

The objective of the LEED project is to successfully integrate existing employer and employee information to provide new insights into the operation of the labour market and business performance. This initiative follows the successful development of similar datasets by a number of European and North American official statistical agencies (eg US, Canada, France, Sweden, Germany, etc). LEED draws on existing, full coverage administrative data from the IRD PAYE and income tax systems, together with business data from SNZ’s Business Frame.

Initial funding for this work was obtained from the Cross-Departmental Research Pool for two years from June 2002 to July 2004. The Department of Labour has provided additional funding.

---

1. The SNZ LEED team wishes to thank the staff of the IRD, the Department of Labour’s Labour Market Policy Group and the United States Bureau of the Census’s LEHD team for their invaluable contribution to this project.
and support from the Skills Action Plan to facilitate input from international and subject area experts. Access to this expertise has made a significant contribution to the development of SNZ’s capability in this area and has enhanced the speed and quality of the development.

3. Prototype Official Statistical Outputs

Stage 1 of the project has attempted to identify the potential of the data to provide new official statistics that inform debate, decision-making and research, both within government and by the wider population, without increasing respondent load. After discussion with key users, and in line with similar developments internationally, priority has been given to assessing the potential to provide new measures of labour market dynamics, as well as enhanced measures of business dynamics.

Considerable uncertainty currently exists regarding the process by which the New Zealand labour market adjusts to economic and social change. Official statistics exist that measure the net change in employment between two points in time; however, the mechanism by which the market moves between those two points is largely unknown. For example, no measures are produced that show the number of new jobs created, the number of existing jobs destroyed, or the number of workers who change jobs, or enter or leave the labour force.

The production of prototype series from LEED has shown that the data can support a range of new statistical outputs that will answer these and many other questions. The proposed outputs and the information gaps they address are outlined in Table 1:
**Feasibility Report: Linked Employer-Employee Data (LEED)**

### Table 1

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Information gap</th>
<th>Examples of questions the data has the potential to answer</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Job flows - job creation and destruction.</strong>&lt;br&gt;The number of jobs created/destroyed by either new/ceased businesses or the expansion/contraction of employment by existing businesses.</td>
<td>Statistics on net changes in jobs, as published from measures such as the Quarterly Employment Survey, are based on a comparison of the number of jobs at two points in time. Such measures do not provide any information on how many jobs were created or destroyed between these two points (the gross creation and destruction of jobs that occurs to achieve a given net change).&lt;br&gt;This information is important in understanding business and labour market adjustment processes and outcomes.</td>
<td>What industries/sectors are creating/destroying jobs?&lt;br&gt;In what regions is job creation/destruction occurring?&lt;br&gt;To what extent do job creation/destruction rates vary by firm size or other firm characteristics?</td>
</tr>
<tr>
<td><strong>Worker flows - employee accessions and separations.</strong>&lt;br&gt;The number of new employees added to or removed from the payroll of an employer during a period.</td>
<td>Below the level of job flow statistics are worker flow statistics, which measure flows of individual employees into and out of businesses. Earnings outcomes for workers can also be assessed.&lt;br&gt;These statistics report how individuals and firms adjust to changing economic signals.&lt;br&gt;No information has previously existed on these flows, which provide an insight into the relationship between business performance and work force compositions and the dynamics and outcomes of employees' labour market experience.</td>
<td>What industries are gaining/losing the most workers? Is employment expanding or contracting as a result?&lt;br&gt;Which industries are gaining/losing older workers/younger workers? Are there differences by sex?&lt;br&gt;What type of workers (older/younger, male/female) have higher/lower turnover rates?&lt;br&gt;What regions are gaining/losing employees?&lt;br&gt;Are new workers being drawn from a pool of previous employees or are they new to the firm?&lt;br&gt;What are the earnings outcomes for these different groups (eg incumbent workers v new hires)?</td>
</tr>
</tbody>
</table>

2. In theory, the QES could produce estimates of these flows although they would be based on sample data and would have sample, coverage and possible other quality constraints.
<table>
<thead>
<tr>
<th>Statistic</th>
<th>Information gap</th>
<th>Examples of questions the data has the potential to answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employment tenure</td>
<td>Employment tenure is defined as the length of time an individual is employed continuously by a specific employer.</td>
<td>Are some industries characterised by short-term employment? Are some workers more likely to be in short-term employment than others? Does this persist over time? Are there differences in earnings outcomes for these different groups? Is there a relationship between employee turnover and business performance?</td>
</tr>
<tr>
<td></td>
<td>This is a new statistic that is designed to measure worker stability in the labour market. It can also provide insights into periods of non-employment</td>
<td></td>
</tr>
<tr>
<td>Multiple job holding</td>
<td>The number of different businesses in which a worker is employed at a point in time.</td>
<td>How many people hold more than one job and what are their earnings in each? Does multiple job holding vary by the age, sex or industry or region? How many multiple jobholders are also receiving income support? What is the income distribution of multiple jobholders?</td>
</tr>
<tr>
<td></td>
<td>Currently surveyed in the HLFS. Information collected is limited and focuses on main job. No information on other income sources eg benefit.</td>
<td></td>
</tr>
<tr>
<td>Business demography</td>
<td>Statistics on the demographics of business including counts of births and deaths. These statistics often incorporate a size measure and can be produced at different structural (ie Geo/Ent) and geographic levels.</td>
<td>What is the probability that a new business will survive after 1, 2 or 3 years of operation? What proportion of businesses grow or contract over time? What are the characteristics of businesses that grow/contract over time? How are these factors related to their workforce composition? How do these measures vary by industry and region?</td>
</tr>
<tr>
<td></td>
<td>Existing statistics provide an annual snapshot of the structure and characteristics of New Zealand business. The figures have known weaknesses in relation to inter-period comparisons and aggregate counts of firm births/deaths. LEED offers the potential to address these weaknesses and to move from annual to quarterly outputs.</td>
<td></td>
</tr>
</tbody>
</table>
In addition, the LEED data has potential to:

- Contribute to the development of an economy-wide measure of labour volume that will also enhance the measurement of productivity. For example, employees’ monthly earnings could be deflated directly to form a volume series, or hours could be modelled using other data sources.
- Replace direct surveying for small and medium-sized business currently in the QES sample, reducing respondent load and increasing the quality of estimates.
- Support detailed industry and regional statistics – universal coverage effectively eliminates sample error.
- Support outputs at a finer level than published to feed into official derived statistics such as the National Accounts.

The research findings reported below indicate that LEED would also support new measures of stable and sustainable employment, during and after government benefit receipt, including information on earnings outcomes.

A broad development programme is now being prepared in conjunction with other government agencies to guide and prioritise the development of both additional official statistics and policy-relevant research. This programme will form a part of the project's business case for new initiative funding.

3.1. **Summary of anticipated output quality**

Table 2 below summarises the expected output quality:
### Table 2

**Expected Output Quality**

<table>
<thead>
<tr>
<th>Quality dimension</th>
<th>Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Relevance</strong></td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>• provides powerful new insights into labour market and business dynamics</td>
</tr>
<tr>
<td></td>
<td>• high demand from a wide variety of users for:</td>
</tr>
<tr>
<td></td>
<td>• proposed official statistics</td>
</tr>
<tr>
<td></td>
<td>• enhancement of existing SNZ measures</td>
</tr>
<tr>
<td></td>
<td>• policy-relevant research based on unit record data</td>
</tr>
<tr>
<td></td>
<td>• no viable alternative source</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td>Very good</td>
</tr>
<tr>
<td></td>
<td>• input data generally very good but needs some cleaning</td>
</tr>
<tr>
<td></td>
<td>• links will be robust after probabilistic matching applied</td>
</tr>
<tr>
<td></td>
<td>• population coverage high for employing businesses and employees</td>
</tr>
<tr>
<td></td>
<td>• self-employed still to be investigated</td>
</tr>
<tr>
<td></td>
<td>• tax reporting units consistent with SNZ units, except for ‘group returns’</td>
</tr>
<tr>
<td></td>
<td>• some modelling needed for regional breakdowns</td>
</tr>
<tr>
<td></td>
<td>• trade-off may be required between classification accuracy and timeliness</td>
</tr>
<tr>
<td></td>
<td>• trade-off may be required between the accuracy of longitudinal links and timeliness</td>
</tr>
<tr>
<td><strong>Timeliness</strong></td>
<td>To be determined</td>
</tr>
<tr>
<td></td>
<td>• time lags exist in reporting and processing of tax and BF data</td>
</tr>
<tr>
<td></td>
<td>• trade-off may be required between accuracy of classification and timeliness</td>
</tr>
<tr>
<td></td>
<td>• trade-off may be required between the accuracy of longitudinal links and timeliness</td>
</tr>
<tr>
<td><strong>Accessibility</strong></td>
<td>High for aggregate outputs</td>
</tr>
<tr>
<td></td>
<td>• once SNZ standard processes are in place including web customised access</td>
</tr>
<tr>
<td></td>
<td>• will need further work to ensure confidentiality of detailed tables</td>
</tr>
<tr>
<td><strong>Restricted for unit record data</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• highly sensitive tax and business data</td>
</tr>
<tr>
<td></td>
<td>• may need development of new confidentiality processes</td>
</tr>
<tr>
<td><strong>Interpretability</strong></td>
<td>Very good</td>
</tr>
<tr>
<td></td>
<td>• supported by quality meta-information</td>
</tr>
<tr>
<td></td>
<td>• project emphasis on understanding and documenting source data</td>
</tr>
<tr>
<td></td>
<td>• project emphasis on understanding and documenting LEED data and outputs</td>
</tr>
<tr>
<td></td>
<td>• user education needed as measures are new to NZ</td>
</tr>
<tr>
<td><strong>Coherence</strong></td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>• concepts and definitions for employee tax data fit well with SNZ standards</td>
</tr>
<tr>
<td></td>
<td>• standard BF concepts and definitions available for businesses</td>
</tr>
<tr>
<td></td>
<td>• data consistent from start of series in 1999</td>
</tr>
<tr>
<td></td>
<td>• internationally recognised measures used for outputs</td>
</tr>
</tbody>
</table>
4. Research Potential

Investigations have confirmed the potential of the data to support policy-relevant research. Preliminary research has produced for the first time information on the dynamics and earnings outcomes for government beneficiaries moving into and out of employment, filling a critical information gap. This work provides a first step towards producing official indicators of stable and sustainable employment. This research therefore has a direct outcome in relation to improving the scope of official employment statistics.

The datasets created have the potential to provide further information on a wide range of employment and firm-based policy questions. With further development, this work should lead to a range of official statistics that: will:

- assist in developing a picture of people’s dependence on benefits
- provide information on the employment/earnings outcomes of a range of government social assistance systems, and allow for more technical evaluation of proposed change
- provide further information on firm performance and outcomes.

The development work on these statistics will benefit from a range of base and technical research into questions and issues inherent in these areas. The current research has highlighted the benefit of high-level input from technical experts that reside in other organisations and it is anticipated that this need will continue.

One of the objectives of this investigation has been to identify specific strengths and weaknesses in the data when used to undertake policy research. The team noted the following points in relation to the current work.

Strengths:

- Universal coverage of benefit recipients (analysis could be broadened to incorporate other types of assistance ie. ACC outcomes).
- Ability to undertake longitudinal analysis across employment/benefit boundaries – facilitates cohort/flow analysis.
- Ability to support measures of duration in particular states.
- Ability to identify joint employment earnings/benefit receipt.
- Ability to answer a range of questions around earnings of those that leave a state of dependency, i.e., do they have higher incomes or not?
- Ability to progress, for the first time, the development of stable employment outcome indicators – including actual earnings outcomes, multiple job holding and the patterns of movement through subsequent employment.
- Ability to provide a basis for moving beyond this into sustainable employment outcome indicators.
- Technical expertise - the information requires a high level of technical competence. The project has assisted in developing public sector knowledge in this area. Staff will carry this knowledge with them back to their original organisations.

3. To enable this work to be progressed SNZ seconded two researchers from DoL’s Labour Market Policy Group. Senior research staff from The Treasury and DoL provided expert peer review and input. As a consequence of interim data issues and available time, the research results are illustrative of the potential information value that can be derived from LEED rather than authoritative statistics. Many of the base data issues identified will effectively be eliminated once the proposed cleaning and repair processes have been implemented.
Other and wider benefits
These have included:

- Building public sector capability in regard to the LEED infrastructure – several organisations have gained knowledge of the workings of the LEED infrastructure and the strengths and weaknesses of the infrastructure, including data access, security, privacy and confidentiality provisions.
- Highlighting the gains that can come from cooperative work –
  - broader testing of potential information and IT infrastructure
  - additional quality testing enabling the building of a wider infrastructure to meet multiple needs
  - additional subject-matter expertise available to SNZ.

Limitations of the current work

- The bundling of benefit types (e.g., unemployment, sickness, invalids, widows, DPB) that appear heterogeneous in relation to their employment/benefit behaviour has impacted on this initial research.
- 'Raw' status of the base data – this is an interim issue but one that will require further work.
- Absence of important socio-demographic variables such as ethnicity.
- The highlighting of the existing IT system's inability to meet some technical research demands.

Next steps for the current work

There appear to be a number of relevant areas for further investigation within the current work, including:

- The investigation of detailed industry and regional labour market indicators of stable employment outcomes. This level of detail potentially available within the LEED infrastructure is often important for regional planners and policy makers when advising decision-makers on policy change.
- An investigation of firm-specific versus individual-specific effects on employment outcomes and the income of individuals.
- Work into how low income earners escape low paid jobs; that is, to escape low pay do individuals change jobs, locations or industries? Or do a number of people escape low pay through pay increases (or increased hours of work) from their current employer?

Further avenues for research\(^4\) appear to be quite broad and potentially could include:

- Further investigation of firm dynamics and firm growth (building on the current LEED output) aimed at developing more reliable official statistical measures of firm and economic performance and productivity.
- Development of official measures of labour input volumes. In order to produce reliable productivity measures, it is important to have high quality, timely labour volume measures. LEED data may in the future be able to supplement current employment measures to provide more accurate and timely labour input measures that, through the BF link, could be made consistent with the National Accounts framework.

\(^4\) Avenues for broader research will always need to be assessed on their individual merits and associated risks/costs.
5. Confidentiality

The longitudinal linking of employer and employee data presents new confidentiality challenges. These challenges arise from the universal nature of the dataset created and the subsequent ability to track both business and individual characteristics over time.

Despite these concerns, existing practices such as the random rounding of counts will provide a high level of protection for individual data in official statistical outputs.

The protection of business data in a small economy such as New Zealand is more problematic. Existing practices, such as the application of cell suppression for business magnitude data, are in theory applicable in this situation. However, they are likely to require unsustainable levels of manual input and result in the unnecessary suppression of significant volumes of valuable data.

New practices are now emerging internationally that confidentialise information through the addition of statistical ‘noise’ to unit record data. These techniques appear robust and efficient in protecting both standard and customised outputs. This approach has been approved by the United States Bureau of the Census Disclosure Review Board and has been adopted by the USCB LEHD team to successfully protect similar data. SNZ specialists are currently reviewing these practices in a New Zealand context.

It is proposed that detailed work be undertaken during the next six months on the application of these techniques to ensure the confidentiality of any LEED outputs.

6. Data Security

SNZ and IRD have a data supply agreement that governs the transfer of data for statistical purposes to SNZ. This process is undertaken under section 81(4)(d) of the Tax Administration Act 1994. In Stage 1, a one-off data transfer process was employed. As agreed with IRD, the data was subject to a high level of encryption and transferred by secure courier. The key to unencrypt the data has been physically secured and is released only for specific use. Should the project move to production, a formal assessment of the data specification and terms and conditions of supply will be undertaken, and a supply arrangement agreed upon.

SNZ's standard security measures and protocols govern the management of the LEED data. In terms of the requirements for maintaining confidentiality and privacy of information, SNZ is required to comply with the confidentiality provisions of the Statistics Act and also the Security in the Government Sector (SIGS) protocols. To this end, SNZ operates a ‘need to know’ policy, which prohibits any staff, other than those actually working on a particular survey or process, from seeing confidential information.

Additional security arrangements for the LEED data include:

- all servers and backup tapes being held under SNZ’s highest level of physical security
- all LEED data directories and associated workspaces being secured on a server dedicated to integration (access is only authorised for those project personnel who need to access data for specific tasks and to selected IT administrators who are required to maintain the IT system)
- the date/time/user and mode of access (read/write) to all datasets in the LEED directories being logged
- no data being copied or stored in non-LEED directories or servers
- all LEED-related network traffic being encrypted when this facility is implemented on the SNZ WAN, and
• a quarterly audit being undertaken by the SNZ Security Office to ensure compliance is maintained.

These extra security settings were established after a review of the requirements for protecting the data to be used in this project.

7. Access

It is likely that some individuals will have concerns regarding the use of information about them in projects such as LEED, including that:

• information might be used in a manner that is detrimental to their personal circumstances
• information might be released that identifies them and aspects of their personal affairs
• unrelated information might be collected about them in an ever-growing database for non-specific purposes.

SNZ has addressed these concerns by ensuring that:

• the data is only used for statistical purposes
• direct identifiers are removed as early as possible during processing
• no identifiable information will be released
• the development meets the requirements of the Privacy Act, and
• formal authorisation is required, including consultation with the Privacy Commissioner, before additional data can be linked to the current framework.

Outputs from LEED will be in one of two forms:

1. Published official statistics
2. Statistical research results drawn from anonymised unit record datasets.

The production of official statistics will be undertaken in accordance with standard SNZ production and publication processes. No issues of concern appear to exist relating to these outputs.

Access to micro record data for bona fide statistical research is seen as an important component of the programme. It is apparent from the initial research findings and discussions with potential users that a significant proportion of the value derivable from LEED arises from such research. Section 37c of the Statistics Act provides the Government Statistician with authority to provide access for approved government researchers to unidentified unit record data. The potential for access by private researchers is more constrained and can in some cases be provided if the research contributes directly to the work of SNZ.

As data steward of the base LEED data, IRD has agreed that SNZ will manage requests for microdata access in terms of the provisions of the Statistics Act. However, IRD has also noted that, while such access may not be unlawful under the Tax Administration Act, the Act does require the secrecy of tax data in all but a few circumstances. SNZ has one of the few exemptions to a general prohibition on tax data being provided to other parties. Further work is required to clarify access arrangements beyond SNZ, ie Data Laboratory access that will meet IRD’s legislative requirements. At present any microdata access requests will be forwarded to IRD for comment prior to approval.
Following the McLennan review, SNZ has been working to enhance its existing processes to provide effective and safe microdata access mechanisms. This involves ensuring the protection of individual and business privacy and use. SNZ’s objective is to achieve the best outcome for both the users and suppliers of the statistics. This means:

- allowing the use of microdata in appropriate ways to meet justified research demands, while
- protecting the supply of information from respondents, and
- complying with relevant legislation.

SNZ is developing a range of options based on best international practice that would allow microdata to be used to meet research needs while protecting the confidentiality of the data. For data such as LEED, new techniques involving synthetic datasets, remote access, etc are being explored. The combination of business and individual records on a longitudinal basis in a small country such as New Zealand presents confidentiality challenges that mean it is unlikely that any unit record dataset will be able to be released outside SNZ. This may be inconvenient to some researchers but will nevertheless provide the potential for controlled access ensuring that extra value is derived from use of the data.

Users and data suppliers are seeking clarification of the boundaries and processes that will govern LEED microdata access before committing to support the development. The revision of policies and processes governing microdata access is a corporate issue, which SNZ is progressing. This process will include consultation with data suppliers and users before a final strategy is determined.

8. Data Quality Issues

8.1. IRD data

In order to administer the tax Acts, IRD collects a range of data related to employment activity in New Zealand. This information provides the backbone of the prototype LEED dataset. The IRD collects this information to support the efficient operation of the New Zealand taxation system. The use of this information as a base for the production of official statistics places new and quite different demands on this data, both in terms of the variables of interest, and its alignment with relevant statistical concepts. Any comments made relate to this new use of the data and should not be taken as indicative of any inability of the data to meet its primary operational function.

Input data was sourced from the IRD to allow the investigations specified in the Stage 1 work plan to be completed. The data relates to a three-year period from April 1999 to March 2002. Prior to April 1999, employers did not provide individual employee data to the IRD on a monthly basis. Therefore the series cannot be produced before this point.

The data obtained was in the form of identified unit records drawn from monthly payroll, annual income tax and registration files. This included information on individuals:

- names, addresses, dates of birth, IRD numbers, resident indicator (NZ or overseas)
- taxable earnings (plus earnings not liable, and lump sum indicator) for work performed including social security payments taxed at source
- tax deductions (PAYE, withholding tax, family support tax credit, student loan indicator and amount)
- employment start and finish dates.
This data forms a complete (excluding undeclared income), longitudinal dataset, detailing employee labour market earnings over the reference period. The annual income tax returns provide a similar record set for the self-employed. As the main social security benefits are also taxed at source, monthly information on individual benefit recipients is also contained in this data.

To date, investigations have focused on the EMS data and supporting client information. The EMS records monthly payments by employers to each of their employees. Detailed investigation of the annual returns relating to the self-employed has been deferred to Stage 2.

8.1.1. Major findings

In Stage 1 of the project, the team has investigated the quality and characteristics of the base data in relation to anticipated statistical use:

- The EMS and supporting IRD data has proven to be of high quality and, in most instances, the collection variables support, or will support, following cleaning and repair, the statistical measures proposed.
- Where necessary, imputed values can be produced with confidence. Remaining inaccuracies will not significantly affect the proposed output statistics.
- While detailed work has not been completed on the self-employed income data, it appears likely that the annual returns will be of similar quality and allow for the employer-employee dataset to be augmented by a comprehensive annual measure of self-employment and earnings.
- The use of complete, rather than sample, data is required to support the proposed outputs and the anticipated use of this data (ie to support reporting on the longitudinal dynamics of employment and business performance, regional estimates, potential links to sample survey data or other administrative sources, and the possible identification of sub-populations for targeted surveys).
- By SNZ current standards, the base IRD dataset is large (in total the three-year extract comprises about 217 million records with an estimated 50 million new records or updates required per annum).

The strength of the LEED data in general and of the EMS in particular is in its coverage. Every taxpayer that pays tax at source on income related to labour input is captured. Being effectively a census, outputs can potentially be produced for very small groups of interest without introducing sample error. Each employer and employee carries a unique identifier, and the two are linked. The identifiers are generally stable over time, which means that the data is longitudinal, ie the data enables a dynamic picture of the relationship between employers and employees to be developed.

Despite the overall high quality of the base data, significant cleaning and transformation will be required before robust official statistical series can be produced from the ‘raw’ data.
8.1.2. IRD data issues

In some areas, the use of administrative data is limited because gaps or inconsistencies exist between the ideal statistical concepts and the source data. While some inconsistencies do exist for employment measures, none is considered significant enough to undermine the validity of statistical measures based on this data. The following are the most significant of these issues:

- The data contains a number of critical links. In general these are of a high quality although repair and cleaning is still required. This is a non-trivial set of tasks, which is reported on separately in section 9 below.

- While the IRD and SNZ make similar distinctions between employees and the self-employed, the distinction is not easily applied in statistical measures, given the different reporting arrangements for the two groups within the IRD data, ie monthly EMS and annual IR3. In addition, to overcome the time lag inherent in the filing of IR3 returns, it will be necessary to investigate the potential to develop an ongoing activity indicator (and possibly model earnings) from returns of GST turnover.

- The EMS captures all persons engaged in paid employment in the month covered by the return, regardless of the timing, or the length of that employment. In contrast, many official labour market statistics are based on employment at a specific point in time (often a specific pay period in a quarter). To the extent that employees change jobs during the EMS reference period (one month), they will be counted as having two jobs during that period, which will upwardly bias both the number of jobs, and the number of multiple job holders. A prototype method to address this issue has been developed but has yet to be implemented.

- About 1 percent of EMS employee records for a month currently records earnings of less than $8 (the adult minimum wage). It appears some of these entries may not relate to current employment. Further investigation is required to establish a minimum earnings cut-off.
### 8.1.3. Cleaning of the IRD data for statistical purposes

The following table summarises work undertaken to derive or clean key output variables:

**Table 3**  
**Fitness for Use of Raw and Cleaned Data**  
(excludes linking issues – see section 8)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Fitness for use of the raw data for statistical analysis</th>
<th>Need and feasibility of cleaning.</th>
<th>Fitness for use of cleaned data for statistical analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>This field does not exist in the IRD tables.</td>
<td>This is a critical output variable. It can be derived from sex-specific titles or first names from IRD tables.</td>
<td>98.5% derived from title, 1.5% derived from first name.</td>
</tr>
<tr>
<td>Date of birth</td>
<td>This field exists in the IRD tables. Less than 4% of employee IRD numbers have missing date of birth.</td>
<td>This is a critical output variable. Missing values can be filled using nearest neighbour imputation.</td>
<td>96.3% from IRD records, 3.7% imputed</td>
</tr>
<tr>
<td>Start and finish dates of jobs</td>
<td>This field exists in the IRD tables, but is not well used.</td>
<td>These dates are essential for the production of point in time employment statistics. It can be randomly imputed within the first or last month of non-zero payment.</td>
<td>50% of end dates are as supplied on the EMS. All other dates have been imputed avoiding unnecessary overlaps of jobs and periods of unemployment.</td>
</tr>
<tr>
<td>Gross earnings</td>
<td>1,874 (0.002%) records of non-zero gross earnings have a detectable error when compared with the PAYE amount and the individual's earnings history</td>
<td>This is a critical output variable. Unlikely values can be replaced with values imputed by applying a typical rate to the PAYE amount.</td>
<td>All detected errors have been replaced with statistically consistent values.</td>
</tr>
<tr>
<td>Addresses of individuals</td>
<td>Addresses are time-specific, and there may be several categories of address active for individuals at any one time. Most addresses are coded to NZ Post Code. IRD have recently established an automated individual address updating process based on NZ Post address maintenance - we are investigating the quality of this address maintenance.</td>
<td>Demand is anticipated for regional outputs to as fine a standard level as possible (subject to privacy &amp; confidentiality). A history series of &quot;best&quot; addresses for all employees has been produced. Testing is in progress to convert these to a geographical classification, possibly as fine as Urban Area.</td>
<td>To be tested</td>
</tr>
</tbody>
</table>

In future, additional firm characteristics drawn from the BF will be incorporated, such as industry and institutional sector. No information exists within the IRD system that would enable the derivation of individuals’ ethnicity or other important socio-economic demographic variables (eg qualification, occupation). If sufficient need for this variable exists it may be technically possible to
establish a link to existing survey sources (eg HLFS) or other administrative sources (MSD) to provide at least partial coverage. Any additional linkages to LEED require executive approval. As a consequence no further work is currently planned on this issue.

8.2. SNZ Business Frame data

8.2.1. The Business Frame's ability to support the demands of LEED

Investigations have shown that the SNZ Business Frame (BF) will support the main requirements of the LEED project for a comprehensive longitudinal frame. In conjunction with the BF section, LEED has been able to realise the inherent potential contained in the existing BF history tables to produce a robust longitudinal Business Frame (LBF) series at geographic (Geo) and enterprise (Ent) unit level.

The new LBF, with its economy-wide coverage and basic data items such as employment, location, industrial activity and firm ownership relationships, has many potential uses including:

- as a base for an enhanced measure of business demography (eg birth/death rates, enterprise survival and age profiles, etc)
- to identify and track the components of enterprises involved in structural change
- as a frame for linking LEED and other business surveys, such as the Annual Enterprise Survey, allowing the use of business survey data longitudinally for productivity and performance measurement.

To support the LBF, the BF section should be funded to undertake a systematic review of the current BF updating and recording practices in relation to, the history tables, transferring Geos, and other fields useful for data integration. Current practices could then be refined to ensure the system is explicitly tuned to support longitudinal linking. Our findings suggest that this should not be a major task.

Recent extensions to frame coverage have moved the BF closer to its target population. Despite this, a large number of very small, self-employed businesses will still be excluded under the proposed coverage rules. It is possible that the extension of coverage to include these micro businesses may be useful from a user perspective. The costs and benefits of creating a so-called 4th Tier on the BF could be investigated if sufficient user interest exists.

When fully developed, LEED processes will provide a largely automated and systematic process for the identification of duplicate units birthed to the frame. The results of this process will identify so-called predecessor/successor units among BF deaths and births providing a more accurate picture of business survival and growth.

While detailed work has yet to be undertaken on the implications of moving from a static to dynamic mode, it is likely that the current updating processes employed for the BF will result in the tail of the LBF series being subject to revision. The extent of this updating and its implications for the proposed output series has yet to be tested. Despite this uncertainty, it appears preferable to incorporate the best available information into the core LEED framework, setting revisions practices for output series independently, based on the likely impact of different revision practices and the needs of users.
9. Unit Record Links

9.1. Description of links

There are seven unit record links within the LEED dataset. Two sources of data are brought together – the tax administration data from IRD, and SNZ Business Frame data. Two different entities are involved – businesses and people. Links over time exist within each source and between these two entities. The arrows in figure 1 show these links.

**Figure 1**

*Unit Record Links in LEED Data*

The link between the tax and BF data is:
- between the same business at enterprise level in a given time period (link 1).

Links within the tax data are:
- between a person (ie employee) and business (ie employer) in a given time period (link 2)
- for the same person over time (link 3)
- for the same business over time (link 4).

Links within the BF data are:
- for the same business over time at enterprise level (link 5)
- for the same business over time at Geo level (link 6)
- between an enterprise and its Geos in a given time period (link 7).

The Longitudinal Business Frame discussed above establishes the links within the BF data.
9.2. Linking Issues

The IRD number issued by IRD to businesses and individuals is a high quality numeric identifier available on both the tax data and the BF. In the first instance, an exact match on IRD number will create the links in all the cases listed above. However, a number of problems arise if we rely solely on the reported IRD number, which will have significant impact on some of the key outputs and analysis. The main issues are:

1. **Group returns** – A group return occurs where a tax employer reports for a number of businesses under one IRD number. In this situation, the reporting unit and statistical unit will not align and an incorrect link will be made to a BF enterprise (estimated to affect about 500–600 enterprises, and about 100,000 jobs – about 70,000 of these jobs are covered by a single return).

2. **Employees to geographic unit** – The integration of the tax data with the LBF means that the employees are linked to the corresponding enterprise on the BF. Because the employer on the tax side is normally an enterprise, it is not possible to link employees directly to their geographic unit for multi-geo enterprises. The ability to provide regional breakdowns will be restricted until a method for assigning a Geo to employees of multi-Geo enterprises is developed. A model that assigns a Geo (and thus an enterprise) to an employee may also best solve the group returns problem.

3. **True births and deaths** – There is widespread interest in comparing the characteristics of new businesses (births) and businesses which cease operation (deaths) compared with continuing businesses. Births appear in the data as new IRD numbers, and deaths as an IRD number that is no longer used. There are a number of reasons (such as a change in legal status or change in ownership) that mean a continuing business can be issued with a new IRD number and stop using the old one. These changes in IRD number for purely administrative reasons inflate the number of births and deaths, and consequently upwardly bias business demographic and job and worker flow statistics. The absolute number of businesses wrongly classed is small in relation to the total number of businesses, but because there are also only a small number of births and deaths, the bias is likely to be significant.

We have identified some of the spurious births and deaths by following groups of common employees in the tax data, and by using information from surveys already held by the BF. We plan to further that work by using probabilistic matching methods that can compare other information known about businesses such as name and address. From work to date, we conclude that the impact is large. A conservative estimate is that over 20 percent of deaths are due to this administrative churn. This level of error would seriously affect any analysis of new and ceased firms, add about 3-4 percent to job and worker flow rates, and contribute 40 percent of job and worker flow rates due to deaths.

4. **Errors in employee IRD number** – There are 2 percent of employee records where the employee IRD number is missing and a smaller number of records where the employee IRD number is recorded incorrectly. In cross-sectional analysis, this level of error would often be a minor nuisance, but we are using the employee IRD number to create longitudinal links. Similarly to the birth and death issue, the absolute number of records involved at any particular time is small, but even a small number of missing or incorrectly identified records can have serious consequences for longitudinal analysis. Where an employee IRD number is missing or wrongly recorded, either a record for a person is missing, creating a break in their employment history, or an additional record is added to their employment history. False transitions are created, and measures of flows are increased. We expect that the effect would
be greatest on smaller flow measures and on analysis focussed on a full employment history. Work to date investigating incorrect employee IRD numbers suggests that around 10–15 percent of individuals have one or more records with errors. We are currently working on using probabilistic matching methods that compare employee name, earnings and job histories to identify and correct errors.

Each of the issues above involves work in specialised statistical areas for which mathematical statisticians are needed who have the ability to apply probabilistic matching techniques. The mathematical theory is well developed, but we must address the problems posed by the new dataset we are creating for LEED.

Initial development work is being carried out on small samples of data. The methods developed then need to be applied to the full dataset and an analysis of the effects on key statistical outputs carried out. No doubt there will still be some problems remaining before we can escalate to a full production environment.

Experience at the US Bureau of the Census and progress already made by the LEED project team leaves us confident that practical methods for reducing these biases can be developed.

10. Privacy

10.1. Privacy assessment

In most instances data linking will increase the information that is available about an individual or business beyond that which they might be aware of. Other privacy risks associated with data integration projects may also exist, such as the security of stored data. Careful management of the information and linking processes is needed to ensure that privacy is protected and the security of the information is maintained while achieving the benefit from use of the data.

The information created by the LEED project will be in the form of unidentified unit record datasets that will be used to produce official statistics and support statistical research. No information will be published from the datasets that would allow for the identification of any individual or business.

A Privacy Impact Assessment (PIA) has been prepared for this project in accordance with SNZ's data integration protocol requirements. Over the last two years, SNZ has held ongoing discussions with the Office of the Privacy Commissioner regarding appropriate safeguards and processes to govern data linking for statistical purposes. The risks to individual privacy arising from the LEED project have been identified and processes developed that will ensure that privacy is not jeopardised.

The PIA found that the current project:

- complies with Cabinet decision [CAB (97) M 31/14] regarding cross-agency statistical data integration using data for unrelated purposes
- complies with the provisions of the Statistics Act 1975, ensuring among other things that information furnished to the Department is kept securely, access is restricted (on a 'need to know basis') and any publication avoids disclosing identifiable information
- complies with the provisions of the Privacy Act 1993, ensuring that privacy concerns of individuals and businesses are not set aside for the potential value/perceived benefits of record linking
• does not link any information relating to individuals that was not already held in a linked form in the IRD system
• does result in additional privacy risks, particularly as the data is transferred from IRD to SNZ and during the period the data is temporarily identified during processing and cleaning; however, these risks have been managed to the lowest possible level
• involves the potential risk of negative public perception. The openness, appropriate use and transparency of the processes followed and appropriate consultation have mitigated this risk. Credible responses to public concerns are available if required.

10.2. Potential future privacy issues

It is likely that future development of the LEED infrastructure will involve proposals to link additional administrative or survey data on individuals to the existing framework. Any proposal will require explicit approval by both the Government Statistician and the chief executive of the agencies from which the data was sourced. A further PIA would also be required. The Government Statistician has determined that the default setting for any additional link would be ‘temporary’. The link would only be established and retained to allow the completion of the specific purpose for which approval was given. There may be instances where an additional link adds ongoing value to the core framework and outputs. The retention of any additional link on an ongoing basis would require specific approval and would be subject to an ongoing evaluation as to whether the link should continue to be maintained.

11. Information Technology Issues

The project has been well supported by SNZ Information Management Services (IMS) during Stage 1. However, the data testing and evaluation undertaken by the data integration projects (including injury statistics) have revealed IT constraints that will need to be resolved as part of a generic system solution. IMS specialists have concluded that the existing IT systems will not support LEED requirements in a production environment.

SNZ has progressively been increasing its reliance on administratively-sourced data in a wide range of areas in keeping with a strategic push to lower respondent load for business and an increase in the quality of outputs. The data integration projects requirements are a further stage in this process, and are unique in terms of the data volumes involved and complexity of processing undertaken.

Current performance issues have been managed by implementing short-term upgrades consolidating data management on a dedicated server. A range of medium to long-term options is now being considered. IMS and external warehousing expertise have been engaged to assess the IT architecture options available to support these projects for the next 3–5 years on a production footing. This work will be completed by 31 October 2003.
12. Remaining Project Risks

Table 4

<table>
<thead>
<tr>
<th>Risk</th>
<th>Management</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absence of baseline funding.</td>
<td>A business case and new initiative funding bid is being prepared for 2004/05.</td>
</tr>
<tr>
<td>Uncertainty remains amongst users and potential sponsors regarding the feasibility of and ground rules for microdata access.</td>
<td>This issue is likely to require the development and implementation of new access methods, ie synthetic data, remote access, in-house research team etc. Consultation will be required with both data stewards and users to develop an effective solution.</td>
</tr>
<tr>
<td>Development of critical mass of expertise in new and complex development areas:</td>
<td>Capability is being fostered in both Wellington and Christchurch MAIS teams. Funding is being sought to develop subject area capability.</td>
</tr>
<tr>
<td>• linking (new, complex - learning new skills while progressing task - slows progress, increases risk)</td>
<td></td>
</tr>
<tr>
<td>• confidentiality of output data (longitudinal, census of both businesses and people)</td>
<td></td>
</tr>
<tr>
<td>• subject-matter expertise in new outputs</td>
<td></td>
</tr>
<tr>
<td>• longitudinal analysis techniques</td>
<td></td>
</tr>
<tr>
<td>• existing momentum vulnerable to key staff turnover.</td>
<td></td>
</tr>
<tr>
<td>Transition risks may arise when transferring development to Wellington.</td>
<td>May result in a temporary loss of momentum as new staff are recruited and gain understanding of programme. Potential to mitigate by retaining overlap with Christchurch-based team and progressive handover.</td>
</tr>
</tbody>
</table>

13. Conclusion

LEED has the potential to make the single most significant advance in recent in statistical information. In doing so, it will make use of existing information and make no or minimal demands on respondents for additional information. It therefore fits well with the strategic goals of SNZ and government.

The creation of a LEED database is feasible. While significant work is required to transform the base data from an administrative to a statistical dataset, work to date confirms that the data can be cleaned to a standard where it will support the production of official outputs and support policy research.

Potential exists for subject area divisions within SNZ to extract further value from the LEED framework. To achieve this potential, it is likely that additional survey or administrative data will need to be linked to the core LEED framework. An overarching strategic development programme is required to ensure that a logical framework is established that targets future development to areas that provide the maximum contribution to SNZ long-term goals. This requires agreement at executive level of the relative prioritise for different development options.
14. Next Steps

If approval to progress is given the project will undertake the following steps:

- A business case including full costings will be prepared seeking further funding by 30 November 2003.
- The existing development team will complete the development and testing of the proposed methods and implement these in the test environment. Prototype outputs will be rerun and released for review.
- A transition plan will be prepared to establish the programme in SNZ’s Wellington office in the first quarter of the 2004/05 financial year.
- This team would be responsible for the establishment of a production system and release of the first official statistical outputs. Target date mid to late 2005.