Record Linkage in LEED
An Overview

Christine Bycroft
Statistics New Zealand

30 September 2003
Contents

1. Introduction .................................................................................................................................. 3
   1.1 LEED (Linked Employer-Employee Data) ........................................................................... 3
   1.2 Record linkage concepts .................................................................................................. 3
   1.3 Links in LEED data ..................................................................................................... 4
   1.4 Linking methods .......................................................................................................... 5

2. Linking Businesses .................................................................................................................... 6
   2.1 Link from tax employer to BF enterprise ......................................................................... 6
   2.2 Linking EMS employers over time ............................................................................... 9
   2.3 Linking BF enterprises and geos over time .................................................................. 10

3. Linking EMS employees over time ........................................................................................ 11

4. Putting it all together .............................................................................................................. 12

Appendix Linking employers on the EMS to enterprises on the BF: The logical flow .......... 14
   Case 1: The exact match is correct ....................................................................................... 14
   Case 2: Predecessor/successor on EMS ............................................................................ 15
   Case 3: Links where a change in IRD number occurs at different times – predecessor and
   successor recorded on BF ................................................................................................. 15
   Cases 4 and 5: Employing on EMS, but not birthed (or already ceased) on BF............. 16

References ..................................................................................................................................... 16

List of Figures

1. Unit record links in LEED data............................................................................................... 4
2. Results from an exact match on IRD number between EMS employer and BF
   enterprise (illustrative only – not to scale) ......................................................................... 7
3. Predecessor and Successor Fields for Continuing Businesses on the EMS ...................... 10
4. Predecessor and Successor Fields Identifying Continuing Businesses on the BF ............ 11
5. Work Flow for Creating Integrated LEED Data .................................................................. 13
6. Links Where the Exact Match is Correct .............................................................................. 14
7. Links Where a Change in IRD Number Occurs at Different Times – EMS Records
   Predecessor and Successor Recorded on EMS ................................................................ 15
8. Links Where a Change in IRD Number Occurs at Different Times – BF Records
   Predecessor and Successor Recorded on BF .................................................................... 15
9. Links Where Birth Occurs First on EMS, or Death Occurs First on BF ......................... 16
1. Introduction

1.1 LEED (Linked Employer-Employee Data)

LEED is a data integration project that uses unit record linkage to establish relationships between entities. The resulting LEED datasets provide a rich and powerful source of information on business and employment dynamics. This richness arises in part from the variety of links available within the combined data. The number of different links also means that the record linkage process itself is complex.

The accuracy of these links can have a significant impact on the quality of any derived statistics. As a consequence, it is important to have investigated and, if necessary, repaired links within the base data to a standard that supports the intended use of this information.

This paper provides a brief introduction to the concepts and practices employed in record linking. It then outlines the main links within LEED and the issues associated with these links. This is one of a series of technical papers that have been written to support the project's Stage 1 Feasibility Report. More information on the project's scope and objectives are available in these associated papers.

1.2 Record linkage concepts

Gill (2001) provides a good introduction to the process of record linkage. He states (at section 1.1, p 13):

The term record linkage is simply the bringing together of information from two different records that are believed to belong to the same person, family or entity. The records to be compared may come from a single data file or multiple data files. They may relate to persons or to other entities, such as households, business establishments or companies. If the two records agree on all variables, and are unlikely to have done so by accident, the level of assurance that the link is correct will be high. Conversely, if most of the variables disagree there will be little doubt that the linkage is wrong. For intermediate situations, the methodology must predict whether the link is true or false.

The two most commonly used methods for record matching are the exact, or deterministic, method and the probabilistic method. Exact methods rely on the exact comparison of common variables – records either match if the comparison agrees exactly, or don't match. The main requirement for exact matching is the presence of a variable that is ideally (i) universally available, (ii) fixed, (iii) easily recorded, (iv) unique to that individual, and (v) readily verifiable. (Gill, 2001, section 3.1, p 27).

Probabilistic methods use a combination of partial identifiers, for example, names and addresses, which are used to compute scores called ‘weights’ for each potential match based on probabilities. The resulting weight is then compared with a preset threshold to decide whether to mark this pair as a true match. (Gill, 2001, section 3, p 27).

Any linking method can result in two sorts of errors. First, it is possible to link records that in reality are not the same person/business. In this case, the match is a false one, called a 'false positive match'. Second, it is possible to not link two records that do in fact belong to the same

1. This paper has been compiled based on the information held by the development team at the time of writing. As such any results are to be regarded as provisional in nature and subject to possible revision.
person/business. In this case, we have not been able to find the match that does exist, and the error is called a 'false negative match'. 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.

1.3 Links in LEED data

There are two sources of data for LEED – the tax administration data supplied by the Inland Revenue Department (IRD), and Business Frame (BF) data from SNZ. Two different entities are involved – businesses and people. There are links over time that are formed within each source. And finally there is a link between each source. These are shown by the arrows in figure 1.

**Figure 1**

<table>
<thead>
<tr>
<th>IRD data - tax records</th>
<th>SNZ data - Business Frame</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Time 1</strong></td>
<td></td>
</tr>
<tr>
<td>Employee A</td>
<td>Employer Z</td>
</tr>
<tr>
<td>Employee B</td>
<td></td>
</tr>
<tr>
<td><strong>Time 2</strong></td>
<td></td>
</tr>
<tr>
<td>Employee B</td>
<td>Employer Z</td>
</tr>
<tr>
<td>Employee C</td>
<td></td>
</tr>
<tr>
<td><strong>Time 3</strong></td>
<td>Employee Z</td>
</tr>
<tr>
<td>Employee C</td>
<td></td>
</tr>
<tr>
<td>Employee D</td>
<td></td>
</tr>
</tbody>
</table>

The link between the tax and BF data is:

(1) between the same business at enterprise level at a given time period.

Links within the tax data are:

(2) between a person (ie employee) and business (ie employer) at a given time period

(3) for the same person over time

(4) for the same business over time.

Links within the BF data are:

(5) for the same business over time at enterprise level

(6) for the same business over time at geo level

(7) between an enterprise and its geos at a given time period.

2. An enterprise is a business or service entity operating in New Zealand, including a company, partnership, trust, estate, incorporated society, producer board, local or central government organisation, voluntary organisation or self-employed individual.

3. A geographic unit (geo) is a separate operating unit engaged in New Zealand in one, or predominately one, kind of economic activity from a single physical location or base.
The three links involving the business – between the tax data and the BF, over time within the tax data, and over time for enterprises within the BF – should ideally be consistent with each other.

1.4 Linking methods

All people and businesses that register for tax purposes with the IRD are issued with an IRD number. For business entities, this number is transferred from the tax registration system to the corresponding BF unit. This IRD number provides a unique numeric identifier common to both sources, which can be used to establish the different types of unit record links. Fortunately, the IRD number is a very good numeric identifier as it is issued and maintained to a high standard by the IRD and largely meets the criteria for use as an exact matching variable in this setting.

However, as with all identifiers, there are problems, including missing values and incorrect coding of IRD numbers. For businesses there are more complications. Two different IRD numbers can refer to the same business where the IRD number is changed for purely administrative reasons (either within a data source at different times, or between the two data sources). Also, the same IRD number can refer to different businesses where there are different reporting units in the tax system and the BF. This means that simply using an exact match on an IRD number would result in the creation of some false links, and would miss some correct links.

For each of the links listed below we are developing methods for improving the quality of the match that would be obtained by simply using an exact match on the IRD number as originally supplied in the source data. Analysis of the source data supports the need to address these issues, as they will impact on the quality of proposed outputs. This is not to suggest that the data quality is not suitable for its primary purpose within the tax system, only that it requires enhancement to support the different demands of statistical use.

The critical links are:

- For the same business between the tax data and BF at the enterprise level in a given time period (the common identifier is employer IRD number) – link (1).
- For the same person over time within the tax data (the identifier is employee IRD number) – link (3).
- For the same business over time within the tax data (the identifier is employer IRD number) – link (4).
- For the same business over time within the BF, which has two parts – the enterprise level and geo level (the identifiers are employer IRD number, and enterprise number and geographic unit number assigned by the BF) – links (5) and (6).

The remaining two links do not need further work:

- The link between employer and employee on the tax system: This is established by the tax record structure where employee IRD number and employer IRD number are on the same record – link (2)
- The BF link between enterprise and geo: This is established by the BF and is assumed to be correct – link (6).

There is a common general approach on the part of the LEED project team to improving the quality of the links in LEED, and most of the methods involve the use of probabilistic matching techniques.
The following factors are considered in the process of probabilistic and deterministic matching:

- The type of link, e.g. within one data source or between two data sources.
- The impact of false positive matches (i.e., records that are linked, but should not be).
- The impact of false negative matches (i.e., records which should be linked, but aren't).
- The additional variables available for matching.

2. Linking Businesses

2.1 Link from tax employer to BF enterprise

The key link for businesses is the link between the tax data and Business Frame. The Employer Monthly Schedule (EMS) is filed by all employers who have employees paying tax deducted at source (PAYE). For Stage 1 of the LEED project, outputs are based on the population of EMS filers, but exclude most of those who are self-employed. The remainder of this paper deals only with EMS tax data, with any data for the self-employed removed. Including the self-employed in the linking raises further issues, which will be dealt with in Stage 2 of the LEED project.

For convenience, the EMS business entity is referred to as an 'employer', to distinguish it from the BF business entity, which conforms to the SNZ standard structure of enterprise and geographic unit. Note that 'employer' on the EMS also includes government departments, such as the Ministry of Social Development, which pay benefits that have tax deducted at source. EMS 'employees' also include beneficiaries that receive these payments. Usually the EMS reporting unit is equivalent to an enterprise, but the geographic unit structure is not derivable from the tax system.

The EMS employers are the population master file for the match between EMS employers and the BF. Matching EMS employers to the BF provides the standard business structure and classification variables for outputs. We link to BF enterprises (and through them to geographic units), and add the industrial classification codes (ANZSIC), sector and geography information. We could also potentially link to SNZ business surveys.

A false positive match occurs when an EMS employer has been linked to the wrong BF enterprise. As a result there may also be wrong values for BF variables and the wrong business structure, and there will be links to the wrong survey respondent. Where an employer on the EMS is not matched to the BF, no variables from the BF exist (though there may be industry and geographic variables assigned by the tax system, which can be used as a proxy) and there will be no link to any survey responses. The BF variables are used in table outputs, mainly as categories to provide comparisons, for example by industry and by region. Having missing data is preferable to providing outputs that use wrong values, of which users are unaware. The consequences of making a false positive match have been deemed by the LEED project participants to be more serious than having no match, and the priority is to minimise false matches.

Figure 2 shows the matches that will be made between the EMS and BF, as well as those that will be missed, when the matching method used is an exact match on IRD number.
Results from an exact match on IRD number between EMS employer and BF enterprise (illustrative only – not to scale)

- out of scope for BF
- in scope, but not available for BF
- timing difference with administrative change to IRD number
- EMS employing, ceased on BF
- EMS employing, not birthed on BF
- different IRD number on BF
- error in IRD number

- BF non-filer enterprise in EMS group return
- timing differences in change to IRD number
- birthed on BF, not yet employing
- different IRD number
- error in IRD number
- not an employer

- correct matches

plus

- EMS group return match to single BF enterprise
False positive matches using an exact match on IRD number

An exact match on an IRD number gives a match for around 90–95 percent of all employers on the EMS to BF units for a given month. Some of these are false matches, which can arise for two reasons:

1. one IRD number (or both) contains an error
2. both sources state the correct IRD number, but refer to different reporting units.

Errors in the employer IRD number in the tax data should be very rare because of the high quality control systems used by the IRD.

Different reporting units are a major concern. This occurs when several businesses report their tax as one unit, known as a group return. The BF units are enterprises, just one of which will have the same IRD number as used on the group return. An exact match on the IRD number will assign the whole group’s employees to one component of the group structure on the BF. This will usually mean that the wrong industry is assigned from the BF, and understating the job and worker flows at the enterprise level will be under-stated.

False negative matches using an exact match on IRD number

An EMS employer may not match to the BF for a number of reasons, including cases where the employer is not expected to be on the BF. There are six reasons for this:

1. The EMS employer does not meet the coverage criteria for BF inclusion, or is not available to the BF.
2. Error in the IRD number for either source (including missing values).
3. On the BF, the same reporting unit but a different IRD number on the EMS because the BF is using a GST number.
4. On the BF, the same reporting unit but a different (though correct) IRD number on the EMS because there have been timing differences in the change of IRD number.
5. Employing unit on the EMS, but already ceased on the BF.
6. Employing unit on the EMS, but not yet birthed on the BF.

Notes on these reasons:

1. Cases where we do not expect to find the EMS employer on the BF can be identified. An EMS employer will not be on the BF if it does not meet the in-scope coverage criteria for BF inclusion, or is in-scope, but not available to the BF because, for example, the address is missing.
2. Errors in an employer's IRD number in the tax data should be very rare because of the high quality control systems used by the IRD.
3. The IRD registers businesses for GST and PAYE, and sometimes issues a different IRD number for each. In a small number of cases, the BF may be using a GST number that will not match the PAYE number on the EMS.
4. The same reporting unit may change an IRD number from one period to the next for administrative reasons, for example a change of ownership. If the change does not occur at the same time on both sources, there will be no match for one or more periods. These matches can be made once the links over time are established, using predecessor and successor fields (see section 2.2). The link over time identifies continuing businesses, among those which appear to be births and deaths because of IRD number changes.
5 and 6. Once continuing businesses are matched, remaining births on the EMS can be searched for on the BF in future periods. Similarly, remaining deaths can be searched for on the BF in previous periods.

Group returns on the EMS are found on the BF as one of the contributing enterprises, unless they are missed for one of the other reasons listed above. However, the remaining enterprises on the BF, which are also part of the group, will have no match to an EMS employer.

BF enterprises without a match

There will be BF enterprises that are not matched to the EMS because the enterprise is not an employer, and therefore not on the EMS. Other BF enterprises should be, but are not, matched to an EMS employer, either because they are the missing part of the false negative match, or because the corresponding EMS employer is incorrectly matched to another BF enterprise. And as noted above, there are enterprises on the BF that have their tax reported as part of a group return; these enterprises will have no match to an EMS employer.

Details of work undertaken to address this issue will be reported in a separate paper.

2.2 Linking EMS employers over time

For longitudinal analysis, we need to follow the same employer over time. Normally, an employer reports their tax using the same IRD number each month so that creating the longitudinal link is simply a matter of matching on that employer's IRD number.

False positive match

A false positive match could occur if there were an error in the IRD number. However, errors are rare because the tax administration system is geared towards collecting tax from all employers, and it relies on the accuracy of the IRD number to do this.

False negative match

A firm birth (or death) is an economic concept that is said to occur when new factors of production are assembled from scratch (or broken up).

While an IRD number will be issued to new employers at birth, there are a number of instances in which the commencement or cessation of use of an employer IRD number will not tie in with a firm's birth or death. The longitudinal link is broken when an employer continues in business but reports its tax under a different IRD number. The new IRD number then appears as a new business (a birth), while the previous IRD number appears to be a business that has ceased employing (a death).

There is great interest in the performance and employment dynamics of new and ceased businesses, compared with continuing businesses. Even though the number of continuing businesses that appear as births and deaths may only be a small proportion of the total number of employers, this can significantly inflate the number of births and deaths, inflate job and worker flows and cause the incorrect identification of cohorts of births. Therefore, it is important to minimise the false negative matches in the longitudinal link, while avoiding the introduction of false positive matches.

A continuing business retains many of the same characteristics from one time period to the next, regardless of the IRD number change. This allows the development of methods that match deaths at one period to births at the succeeding period. We are working with records from the same file, and therefore have the same information available for the deaths and births. For the
EMS, we know who the employees are, as well as the business’s name, address, industry and tax codes.

When a match is found, the link is identified using predecessor and successor variables. A death has the IRD number of the matched birth recorded in the successor field; the birth has the IRD number of the matched death in the predecessor field. See figure 3 where emp1 in month2 is the same business as emp6 in month3. The predecessor and successor fields also enable a link to be made between the EMS and BF when the change in the IRD number does not occur at the same time for both sources.

**Figure 3**

**Predecessor and Successor Fields for Continuing Businesses on the EMS**

<table>
<thead>
<tr>
<th>ems_ird_nbr</th>
<th>period</th>
<th>pred</th>
<th>succ</th>
</tr>
</thead>
<tbody>
<tr>
<td>emp1</td>
<td>month2</td>
<td></td>
<td>emp6</td>
</tr>
<tr>
<td>emp2</td>
<td>month2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>emp3</td>
<td>month2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>....</td>
<td>....</td>
<td></td>
<td></td>
</tr>
<tr>
<td>**</td>
<td>**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>emp2</td>
<td>month3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>emp3</td>
<td>month3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>emp6</td>
<td>month3</td>
<td></td>
<td>emp1</td>
</tr>
</tbody>
</table>

emp6 is successor for emp1

emp1 disappears

emp1 is predecessor for emp6

The simplest case is when there is a one-to-one match between the unit that appears as a death and the unit that appears as a birth. For example, where an employer at a single geographic location starts to report tax under a different IRD number because of a change in legal status, or change in ownership.

When there is a change in IRD number for employers who report tax for several different geographic locations, there may be a simple one-to-one match between the new and old structure. However, when mergers, takeovers and breakouts are involved, the situation becomes more complex. There can be changes to the IRD number for continuing employers where the relationship between the IRD numbers for the previous and succeeding employers is one-to-many, many-to-one or many-to-many. It is not yet clear how these complex re-arrangements should be treated in a longitudinal sense. On the EMS, we will only attempt to identify continuing businesses where the match is one-to-one. The more complex situations require the ability to match individual geographic units that can only be identified on the BF.

Details of work undertaken to address this issue are reported in Kelly(2003)

2.3 Linking BF enterprises and geos over time

The BF series used for LEED was created from the BF and its history tables as a true longitudinal series. The normal BF is dynamic, and only holds the most up-to-date values. The new Longitudinal BF currently contains selected information about businesses for each month from April 1999 to March 2002 (see Seyb, 2003, for more details).

However, similar comments apply to the longitudinal linking of businesses on the BF side. Here we have the advantage of knowing the business structure, since enterprises and geographic units
are both available on the BF. A one-to-one match between a supposed death and birth can be made at geo level, and the more complicated issue of identifying continuing enterprises and classifying the type of continuity, can be addressed from knowledge of the geos. This ability to better distinguish between continuing businesses and real births and deaths on the BF is another attribute of an EMS employer that we gain from linking the EMS tax data to SNZ’s BF.

In fact, many continuing businesses with a changed IRD number are already identified on the BF, though the ability to do this depends on responses to surveys. Additional links can be made using probabilistic techniques, where a match is made between geo deaths in one period and births in the next period. A paper by Jarmin and Miranda (2002) describes a similar approach taken in the US for a longitudinal business register. The match found will be held in predecessor and successor fields (figure 4). The BF holds variables such as name, address and phone number for every geo. Again, the emphasis is on minimising false negative matches, since we want births and deaths to reflect reality, not administrative changes. At the same time we do not want to introduce false positive matches.

Figure 4

| Predecessor and Successor Fields Identifying Continuing Businesses on the BF |
|-------------------------------|-----------------|-----------|
| bf_ird_nbr | period | pred | succ | ent_nbr |
| bf_nbr2 | month4 | bf_nbr8 | ent2 |
| bf_nbr3 | month4 | ent3 |
| ... | ... | ... | ... |
| ** | month5 | ** | bf_nbr2 disappears |
| bf_nbr3 | month5 | ent3 |
| ... | ... | ... |
| bf_nbr8 | month5 | bf_nbr2 | ent8 | bf_nbr2 is predecessor to bf_nbr8 |

Note: The pred and succ information can be held equivalently by using the BF assigned ent_nbr.

3. Linking EMS employees over time

For the successful analysis of employee dynamics, we need to follow the same employee between different employers and over time. An employee reports their tax under the same IRD number each month, so creating the link is a matter of finding a match to that number.

The IRD’s system of issuing IRD numbers ensures that each person can have only one IRD number, and each IRD number is issued to only one person. However, the recording of the employee IRD number each month on the EMS return is subject to error. 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. When the correct value for an employee IRD number is not available, it appears as though the person is not working for that employer in that month. This leads to an increase in worker flow statistics, and can significantly affect other types of economic analysis. The methods proposed to improve on an exact match on employee IRD number as provided in the source data are based on those used by the Longitudinal Employer-Household Dynamics (LEHD) project at US Bureau of the Census (see Abowd and Vilhuber, 2002).

Probabilistic matching will be used to identify incorrectly recorded IRD numbers, using the names recorded each month on the EMS. This is a de-duplication exercise, where a name that is not a
duplicate of the other names given for an IRD number is assumed to contain an IRD number error.

The process then requires finding the correct employee IRD number for all those that are missing or found to be wrong. Where errors are occurring at random, they will largely be one-period job histories, known as 'plugs'. We use the fact that when the IRD number is wrong there will be a missing value in the person's job history, which appears as a 'hole'. Using probabilistic techniques, we attempt to match one-period job histories, or plugs, to the holes. Of course there are many valid reasons for interruptions to long periods of employment, and valid reasons for short spells of employment, but we are able to reduce the number that are caused by the miscoding of identifiers in the data.

For the one-period job history, we have the name variable, gross earnings and tax deductions variables, the employer's IRD number, month and sometimes the (incorrect) employee IRD number. These same variables are available for the periods on either side of the hole in the long job history, and are used to calculate expected values for the period of the hole.

Further work may be needed to develop methods to make corrections to missing values that occur at the beginning of a job history, or where there is miscoding in more than one period.

Details of work undertaken to address this issue will be reported in a separate paper.

4. Putting it all together

Figure 5 shows the sequence used to establish all the links involved with the LEED data. This is the proposal given our level of knowledge at the time of writing. Further work may lead to changes. (The numbers in circles refer to the links shown in figure 1.)
Figure 5

Work Flow for Creating Integrated LEED Data

1. **BF - EMS**
   - Deal with group returns (and other false positives from exact match?)

3. **EMS - EMS**
   - Establish Employee longitudinal link (e.g., Repair Employee IRD number)

1. **EMS - EMS**
   - Establish EMS employer longitudinal link (pred/succ)

1. **BF - EMS**
   - Establish exact BF-EMS matches which are OK

1. **BF - EMS**
   - Add matches from EMS pred/succ

1. **BF - EMS**
   - Add matches from BF pred/succ

1. **BF - EMS**
   - Add matches from BF births and deaths

1. **BF - EMS**
   - Add Probabilistic matches???
Appendix
Linking employers on the EMS to enterprises on the BF: The logical flow

The links between the tax record for an employer on the EMS and the correct enterprise on the BF will be held in a link file. This link file will consist of one record for each employer for each period on the EMS. Each employer has the IRD number from the EMS, the period, and the record on the BF to which it should be linked. Unique records are identified by two fields – the IRD number and period. There is also a field that records how the match was made.

The tables below show the logical flow for how these links are made for five main cases. The first is where the exact match on IRD number * period between the EMS and BF is correct. The second and third show how using the predecessor and successor fields on the EMS and BF enable further matches to be made when the timing of a change in IRD number is different between each source. The fourth is the case where employing business are already ceased, or not yet birthed, on the BF.

The variable ems_ird_nbr is the IRD number from the EMS, while bf_ird_nbr is the IRD number held on the BF. The identifier assigned by the BF for enterprises is ent_nbr.

Case 1: The exact match is correct

Where the exact match on IRD number for the period is correct, the link is simply the BF IRD number (which equals the EMS IRD number) for the same period. This is shown for the period month1 in figure 1.

Figure 6
Links Where the Exact Match is Correct

<table>
<thead>
<tr>
<th>EMS</th>
<th>Link to BF</th>
<th>BF</th>
</tr>
</thead>
<tbody>
<tr>
<td>ems_ird_nbr period</td>
<td>link_bf_ird_nbr link_period match_type</td>
<td>bf_ird_nbr period</td>
</tr>
<tr>
<td>emp1 month1</td>
<td>bf_nbr1 month1 exact</td>
<td>bf_nbr1 month1</td>
</tr>
<tr>
<td>emp2 month1</td>
<td>bf_no2 month1 exact</td>
<td>bf_nbr2 month1</td>
</tr>
<tr>
<td>... ...</td>
<td>... ...</td>
<td>... ...</td>
</tr>
<tr>
<td>empN month1</td>
<td>bf_noN month1</td>
<td>bf_nbrM month1</td>
</tr>
</tbody>
</table>
Case 2: Predecessor/successor on EMS

Longitudinal linking of the EMS will identify some of the situations where there is a change in IRD number, but the same employer. This is captured in the predecessor and successor fields. If the change in IRD number occurs at a later time on the BF, then there will not be an exact match for the months when the BF is still using the previous IRD number. The enterprise that should be linked is on the BF and can be found using the predecessor field. This is shown in figure 7. In month3, emp1 starts reporting as emp6 on the EMS. However, it is still reporting under emp1 on the BF. The link_bf_ird_nbr is then bf_nbr1 for month3.

Figure 7
Links Where a Change in IRD Number Occurs at Different Times – EMS Records Predecessor and Successor Recorded on EMS

<table>
<thead>
<tr>
<th>EMS</th>
<th>Link to BF</th>
<th>BF</th>
</tr>
</thead>
<tbody>
<tr>
<td>ems_ird_nbr</td>
<td>period</td>
<td>pred</td>
</tr>
<tr>
<td>emp1</td>
<td>month2</td>
<td>emp6</td>
</tr>
<tr>
<td>emp2</td>
<td>month2</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>empN</td>
<td>month2</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>emp2</td>
<td>month3</td>
<td></td>
</tr>
<tr>
<td>emp3</td>
<td>month3</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>empN</td>
<td>month3</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>emp6</td>
<td>month3</td>
<td>emp1</td>
</tr>
</tbody>
</table>

Case 3: Links where a change in IRD number occurs at different times – predecessor and successor recorded on BF

Similarly, if the change in reporting number occurs first in the BF, the link can be found if the pred/succ fields are available on the BF. This is shown in figure 8 where emp2 continues reporting on the EMS in months 4 and 5, but on the BF is reporting as bf_nbr8 in month5. The pred field on the BF allows the correct link to be made.

Figure 8
Links Where a Change in IRD Number Occurs at Different Times – BF Records Predecessor and Successor Recorded on BF

<table>
<thead>
<tr>
<th>EMS</th>
<th>Link to BF</th>
<th>BF</th>
</tr>
</thead>
<tbody>
<tr>
<td>ems_ird_nbr</td>
<td>period</td>
<td>pred</td>
</tr>
<tr>
<td>emp2</td>
<td>month4</td>
<td>bf_nbr2</td>
</tr>
<tr>
<td>emp3</td>
<td>month4</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>empN</td>
<td>month4</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>emp2</td>
<td>month5</td>
<td>bf_nbr8</td>
</tr>
<tr>
<td>emp3</td>
<td>month5</td>
<td></td>
</tr>
<tr>
<td>emp4</td>
<td>month5</td>
<td></td>
</tr>
<tr>
<td>emp8</td>
<td>not on EMS</td>
<td></td>
</tr>
</tbody>
</table>

emp1 not on EMS
link by EMS pred number

bf_nbr8 month5, bf_nbr2 ent8
link by BF pred number
Cases 4 and 5: Employing on EMS, but not birthed (or already ceased) on BF

Once links are made for continuing businesses, we can capture the situation where a birth or death occurs at different times. There will be no match from the EMS to the BF if a business is employing before it appears on the BF. Searching the BF for the first appearance on the BF allows the link to be made to the same IRD number, but at a different period on the BF. When an employing business is already ceased on the BF, the IRD number can be found at a previous period and the link made. In figure 9 emp10 appears on the EMS in month6 but is not birthed on the BF until ** month7. In month7 emp9 still appears on the EMS, but was ceased on the BF in month6.

Figure 9

<table>
<thead>
<tr>
<th>Links Where Birth Occurs First on EMS, or Death Occurs First on BF</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMS</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>ems_ird_nbr</td>
</tr>
<tr>
<td>emp8</td>
</tr>
<tr>
<td>emp9</td>
</tr>
<tr>
<td>emp10</td>
</tr>
<tr>
<td>emp8</td>
</tr>
<tr>
<td>emp9</td>
</tr>
<tr>
<td>emp10</td>
</tr>
</tbody>
</table>

References


