Box 1: A simplified example of probabilistic linking

The following example is taken from the Statistics New Zealand Data Integration Manual.
It looks at two records on two different datasets to see whether they are a match and, therefore, should be linked.

Field	Record A	Record B
Name	Jon Block	John Black
Date of birth	23-11-65	23-11-63
Sex	M	M
Address	89 Molesworth Street	112 Hiropi Street

The linking software assigns m and u probabilities for each field (shown below) which range between 0 and 1. (The calculation of the agreement field weight is log^2[m/u]. The disagreement field weight is log^2([1 – m]/[1 –u]).)

Field	m probability	u probability	Agreement field weight	Disagreement field weight
Name	0.95	0.01	6.57	–4.31
Date of birth	0.9	0.01	6.49	–3.31
Sex	0.95	0.5	0.93	–3.32
Address	0.7	0.01	6.13	–1.72

The fields on the two records are compared (see table below). Field weights with positive values indicate that fields agree, while negative values indicate disagreement. In this example, the field weight of –1.72 indicates that the records do not agree on address.

For simplicity, this example assumes no partial field agreement, although in practice, 'Jon' and 'John' are sufficiently similar that there would probably be some sort of adjustment to the field weights to take account of this, resulting in a lower, but still positive, agreement weight.

Field	File A	File B	Agreement?	Field weight
Name	Jon Block	John Black	No	–4.31
Date of birth	23-11-65	23-11-63	No	–3.31
Sex	M	M	Yes	0.93
Address	89 Molesworth Street	112 Hiropi Street	No	–1.72
			Composite weight (sum of field weights)	–8.41

The field weights are summed: (–4.31) + (–3.31) + (0.93) + (–1.72) = –8.41 (= composite weight). As the composite weight in this example is negative (–8.41), the linking process would determine that these records are a **non-link**. For more information, see Statistics New Zealand, *Data Integration Manual*, 2006, pp.36-44.

For more information

Probabilistic linking theory

- Fellegi, I. and Sunter, A. 1969, 'A Theory for Record Linkage', *Journal of the American Statistical Association*, Vol.64, no.328, pp. 1183-1210.
- Winkler, W.E. 2006, 'Overview of Record Linkage and Current Research Directions', *Research Report Series*, no. 2006–2, Statistical Research Division, U.S. Census Bureau, Washington.

Linking (including m and u probabilities)

- Christen, P. 2012, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*, Springer, Canberra.
- Herzog, T.N., Scheuren, F.J. and Winkler, W.E. 2007, *Data quality and record linkage techniques*, Springer, New York.
- Jaro, M. 1995, 'Probabilistic Linkage of Large Public Health Data Files', *Statistics in Medicine*, Vol. 14, pp. 491-498.
- Samuels, C. 2012, 'Using the EM Algorithm to Estimate the Parameters of the Fellegi-Sunter Model for Data Linking research paper', *Methodology Advisory Committee Paper*, Cat. no. 1352.0.55.120, Australian Bureau of Statistics, Canberra.
- Statistics New Zealand 2006, *Data Integration Manual*, Statistics New Zealand, Wellington.
- Winkler, W.E. 1990, 'String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage', *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 354-359.

Clerical review

- Guiver, T. 2011, 'Sampling-Based Clerical Review Methods in Probabilistic Linking', *Methodology Research Papers*, Cat. no. 1351.0.55.034, Australian Bureau of Statistics, Canberra.
- Bishop, G. and Khoo, J. 2007, 'Methodology of Evaluating the Quality of Probabilistic Linking', *Methodology Research Papers*, Cat. no. 1351.0.55.018, Australian Bureau of Statistics, Canberra.

To provide feedback on this series please email: **statistical.data.integration@nss.gov.au**

Australian Government

National
Statistical
Service

Data Linking

Information Sheet Four

Probabilistic linking

Probabilistic linking is a method for combining information from records on different datasets to form a new linked dataset.

It has been described as a process that attempts to link records on different files that have the greatest probability of belonging to the same person/organisation.

Whereas deterministic (or exact) linking uses a unique identifier to link datasets, probabilistic linking uses a number of identifiers, in combination, to identify and evaluate links.

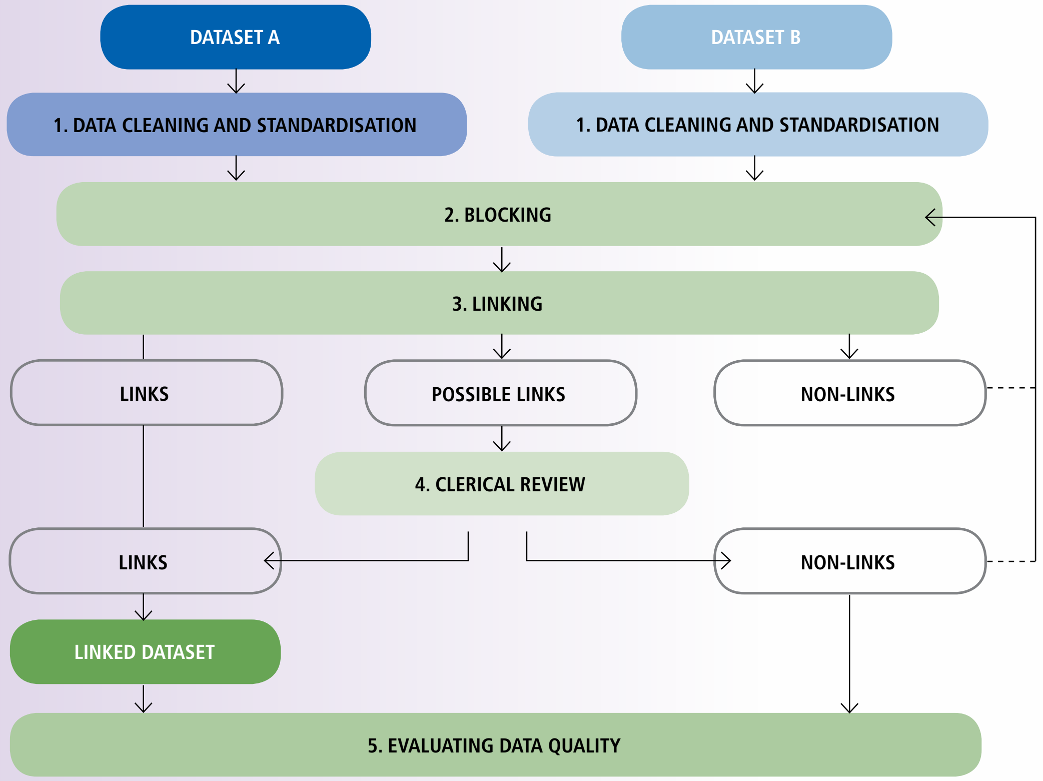
Probabilistic linking is generally used when a unique identifier is not available or is of insufficient quality.

The method derives its name from the probabilistic framework developed by Fellegi and Sunter (1969) and requires sophisticated software to perform the calculations. References at the end of this sheet provide more information about linking algorithms.

The key steps of probabilistic linking (as shown in Diagram 1) are:

1. Data cleaning and standardisation
2. Blocking
3. Linking
4. Clerical review
5. Evaluating data quality

Diagram 1: Key steps in probabilistic linking



Key Terms

Blocking – divides datasets into groups, called blocks, in order to reduce the number of comparisons that need to be conducted to find which pairs of records should be linked. Only records in corresponding blocks on each dataset are compared, to identify possible links.

Fields – types of information, such as name, address, date of birth, on the records in datasets.

Linked dataset – the result of linking different datasets is a new dataset whose records contain some information from each of the original datasets.

Links – records that have been combined after being assessed as referring to the same entity (i.e., person/family/organisation/region).

Unique identifier – a number or code that uniquely identifies a person, business or organisation, such as passport number or Australian Business Number (ABN).

DataLinking. Probabilistic linking

Key steps in probabilistic linking

1. Data cleaning and standardisation

See Sheet 2 in this information series.

2. Blocking

Data linking involves a large number of record comparisons. Ideally, every record on Dataset A is compared with each one on Dataset B to find which record pairs are most likely to be links. However, if every record is compared between two datasets containing 100 000 records each, this would require 10 billion comparisons. Even using advanced computer power, this would take a long time to perform.

As a way to save time, 'blocking' is used to reduce the number of record comparisons required to find potential record pairs.

Blocking is similar to sorting a basket of socks into like colours before trying to locate the pairs. See Diagram 2.

For example, by using gender for blocking, only records with the same sex (males or females) are compared to each other, which usually cuts in half the number of comparisons required.

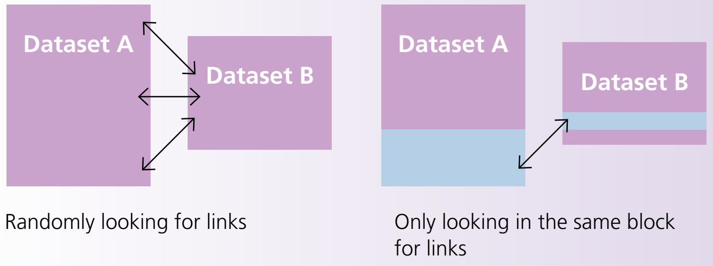
However, gender is not overly useful for blocking as it only separates the dataset into two large blocks, so a lot of comparisons are still required. Ideally, a blocking strategy should result in small, equal-sized blocks on each dataset.

For example, using month of birth would result in 12 blocks (one for each month) and be expected to have a fairly even number of records in each block. A common strategy is to keep the block sizes as small as possible and run multiple blocking passes.

Blocking passes

Sometimes links are missed because the information in the selected blocks was missing or incorrectly recorded.

Diagram 2: Example of blocking


Dataset A
Dataset B
Randomly looking for links
Dataset A
Dataset B
Only looking in the same block for links

To compensate for possible errors in the data, a different blocking field may be used for a re-run of the linking process. This is called a blocking pass. Each time a pass occurs, the links are kept and the remaining unlinked records are subjected to another pass using a different block. Each blocking pass can be based on one or more fields (see Table 1).

The more blocking passes used, the more links are likely to be identified, but there will generally be diminishing returns on subsequent passes.

Devising a blocking strategy

Although all the fields that are common to both datasets could potentially be used to compare the records, some are more appropriate than others, as discussed in Table 1.

Selecting which fields will be used for blocking and linking should be established before the process starts. It is good practice to plan and document the blocking strategy used for the project.

Blocking often requires some trial and error to determine the best outcome. References at the end of this sheet provide more information on blocking.

Table 1: Considerations when choosing which fields to use for linking and blocking

Surnames	Standardising makes this field more useful for linking and blocking. Sometimes parts of the surname and first name (e.g., first two letters of each) may be used as a blocking variable. Note that name changes can occur as a result of marriage and divorce and sometimes surnames and first names are swapped around. Spelling variations can result from errors during recording or transcription.
First names	Inconsistencies can result when both nicknames and formal names are used interchangeably, which can cause discrepancies between data sources.
Gender	Gender is not overly useful for blocking as it only separates the datasets into two large blocks. Gender is often good for linking as it is generally well reported and unlikely to change during a person's lifetime.
Birth date	Birth month and birth day are usually reliable because they do not change over a person's lifetime and are usually well reported. If age is collected, it can be checked against birth year for consistency. There may be different date formats (e.g., MM/DD/YYYY or DD/MM/YYYY) which can be addressed by standardisation. Transcription errors may occur when digits are accidentally transposed. Blocking often uses month and year aggregated, rather than the date of birth.
Age	Age may be checked against birth year, if available. Age groups may be used as blocks. It is not recommended that age be used in the same blocking pass as birth date because there is a direct relationship between the two fields.
Address	As with birth dates, there can be formatting differences in address records, so these generally require standardising. People may change address and sometimes use postal address and street address interchangeably, which affects its usefulness. However, address can still be useful for confirming matches. When address is used in combination with other non-geographical blocks, there is still a chance of identifying links even if people have moved address. For blocking, it is often useful to use an aggregated form of address, such as the suburb or postcode.

3. Linking

Linking involves a number of steps including assigning field probabilities and using these to calculate field and composite weights. The weights reflect the similarities of each pair of records and provide a way to decide whether they are links, or not.

Calculating probabilities

Each linking field (e.g., name, address, date of birth) on a record has two probabilities associated with it. They are called the 'm' and 'u' probabilities.

The m probability is the likelihood of the field values agreeing on a pair of records, given the records refer to the same entity (e.g., person, organisation).

It can be thought of as follows: if two different datasets contain records that refer to the same person, what would prevent the information in those records from agreeing? The answer is usually errors in spelling or missing information.

Therefore, the m probability reflects the reliability of the field, and is calculated as 1 minus the error rate of the field. For example, most people report their gender consistently over time and on different datasets, so the m probability is close to 1 (0.95 in the example on the next page). Address, however, may only have an m probability of 0.7 because different datasets can have different addresses for the same person because they may have moved house.

The probabilistic linking software is used to produce estimates for the m probability. These estimates may be based on prior knowledge of the datasets or similar linking projects, through the identification of a large number of linked and non-linked records that 'train' the software, or by using the EM (Expectation-Maximisation) algorithm (see Samuels, 2012 for more information).

The u probability is the likelihood that the field values on two records will agree, given the two records refer to different entities.

It is essentially a measure of how likely two fields on different records will agree by chance. Another way to think of this is: given two records belong to two different people, what is the probability that they will agree anyway?

The u probability is often estimated by 1/n (where n is the number of possible values). For gender, there are two possible values (male or female) so the u probability for gender is ½ (0.5). For month of birth, the u probability can be estimated as 1/12 (0.08).

Calculating the field weights

Using the m and u probabilities, the probabilistic linking software generates an estimate of how closely the relevant fields agree on each record pair being compared. This is called a field weight.

In practice, field weights may be modified to allow for partial agreement, such as a minor difference in spelling (e.g., Block vs Black in the example in Box 1).

This is achieved using a 'string comparator' to generate a lower, but still positive agreement field weight depending on the specified degree of similarity. Some string comparator options include:

- Exact match where the fields either agree or they do not – no adjustment is made to the field weight (e.g., gender).
- Exact match but the weight is modified so that rarer values are given higher weights than more common values when they agree (e.g., country of birth).
- Approximate string comparison (e.g., name) where the weight depends on the number of characters that differ, allowing for misspellings and poor handwriting (see Winkler, 1990).

For more information about m and u probabilities and field weights, refer to Fellegi and Sunter (1969).

For each possible record pair, the field weights are summed to produce an overall weight – the composite weight. The higher the composite weight, the more likely that both records refer to the same entity. Box 1 provides a highly simplified example of this process.

Determining links based on threshold cut-offs

The composite weight for each record pair is compared to the cut-off threshold. If the composite weight is above the cut-off, the record pair is deemed to be a link. If the composite weight is below the cut-off, the record pair is deemed not to be a link. Sometimes two cut-off thresholds (upper and lower) are used.

A key feature of this methodology is the ability to rank all of the possible links and then, using an 'optimal threshold' algorithm (see Christen, 2012), assign the link to the most optimal record pair based on how well the records match. Various manual and automated methods are available to determine thresholds based on the distribution of the composite weights for the linked records.

4. Clerical Review

Clerical review is a useful tool to manually assess those records without a designated link/non-link status and to examine records close to the thresholds to check if they are links.

However, clerical review is time-consuming and resource-intensive. To minimise the number of records that need to be reviewed, it is necessary to ensure that threshold values are appropriate and the linking software is operating as efficiently as possible. For more information on clerical review see Guiver (2011).

5. Evaluating data quality

This is covered in Information Sheets 2 and 5 (to be released).