Australian Government

National
Statistical
Service

Data Linking

Information Sheet One

What is data linking?

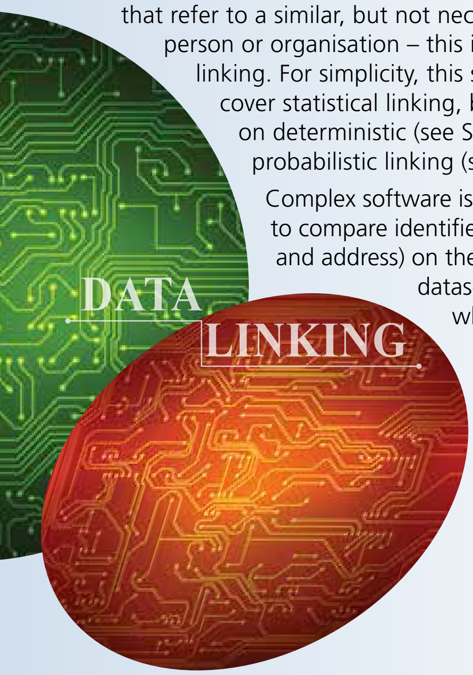
Data linking is used to bring together information from different sources in order to create a new, richer dataset.

This involves identifying and combining information from corresponding records on each of the different source datasets. The records in the resulting linked dataset contain some data from each of the source datasets.

Most linking techniques combine records from different datasets if they refer to the same entity. (An entity may be a person, organisation, household or even a geographic region.)

However, some linking techniques combine records that refer to a similar, but not necessarily the same, person or organisation – this is called statistical linking. For simplicity, this series does not cover statistical linking, but rather focuses on deterministic (see Sheet 3) and probabilistic linking (see Sheet 4).

Complex software is often required to compare identifiers (such as name and address) on the records in both datasets, to assess whether they refer to the same entity. If the identifiers of the records on the different datasets agree, then the records are linked.



Why is data linking important?

Linked datasets create opportunities for more complex and expanded policy and research.

For example, data linking helped to identify the role of folate in pregnancy in reducing neural tube defects, such as spina bifida.

On the business front, in New Zealand, the Linked Employer-Employee Data (LEED) links taxation data with business data to provide information such as the employment outcomes of tertiary education and transitions from work to retirement and from benefit to work.

Data linking has the advantage of utilising information that already exists. Making use of data collections in this way avoids the time and expense of collecting a whole new set of data. It also avoids imposing extra questions on people and organisations when this information already exists.

The Data Linking Information Series

These information sheets provide a broad overview of the main technical aspects of data linking. As this is a complex topic, technical references are provided at the end of each sheet.

The Data Linking Information Series currently comprises:

Sheet 1: What is data linking?

Sheet 2: Preparing for linking

Sheet 3: Deterministic linking and linkage keys

Sheet 4: Probabilistic linking

This series will be expanded in the future to provide further information about evaluating data quality.

Key Terms

Confidentiality – the legal and ethical obligation to maintain and protect the privacy and secrecy of the person, business, or organisation that provided their information.

Data linking – creating links between records from different sources based on common features present in those sources. Also known as 'data linkage' or 'data matching', data are combined at the unit record or micro level.

Deterministic (exact) linking – using a unique identifier to link records that refer to the same entity.

Identifier – for the purpose of data linking, an identifier is information that establishes the identity of an individual or organisation. For example, for individuals it is often name and address. Also see *Unique identifier*.

Source dataset – the original dataset as received by the data provider.

Unit record level linking – linking at the unit record level involves information from one entity (individual or organisation) being linked with a different set of information for the same person (or organisation), or with information on an individual (or organisation) with the same characteristics. Micro level includes spatial area data linking.

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

Australian Government

National
Statistical
Service

Data Linking. What is data linking?

Information Sheet One

What are the main ways to link datasets?

There are a number of different approaches to data linking.

Usually, the most straightforward way is to use a unique identifier (such as a tax file number) present on both files, in order to identify the links between the records on each dataset. This is sometimes referred to as 'deterministic' or 'exact' linking because the unique identifiers on the records either match or they do not – there is no uncertainty.

Where a unique identifier is not available, or is not of sufficient quality or completeness to be relied on alone, an alternative approach is to construct a linkage key, which acts as a proxy for the unique identifier. This key (or code) is created using identifiable information, such as name and address, available on both datasets.

Linkage keys can help to preserve privacy because the key replaces name and address, thereby reducing the chance of identification. Information Sheet 3 in this series has more information on linkage keys and deterministic linking.

Probabilistic linking is another option for linking where a unique identifier is not available. Probabilistic linking is based on a calculation of the likelihood that a pair of records (one drawn from each dataset) refers to the same person/organisation. Complex methods and sophisticated data linking software are used to achieve high-quality results.

Information Sheet 4 in this series has more information on probabilistic linking.

Protecting privacy and confidentiality of a linked dataset

Datasets that contain identifiable information need to be handled with care to protect the identity of a person or organisation. There is an increased risk of identification of an individual/business/organisation when two datasets are linked.

Even if identification is protected (such as by removing name and address) in the original datasets, the result of the linking may provide a combination of characteristics which leads to spontaneous recognition of the identity of a person or organisation (e.g., local area school data showing a cardiac specialist who is the mother of six).

To minimise this risk, data linking should only be conducted in a safe and effective environment ensuring that the methods used are fit-for-purpose. Confidentiality and statistical disclosure techniques are available to manage the privacy risks that can be associated with data linking (see Confidentiality Information Series Sheet 5).

If a data linking project involves Commonwealth datasets and is for statistical and research purposes the project should comply with the *High Level Principles for Data Integration Involving Commonwealth Data for Statistical and Research Purposes* and the supporting governance and institutional arrangements.

See the National Statistical Service (NSS) website for further information, **http://www.nss.gov.au/dataintegration**

For more information

- Christen, P. 2012, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*, Springer, Canberra.
- NCSIMG 2004, *Statistical Data Linkage in Community Services Data Collection*, AIHW, Canberra.
- National Statistical Service 2011, *Confidentiality Information Series*, National Statistical Service, Canberra, http://www.nss.gov.au
- Statistical Data Integration (including the High Level Principles), see National Statistical Service at http://www.nss.gov.au/dataintegration
- Statistics New Zealand 2006, *Data Integration Manual*, Statistics New Zealand, Wellington.
- 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.

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