Australian Government

National Statistical Service

Data Linking

Information Sheet Three

Deterministic linking and linkage keys

Deterministic linking involves the exact matching of information on different records across the datasets being combined for a linking project.

The simplest form of deterministic linking uses a unique identifier, such as an Australian Business Number or a social security number, to determine if the records refer to the same entity. (An entity may be a person, household, organisation or locality.)

This is also called 'exact linking' because the identifier either matches or does not match. This means that if the unique identifier contains any errors, the matches will not be found because the identifiers must be identical on all the datasets being linked.

If it is the case that the unique identifiers may be unreliable for linking purposes, there are a number of other options available. One is to instead use probabilistic linking – see Sheet 4.

Other possible approaches include variations of deterministic linking, such as 'stepwise deterministic linking' and 'rules-based linking'. These techniques use other information on the records to overcome deficiencies in the quality of the unique identifier. For more information, see the references on this sheet.

If a unique identifier is not available, or is not of sufficient quality, it is possible to create a proxy, often referred to as a linkage key.

Creating a linkage key

A linkage key is a code created using a combination of identifying information on each record, such as name, address and date of birth (see Table 1 for an example).

The linkage key usually replaces identifiers on the linked record. If the records in the linked dataset are de-identified (by removing name and address), this helps to protect the identity of the people or organisations in the new dataset.

However, this does not necessarily ensure privacy protection. Even without name and address it may still be possible to recognise a person or organisation, through a set of unusual characteristics in the linked dataset. For example, small-area data (e.g., a suburb) showing a 17-year-old widow with four children could be recognisable to someone living in that area. Therefore, further confidentiality techniques need to be applied before releasing the data.

The Confidentiality series provides more information on privacy and confidentiality.

Table 1 shows how a 12 character key might be built using:

- the second, third and fifth letters of a person's last name, second and third letter from a person's first name
- the second, fourth, sixth and seventh numbers from a person's date of birth (DD/MM/YYYY)
- gender (male is 1 and female is 2)
- the second and third numbers of the postcode.

Table 1: Example of creating a linkage key

Name	Date of birth	Gender	Postcode
John Smith	12/05/1970	Male	5623
Linkage key = MIHOH2597162			

As with the unique identifier, if there is an error or missing information on the records, the linkage key may not match exactly and therefore the records will not be linked.

As linkage keys use identifiers in their creation, technically they could be re-constructed, thereby identifying people in the dataset. Therefore, encryption of the key is recommended as an additional safety measure to avoid the risk of identification or re-identification.

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.

Content data – the term used for the administrative or clinical information on a record (such as medical condition, income, educational attainment), as opposed to identifying information (such as name and address).

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*.

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. Deterministic and linkage keys

Information Sheet Three

An example of deterministic linking using a linkage key

Stage 1: Assigning linkage keys to all records within datasets A and B.

This example uses a linkage key (based on Table 1) for a project looking at educational attainment and earnings, by age and sex.

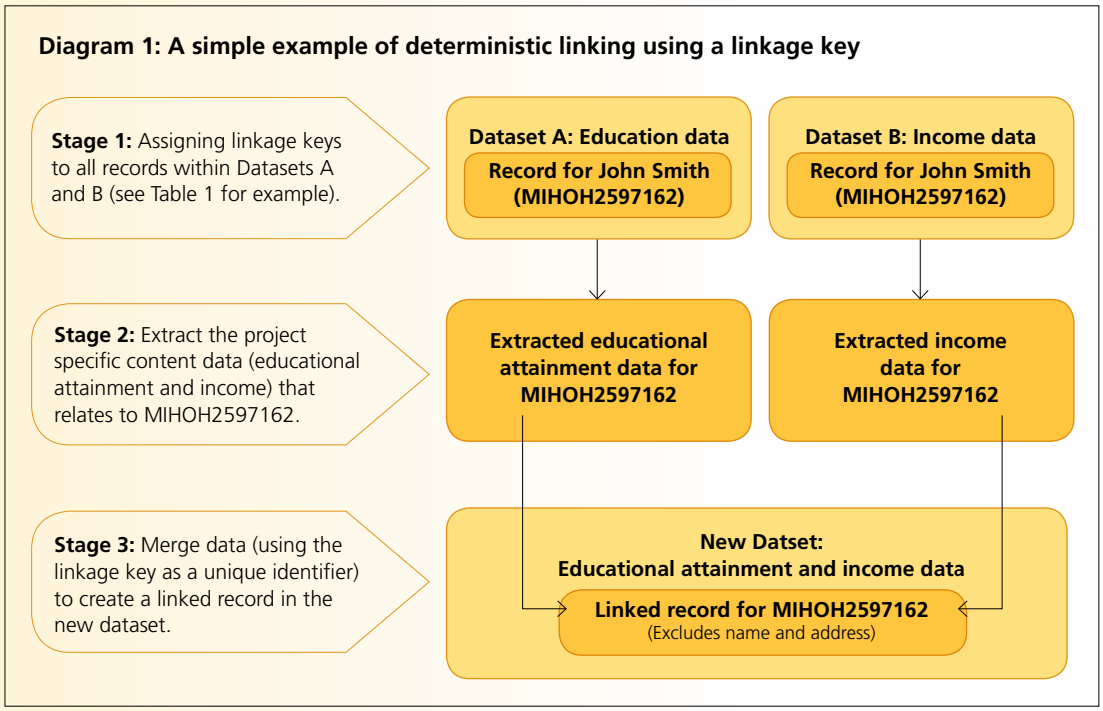
Stage 2: Extracting the content data and unique identifier.

In this example, the linkage key (MIHOH2597162) identifies the records that refer to the same person, in this case John Smith.

Stage 3: Merging to create a linked record.

Using the linkage key, only the information required for the project (highest educational attainment, income, sex and age) is extracted from each record and merged into a new linked record (now known by the identifier MIHOH2597162) in the new dataset.

Diagram 1: A simple example of deterministic linking using a linkage key



For more information

- Australian Institute of Health and Welfare and Australian Bureau of Statistics 2012, *National best practice guidelines for data linkage activities relating to Aboriginal and Torres Strait Islander people*, AIHW Cat. no. IHW 74, AIHW, Canberra.
- Australian Institute of Health and Welfare 2011, 'Comparing an SLK-based and a name-based data linkage strategy: an investigation into the PIAC linkage', *Data linkage series*, No. 11. Cat. no. CSI 11, AIHW, Canberra.
- Australian Institute of Health and Welfare: Karmel, R. 2005, 'Data linkage protocols using a statistical linkage key', *Data linkage series*, No. 1. Cat. no. CSI 1, AIHW, Canberra.
- Bass, J. and Garfield, C. 2002, 'Statistical linkage keys: How effective are they?' *Proceedings of Symposium on health data linkage*, Public Health Information Development Unit, held March 20-21 2002, Sydney, pp. 40-45.
- National Statistical Service 2011, *Confidentiality Information Series*, National Statistical Service, Canberra http://www.nss.gov.au
- NCSIMG 2004, *Statistical Data Linkage in Community Services Data Collection*, AIHW, Canberra.

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