Australian Government

National Statistical Service

Data Linking

Preparing for linking

Significant effort is required to prepare datasets before they can be linked. Pre-linking preparation includes understanding the source dataset, selecting which fields to use for linking, data cleaning and standardising and, where required, secure data transfer.

It has been estimated that about three-quarters of the effort involved in data linking involves preparing the data to ensure the input files are ready for linking (Gill, 2001).

It is necessary to invest the time and effort in data preparation in order to ensure the quality of the linked dataset.

Understanding the source datasets

Before attempting linking, it is important to understand the information in the datasets being linked, to ensure the linking software is comparing 'apples with apples'. This can be achieved in a number of ways, including examining the metadata and checking a sample of the dataset.

Metadata is information about data that provides context or additional information, such as how and when the data was collected, its purpose and quality.

Understanding the metadata helps to avoid combining data that is not compatible. For example, one data source might report earnings on a financial year basis while the other uses calendar year reporting. This would need to be resolved before linking and analysis can occur.

For more information on metadata, there are references listed at the end of this sheet.

Selecting which fields to use for linking

The main requirement when selecting which fields to use for linking is to ensure that they exist on each of the datasets being linked. Other considerations include the quality of that data, such as whether the data contains any errors or missing information. These may be caused by, for example, variations in spelling (Smith and Smyth), incomplete information or spelling errors (e.g., is 'Ray's Burgers' the same business as 'Roy's Burgers'?).

Differences in the fields on different datasets often arise as a result of factors in the data collection phase. Errors may occur if information is supplied by a relative rather than the person themselves, as could be the case, for example, for schools enrolment data or hospital admissions.

Varying levels of quality between datasets often depend on the relative importance of a particular field for a dataset. For example, in mortality datasets the names are critical for identity resolution reasons and so this field is collected and transcribed very carefully and is, therefore, high quality.

Data cleaning and standardising

The terms 'data cleaning' and 'data standardisation' refer to the task of transforming the records on each dataset into a format that allows them to be linked with other datasets.

The key aim of data cleaning and standardising is to ensure all the data being linked is consistent and uniform, so the linking software can work effectively.

Most datasets contain some incomplete, out-of-date or differently formatted information. Computerised records can contain errors because, for example, information has been provided or recorded incorrectly.

The most common data inconsistencies involve variations in spelling of names, such as nicknames (e.g., Rob and Robert), formats of date of birth, data coding and missing information.

Techniques used to resolve these inconsistencies to convert data into standard forms include: editing, standardising, de-duplication and correspondences.

Editing is the identification and treatment of data errors and anomalies. For example, editing may include the removal of an impossible response, such as a birth date in the future and its re-clarification as a missing value.

Where data is missing, sometimes there is enough information available to approximate the data. This is often called 'data repair'. Data repair should be well documented in terms of the decisions arrived at and data used to make the repairs.

This may involve accessing additional information from previous data collections similar to the one being linked, to check for inconsistencies and help to inform decisions about missing information. An example of data repair would involve changing a postcode to match suburb and street information.

Standardising is the formatting of data items so they are consistently represented in all of the datasets. For example, dates can be represented in different formats such as, DD/MM/YYYY or MM/DD/YYYY. One format should be selected and employed across all of the datasets for that project.

De-duplication is removing duplicates from a dataset. For example, when a record incorrectly appears more than once in a dataset with exactly the same data item value, one copy should be retained, and the others should be removed. Alternatively, the same unit may appear more than once in the same dataset but with different values for some of the data items. Sorting and ordering the datasets will highlight these cases. A set of rules should be developed to treat such duplicates in a consistent manner, such as how to decide which records to keep.

Information Sheet Two

Australian Government

National Statistical Service

Data Linking. Preparing for linking

Information Sheet Two

Correspondence involves creating a consistent coding classification across all datasets. A sound understanding of the data is required to produce the best correspondence. Table 1 shows how marital status could be coded in two different datasets.

Dataset 2 could be used as the standard and the first dataset would be altered to be consistent with that coding.

Table 1: Example of a correspondence for marital status

Dataset 1	Dataset 2
1. Never Married	1. Never Married
2. Currently Married	2. Currently Married
3. Separated	3. No Longer Married
4. Divorced	
5. Widowed	

Secure data transfer

When transferring data between organisations, secure data transfer considerations are essential to preserve privacy. They include consideration of the format, method and encryption.

Box 1 shows an example of secure data transfer, based on Statistics New Zealand's key steps in encrypted-DVD courier delivery – a common method of data transfer for large datasets.

Email transfer is often only suitable for small datasets. More secure email systems are being developed that could provide better options for consideration in the future.

Box 1: Example of key steps for secure data transfer by courier

1. Media creation and encryption	The data custodian extracts information from their systems and copies it to a CD/DVD using an encrypted format.
2. Handover to courier	The data custodian hands the media to the courier, with the name and address of the recipient (the organisation doing the linking).
3. Delivery	A key contact person from the linking organisation receives the data and signs for its receipt. All transactions must be recorded.
4. Transfer to IT secure facilities	The data should be personally carried by the key contact person to the relevant secure IT facilities of the linking organisation.
5. Place in safe storage	The data transfer and the database loading are completed and stored securely.
6. Disposal	Once transfer is completed, the linking organisation destroys the media (i.e., shredding of DVDs, deletion of back-up files, and for very sensitive information at high risk, degaussing (demagnetisation)).
7. Confirmation to data custodian	The linking organisation informs the data custodian when the data transfer is successful (or not).
8. Preliminary checks of the data	Once the data has been received, the linking organisation checks the received data to confirm it is fit-for-purpose.

For more information

Understanding metadata

- Metadata Management, see the National Statistical Service, *Keeping your data in good shape*, and accompanying case studies http://www.nss.gov.au

Data quality

- Australian Bureau of Statistics 2011, *Information Paper: Quality Management of Statistical Outputs Produced from Administrative Data*, Cat. no. 1522.0, Australian Bureau of Statistics, Canberra.
- Australian Institute of Health and Welfare Metadata Online Registry (METeOR), metadata, quality statements and National Data Dictionaries http://meteor.aihw.gov.au

Data cleaning and standardising

- Gill, L. 2001, 'Methods for Automatic Record Matching and Linkage and their Use in National Statistics', *National Statistics Methodological Series,* No. 25, Oxford University, Norwich.
- Statistics New Zealand 2006, *Data Integration Manual*, Statistics New Zealand, Wellington.

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