Australian Government

National Statistical Service

Data Linking

Information Sheet Five

Linked data quality

Evaluating linked data quality is a key component of the linking process and essential for any subsequent analysis that uses the linked datasets. A comprehensive understanding of the linked data is built through the measurement, assessment and documentation of the linked record pair quality.

Data quality assessment occurs in three stages. Stage 1 is when the data is received, Stage 2 when the data is cleaned, edited and standardised, and Stage 3 occurs after the data is linked. See Diagram 1.

Stages 1 and 2 are covered in Information Sheet 2 - Preparing for Linking. Stage 3 is covered in this Information Sheet.

For more information on data quality and sound data management practices, see Brackstone (1999) and the National Statistical Service website.

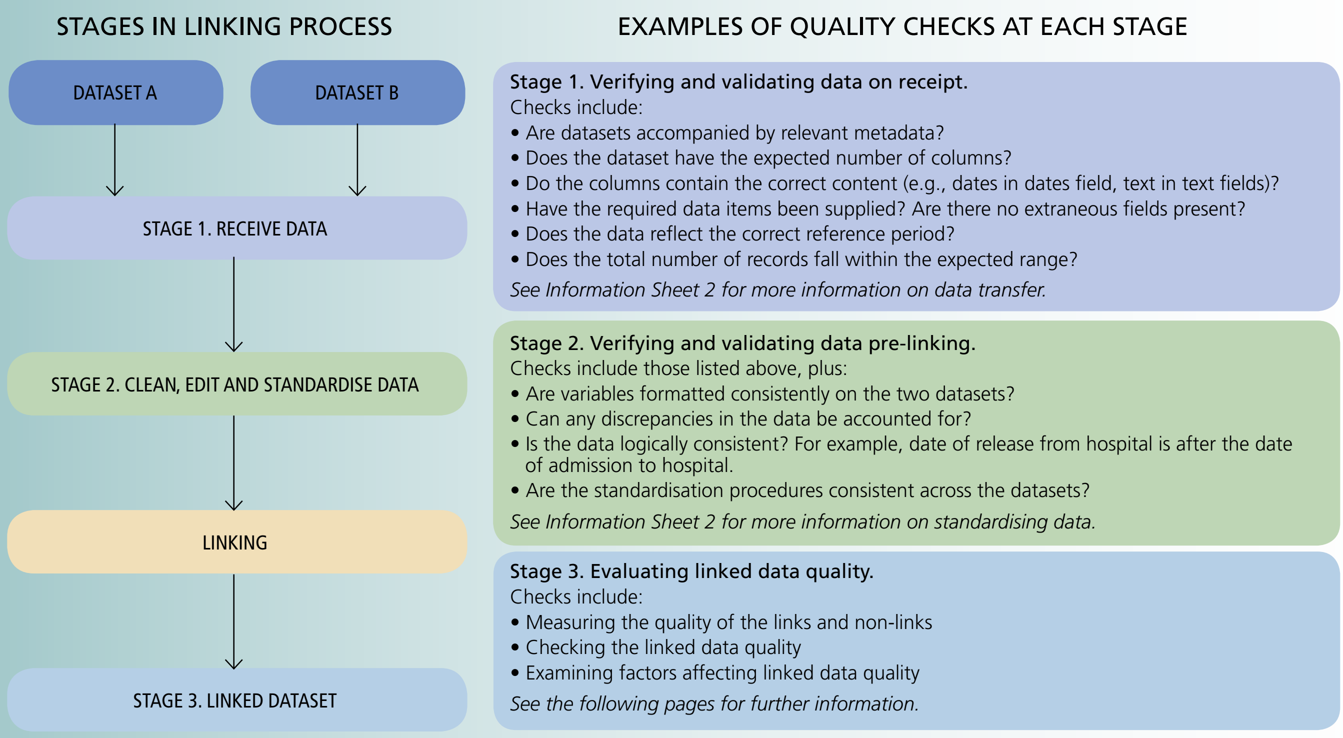
Evaluation of linked data quality

Evaluation of linked data quality requires a number of assessments and techniques. To do this, it is important to understand the difference between a match and a link.

A **match** is a pair of records that refer to the same entity (e.g., person or organisation), whereas a **link** is a pair of records combined by the linkage process irrespective of whether the records refer to the same entity.

Data quality measures calculate the number of records which were correctly linked (i.e., refer to the same entity), and which records were missed. Prior to analysis, it is important to consider and assess the factors that may have influenced the linked data quality.

Diagram 1: Overview of data quality checks for linked data



STAGES IN LINKING PROCESS

DATASET A

DATASET B

STAGE 1. RECEIVE DATA

STAGE 2. CLEAN, EDIT AND STANDARDISE DATA

LINKING

STAGE 3. LINKED DATASET

EXAMPLES OF QUALITY CHECKS AT EACH STAGE

Stage 1. Verifying and validating data on receipt.
Checks include:

- Are datasets accompanied by relevant metadata?
- Does the dataset have the expected number of columns?
- Do the columns contain the correct content (e.g., dates in dates field, text in text fields)?
- Have the required data items been supplied? Are there no extraneous fields present?
- Does the data reflect the correct reference period?
- Does the total number of records fall within the expected range?

See Information Sheet 2 for more information on data transfer.

Stage 2. Verifying and validating data pre-linking.
Checks include those listed above, plus:

- Are variables formatted consistently on the two datasets?
- Can any discrepancies in the data be accounted for?
- Is the data logically consistent? For example, date of release from hospital is after the date of admission to hospital.
- Are the standardisation procedures consistent across the datasets?

See Information Sheet 2 for more information on standardising data.

Stage 3. Evaluating linked data quality.
Checks include:

- Measuring the quality of the links and non-links
- Checking the linked data quality
- Examining factors affecting linked data quality

See the following pages for further information.

Key Terms

False links – records that were incorrectly linked and do not refer to the same entity. See Box 1.

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

Links – combined record pairs assessed as referring to the same entity (i.e., person/family/organisation/region).

Matches – record pairs that actually refer to the same entity (e.g., person/family/organisation/region).

Metadata – provides data users with information about the purpose, processes, and methods involved in the data collection; from design through to communication.

DataLinking. Linked data quality

Measuring the quality of the links

Ideally, the set of links will be identical to the set of matches. However, this is not usually the case as there may be records without a match or a linking error may occur. See Box 1.

Some of the measures used to check the quality of the links involve considerable statistical expertise. These include evaluating the *accuracy, specificity, sensitivity (match rate), precision (link accuracy), false negative rate and the false positive rate.* Refer to Box 2 and Christen & Goiser (2007) for more information on the definition and calculation of these measures.

Measures which rely on knowing the number of records that do not have a match (i.e., are a true negative non-link) can be difficult to calculate. Of the techniques noted above, the *accuracy, specificity, false negative rate* and the *false positive rate* measures are included in this category.

More commonly, measures based on the total number of matches that are linked (i.e. true positive links) are used. These measures include *match rate* and *link accuracy*.

The *match rate* is defined as:

$$\text{Match rate} = \frac{\text{number of true positive links}}{\text{total matches}}$$

The *match rate* measures the proportion of matches linked (i.e., the set of record pairs, which refer to the same entity). For example, correctly linking all matches would mean the match rate would be 100%. However, calculation of the match rate is difficult if the total number of matches is unknown.

Link accuracy is the proportion of links that are matches (i.e., the proportion of links which are correct). To measure this, an estimate of the number of links that are matches should be undertaken.

$$\text{Link accuracy} = \frac{\text{number of true positive links}}{\text{total links}}$$

For example, if 6000 out of 8000 links are correct, the link accuracy is 75%.

Box 1: Classification of matches and links

There are four classifications following the linking process. Records correctly assigned a link status are:

1. *True positive link* – records which refer to the same entity and have been correctly linked.

2. *True negative link* – records which do not have a match and are correctly classified as a non-link.

There are two types of linking errors:

3. *False positive link* – records that have been linked, but do not belong together (also known as *false* links).

4. *False negative link* – records that have not been linked, but do belong together (also known as *missed links* or *missed matches*).

		MATCH	NON-MATCH	
LINK STATUS	LINK	True positive link (matches that are linked)	False positive link (non-matches that are linked)	Total links
	NON-LINK	False negative link (matches that are not linked)	True negative non-link (non-matches that are not linked)	Total non-links
		Total matches	Total non-matches	Total records pairs

Source: Bishop and Khoo 2007

An example of a false positive link

Twin sisters living in the same house with the same initials and surname may have their records incorrectly paired with each other because their linking characteristics are so similar. If the linking software used surname, initials, address, sex and date-of-birth to determine the matches, it could assess their records as a 'link' because of the high level of agreement between the linkage fields.

A description of a false negative link

Even if two records on different datasets belong to the same person, the linking process may not identify them as links if:

1. there are errors in the source data (such as misspellings of name, address) and/or

2. the information changes over time (e.g., if a person changes their name after marriage, the linking software cannot tell if Jane Smith on one dataset and Jane Barnett on another are the same person, if using only name to match).

Additional approaches to checking linked data quality

Prior to analysis, there are a number of additional checks used to assess the quality of a linked dataset. This is to ensure any limitations or issues with the linked data are taken into account during analysis.

Additional checks include comparing linked datasets, clerical assessment, comparing the expected number of links with the actual number of links, and assessing discrepancies in the representation of sub-groups.

There are other more complex approaches, such as simulation-based methods, to assess linkage quality. For more information, refer to Winglee, Valliant and Scheuren (2005).

1. Comparing linked datasets

Results are compared from two separate linking methods applied to the same datasets. This may involve comparing a linked dataset with a 'gold standard file' or 'truth file' (where available) to assess the quality of the linkage strategy. For further information, see Australian Bureau of Statistics (2013).

2. Clerical assessment

Clerical assessment can be undertaken post linkage to check the quality of the link status assigned to the records. Checking link quality involves a review, usually manual, of the linked records. It is a time consuming and subjective process.

Clerical assessment is different to clerical review (see Information Sheet 4) as it examines the quality of the link following the linkage process, rather than determining if the records should be a link. For more information, see AIHW and ABS (2012).

3. Comparing the expected number of links with the actual number of links

An initial assessment of the two source datasets may yield an estimate of the number of matches. A simplified example is where a dataset containing 3000 records is known to be a subset of a dataset with 5000 records. In such a case, the estimated number of linked records on the new dataset is 3000. Therefore, if the number of links on the new dataset were significantly less than expected, this would be cause for investigation.

This method is particularly useful if one dataset is a subset (or near subset) of another, e.g., linking an administrative dataset to a census. However, even in cases where the overlap between datasets is quite small, an accurate estimate of this overlap can often be calculated.

4. Assessing discrepancies in the representation of sub-groups

Checking the characteristics of the linked dataset also provides scope for highlighting any population groups that may have been over or under represented.

Some sub-groups can be hard to correctly link. For example, young adults are more mobile, making geographic fields less reliable and children often have uniform or "not applicable" values on a number of fields (such as education, qualification, occupation, marital status). A linked dataset may be under-represented on younger individuals as a result. Refer to Bishop (2009) and Wright, Bishop and Ayre (2009).

When assessing these characteristics it is useful to consider the key factors that have influenced the linked dataset.

Box 2 – A simplified introduction to the key measures of linked data quality*

Accuracy rate	Accuracy rate is the proportion of all record pair comparisons that are true positive links or true negative links. The denominator for this rate is the number of all record pair comparisons, while the numerator is the number of record pairs that are correctly classified as true matches or false matches.
False-negative rate	False-negative rate is the proportion of all record pairs belonging to the same individuals or entities that are incorrectly assigned as non-links.
False-positive rate	False-positive rate is the proportion of all record pairs belonging to two different individuals or entities that are incorrectly assigned as links.
Precision (Link accuracy)	Precision (Link accuracy) is the proportion of all classified links that are true links as opposed to classified links that are false links. It is calculated by dividing the number of links that are ascertained as true, by the total number of classified links.
Sensitivity (match rate)	Sensitivity (match rate) is the proportion of all records in a file or database with a match in another file that were correctly accepted as a link.
Specificity or true-negative rate	Specificity or true-negative rate is the proportion of all records on one file or database that have no match in the other file that were correctly not accepted as a link.

*Sourced from Australian Institute of Health and Welfare (AIHW) and Australian Bureau of Statistics (ABS) 2012. Also see Christen and Goiser 2007.

Factors affecting linked data quality

A record may be missed or a link error may occur (a 'false link') for three main reasons.

First, the source data may contain missing or incorrect values on key linking fields. Poor quality data used to link the records can lead to missed links, or records linked to the wrong record.

The source data should be standardised prior to linking, to correct as much as possible typographical errors and out-of-date information. For more information and examples, refer to the CHeReL Quality Assurance report (2013).

Improving the source data quality will reduce the number of missed matches and thereby improve the quality of the linked dataset. For more information on editing and standardising, see Information Sheet 2.

Second, a decision point in the linking process is around the trade-off between link accuracy and match rate. A strategy aimed at producing high link quality usually comes at a cost of a reduced number of total links.

This trade-off may be influenced by data availability. For example, as discussed above, there is often limited information available relating to young persons. As a result, there may be a lower degree of certainty regarding the linkage. Therefore, only accepting links that have a high degree of certainty may lead to the linked dataset containing less linked records relating to young persons.

Third, the linking strategy influences the data quality through the blocking strategy used (e.g., the quality of the blocking fields used and the number of passes run), clerical review (such as setting threshold cut-offs) and, if applicable, the quality of the linkage keys created. See Information Sheets 3 and 4 for more details on these methods.

For more information on how the linkage strategy affects the quality of the linked data, refer to Bishop and Khoo (2007), Richter, Saher and Campbell (2013) and Australian Institute of Health and Welfare (2011).

Using documentation to improve quality

Documentation is important, as it collates and enhances an understanding of the linked data, including any factors influencing the quality of the data, prior to analysis.

Data linking projects require the documentation of every stage to allow for relevant evaluation and application in future projects. This is particularly useful when different linking strategies are utilised to assess the quality of the linked data.

Documentation also provides the analyst with a more comprehensive understanding of the linked data, providing context around how it may or may not be used for analytical purposes.

Documentation includes:

- the context and quality of the source data, including metadata statements (see the National Statistical Service website);
- risks to confidentiality and how the risks were managed (refer to the Confidentiality Information Series);
- how data has been edited and standardised before linking (see Information Sheet 2);
- the linking method (see Information Sheets 3 and 4);
- techniques used to overcome issues and/or problems; and
- results and findings from evaluation and quality assessments.

For more information

- Australian Bureau of Statistics 2009, *ABS Data Quality Framework*, Cat. no. 1520.0, ABS, Canberra.
- Australian Bureau of Statistics 2013, 'Assessing the Quality of Linking School Enrolment Records to 2011 Census Data: Deterministic Linkage Methods', Cat. no. 1351.0.55.045, ABS, Canberra.
- Australian Institute of Health and Welfare 2011, 'Comparing an SLK-based and name-based data linkage strategy: an investigation into the PIAC linkage', Data linking series, No. 11 Cat. no. CSI 11, AIHW, Canberra.
- Australian Institute of Health and Welfare (AIHW) and Australian Bureau of Statistics (ABS) 2012, *National Best Practice Guidelines for Data Linkage Activities Relating to Aboriginal and Torres Strait Islander People*, AIHW Cat. no. IHW 74, AIHW, Canberra.
- Bishop, G. 2009, 'Assessing the Likely Quality of the Statistical Longitudinal Census Dataset', *Methodology Research Papers*, Cat. no. 1351.0.55.026, ABS, Canberra.
- Bishop, G. and Khoo, J. 2007, 'Methodology of Evaluating the Quality of Probabilistic Linking', *Methodology Research Papers*, Cat. no. 1351.0.55.018, ABS, Canberra.
- Brackstone, G. 1999, 'Managing data quality in a statistical agency', *Survey Methodology*, December 1999, Vol. 25, No. 2, pp. 139-149.
- CHeReL, *Quality Assurance*, http://www.cherel.org.au/quality-assurance, viewed November 2013.
- Christen, P. and Goiser, K. 2007, 'Quality and complexity measures for data linkage and deduplication', in Guillet, F. and Hamilton, H. editors, *Quality Measures in Data Mining*, Vol. 43 of Studies in Computational Intelligence, Springer, pp. 127-151.
- National Statistical Service, *Data Quality Resources*, http://www.nss.gov.au, viewed May 2014.
- National Statistical Service 2011, *Confidentiality Information Series*, National Statistical Service, Canberra.
- Richter, K., Saher, G. and Campbell, P. 2013, 'Assessing the quality of linking migrant settlement records to 2011 Census data', Cat. no. 1351.0.55.043, ABS, Canberra.
- Winglee, M., Valliant, R. and Scheuren, F. 2005, 'A Case Study in Record Linkage', *Survey Methodology*, Cat. no. 12-001, Statistics Canada.
- Wright, J., Bishop, G. and Ayre, T. 2009, 'Assessing the Quality of Linking Migrant Settlement Records to Census Data', *Methodology Research Papers*, Cat. no. 1351.0.55.027, ABS, Canberra.

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