

Fact sheet

De-identifying research data

Pathways for the Future Program

Centre for Education Statistics and Evaluation

What is de-identification?

De-identification is the process of removing or changing information in a dataset to protect people's identities.

There are generally 2 steps to de-identification:

1. Removing personal information that directly identifies people, such as names, dates of birth and addresses.
2. Making sure the right controls are in place so that it is extremely unlikely that people could be re-identified. This includes storing the dataset securely, controlling access to the dataset, and modifying the data before publication.

For the Pathways for the Future Program, direct identifiers are removed from the research data. Identifiers such as names and addresses are separated from content information, such as subjects studied or qualification level.

De-identification means that the information is no longer considered personal information for the purposes of NSW privacy law.¹

The best method of de-identifying information will depend on the type of data, how it is stored and how it will be used.

Why de-identify data?

De-identification protects a person's privacy, allowing researchers to analyse trends in the wider population without revealing personal information in their research findings, reports or any published information.

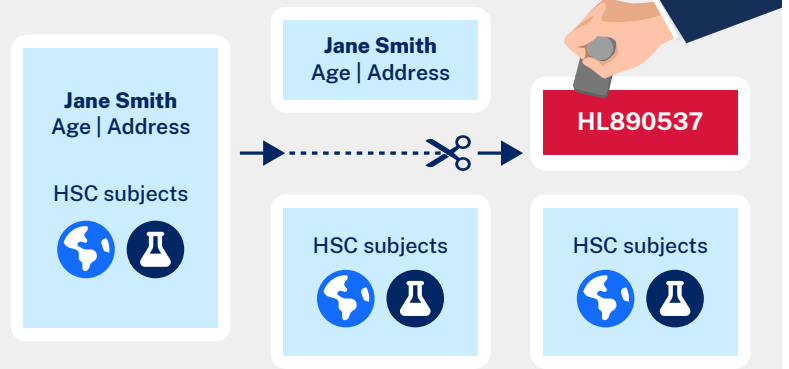
For the Pathways for the Future Program, de-identification protects privacy while allowing the department to use existing data to understand trends in the education and training system. This data is sourced from government agencies and educational and training organisations.

¹ Information and Privacy Commission NSW, *Fact sheet – de-identification of personal information*, May 2020, accessed 24 February 2025; the applicable NSW privacy laws are the *Privacy and Personal Information Protection Act 1998* and *Health Records and Information Privacy Act 2002*.

How data linkage works

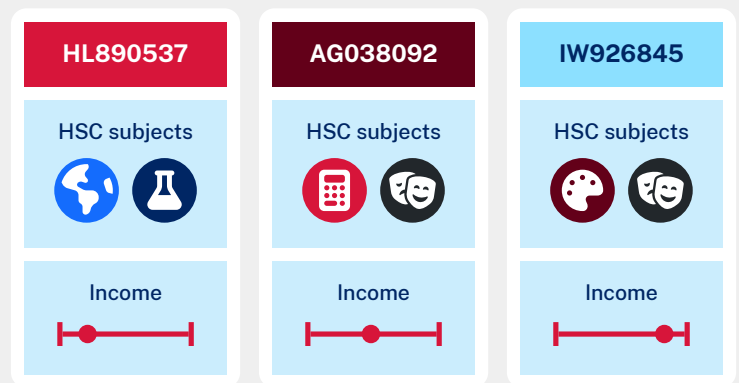
1 Splitting and removing direct identifiers

- Datasets are supplied by government agencies and education and training organisations.
- The linkage agency:
 - splits identifier information (such as names and addresses) from content information (like HSC subjects). Identifier information and content information are handled by different teams
 - gives each individual a random reference number which is used instead of identifying details.



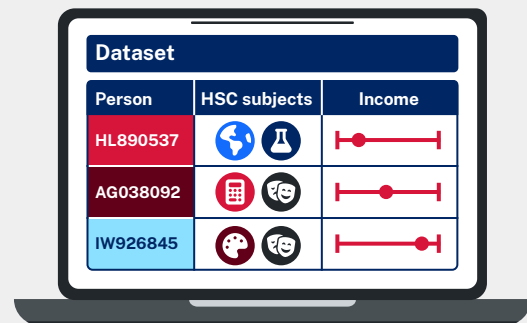
2 Data linkage

- Various matching methods are used to distinguish between different individuals and combine information from different datasets.
- Once the data is linked together, all records for an individual will have the same random reference number.



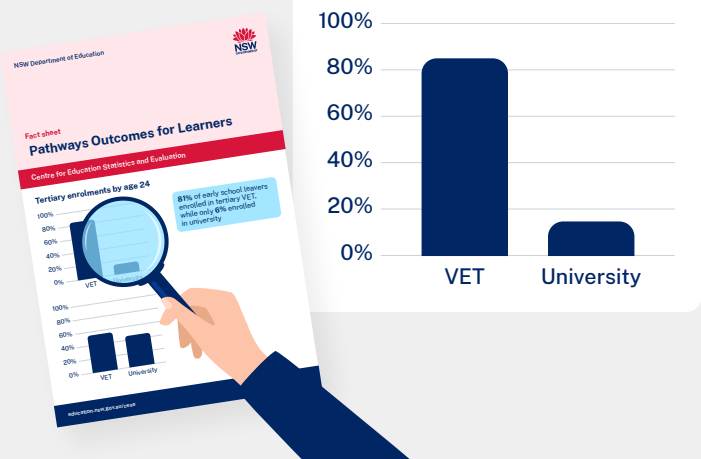
3 Final dataset

- The final dataset is hosted on the secure Australian Bureau of Statistics (ABS) DataLab. It does not contain any personal information that directly identifies people (such as names, dates of birth or addresses).
- Researchers must be approved and meet strict requirements to access the dataset.



4 Publications

- Datasets are used to create publications.
- Publications go through a rigorous output checking process to minimise the risk of re-identification. These controls make it extremely difficult, if not impossible, for a person to be re-identified in the data.



Can information be re-identified?

After personal information is removed, additional controls must be put in place to make it extremely difficult, if not impossible, for a person to be re-identified in the data.

Results are grouped together (aggregated) so individuals cannot be re-identified.

Techniques for further reducing the risk of re-identification can include:

- generalising information and combining categories. For example, instead of listing data for each age (20, 21, 22 and so on), insights can be categorised into age brackets (21–25, 26–30)
- suppressing data, which involves deleting or not releasing cells that may have a higher risk of re-identification
- changing the data in ways that do not impact the accuracy of the results but make it harder to re-identify. For example, rounding involves slightly altering small values in a table so that the original values are unclear. Swapping may involve exchanging values with similar characteristics.

All outputs for the Pathways for the Future Program are subject to the ABS DataLab's output clearance rules, which are designed to reduce the risk of re-identification. These rules are applied to any data being removed from the DataLab.

All results are aggregated or grouped to make sure individuals cannot be identified.

For example, in the [2023 early school leavers fact sheet](#), high-level trends and findings include:

- 30% of students left school early
- 81% of early school leavers enrolled in vocational education and training (VET)
- the top occupation for early school leavers was technicians and trades workers.

Further information

- [Information and Privacy Commission – De-identification of personal information](#)
- [Centre for Health Record Linkage – Our services](#)
- [Australian Bureau of Statistics – Data integration](#)

If you have any questions about de-identification and the Pathways Program, you can contact the team at Pathways.Project@det.nsw.edu.au.