information and privacy commission
new south wales

Fact Sheet **May 2020**

De-identification of personal information

NSW privacy law[1] places obligations on agencies for how they collect, store, use and disclose personal information. Personal information is defined in the *Privacy and Personal Information Protection Act* (PPIP Act) as any information or opinion (including information or an opinion forming part of a database and whether or not recorded in a material form) about an individual whose identity is 'apparent' or can 'reasonably be ascertained' from the information.

This fact sheet discusses the importance of de-identification and offers practical tips to agencies when de-identifying information.

What is de-identification?

The term 'de-identification' is not defined in NSW privacy law. In this fact sheet, it is understood by reference to the meaning of personal information in the PPIP Act. [2]

De-identification means that a person's identity is no longer apparent or cannot be reasonably ascertained from the information or data. De-identified information is information from which the identifiers about the person have been permanently removed, or where the identifiers have never been included.

This means that the information is not personal information for the purposes of the PPIP Act.

The NSW Civil and Administrative Tribunal (NCAT) has confirmed that de-identification of an individual's face, head and neck through pixilation of CCTV footage is no longer personal information because the individual can no longer be identified from the information.[3]

For further guidance, agencies should refer to Fact Sheet, Reasonably Ascertainable Identity.

Why de-identify?

De-identification is important because it can make available data sources to agencies and enable information to be used while preserving an individual's privacy. De-identification is also important to community expectations about how agencies handle personal information.

Importantly, de-identification can protect against an individual's or a group of individuals' identities from being revealed.

Other reasons for de-identification include:

- to protect against breaches of NSW privacy law and assist agencies to comply with the conduct expected by the information protection principles
- to enable an agency to make the best decisions about whether to safely release and share data, including within the agency or with another agency or third party
- to enable agencies to achieve the objective of research by the use of de-identified or anonymised information, without having to claim the exemption in section 27B of the PPIP Act
- to mitigate risk and minimise the harm caused to individuals from a data breach
- to build community trust in how agencies store and handle data.

How can an agency de-identify data?

The key to de-identification is the removal of the 'identifiers' of personal information so that the information is not about an identifiable person.

Examples of a direct identifier include, an individual's name, address, telephone number or Tax File Number.

An indirect identifier allows information to be connected until an individual can be identifiable. Examples can include, a client number, vehicle registration number, or demographic data such as date of birth and gender.

In the privacy context, for information to be truly 'de-identified', it must be extremely difficult, if not impossible, for re-identification of the person to occur.

[1] *Privacy and Personal Information Protection Act 1998* and *Health Records and Information Protection Act 2002*.
[2] PPIP Act, section 4.
[3] *Seven Network Limited v South Eastern Sydney Local Health District* [2017] NSWCATAD 210 at [62].

In research contexts, ‘de-identified’ may be referred to as ‘anonymised’ to describe data for which codes or numbers have replaced names or other personal identifiers.

There is no correct way to de-identify data. It may require the agency to put controls and safeguards in place to avoid the risk of re-identification.

Care should be taken to choose the most suitable method by considering the type of data, its intended use and reasons for its use, and the access environment where the information is to be provided.

Techniques for de-identifying data

Techniques can include:

- redacting information, including through pixelation in video and digital footage
- aggregating data
- removing some variables
- coding or pseudonymising (replacing identifiers with unique, artificial codes)
- hashing (one-way encryption of identifiers)
- generalising (e.g. by replacing precise date of birth with an age bracket)
- suppressing (e.g. by replacing some values with ‘missing’)
- micro-aggregating (e.g. group in fours, so ages 31, 32, 33 and 34 each become 32.75)
- data-swapping (e.g. swap salaries for people within the same postcode, so the aggregate is still valid)
- differential privacy (describing or analysing the patterns of groups within the dataset while withholding information about individuals).

How can re-identification occur?

The risk of re-identification is the key concern when de-identifying information. If re-identification occurs, the information is no longer de-identified information but is personal information.

If the re-identified information has been used or disclosed for a purpose different to the purpose for which the information was collected, then the agency may have breached privacy laws.

Agencies should be aware that altering the information by removing a person’s name, address or other direct identifier, may not necessarily mean that the information (or data) is de-identified.

Re-identification can occur through linkages of the de-identified information with other information or contextual indicators. It may occur through the application of data matching techniques so that data is no longer anonymised.

The NCAT has applied a test of whether or not ‘more than moderate steps’ are necessary to match data from different sources, in order to ascertain an individual’s identity.[4]

Circumstances where there is a risk of re-identification can include:

- coded information will remain potentially re-identifiable to a person or body with the means to link the code back to other identifying details
- flaws or a weakness in the technique used to encrypt information in the dataset, allowing the encryption to be reversed
- highly detailed information in the dataset can create a significant risk that some individuals may be identified by linking with other sources
- unique or rare characteristics of the individual, or a combination of unique or remarkable characteristics, can enable identification
- machine identification, despite the personal information being redacted and not able to be read by a human eye[5]
- public access being given to large datasets, such as complaint and feedback forms made publicly available on an agency website.

Examples of re-identification

The risk of re-identification is presented by the following examples from a range of jurisdictions:

- In 2020, the Australian Federal Court published the names and dates of birth against the pseudonyms given to protection visas applicants on the publicly available Commonwealth Courts searchable database (the pseudonyms were a collection of letters and numbers). The applicants’ personal identifiers could be matched with the information of authorities in foreign countries.
- In 2018, the protection of sensitive client information and feedback on the online database of Family Planning NSW was compromised by a ransomware attack on the website. While the web client feedback form was not connected to internal medical records, it was unclear how much personal information was required to be submitted by client feedback.
- In 2018, a large dataset by Public Transport Victoria was released online which contained

[4] *AIN v Medical Council of New South Wales* [2016] NSWCATAD 5 at [39] – [44].

[5] *AIN v Medical Council of New South Wales* [2016] NSWCATAD 5 at [32].

information from 15 million ‘myki’ travel cards to support a datathon event. Although the dataset was claimed to have been de-identified, the dataset recorded 1.8 billion myki ‘tap on’ and ‘tap off’ events between July 2015 and June 2018, exposing myki card users’ travel histories. The Office of the Victorian Information Commissioner’s investigation found that public transport history can contain a wealth of information about a person’s private life.

- In 2016, the datasets of the Medicare Benefits Schedule and Pharmaceutical Benefits Schedule containing patients’ claims between 1984-2014 were removed from the website of the Commonwealth Department of Health, after it was found that individuals could be identified. While the provider ID numbers were encrypted or put into special code, the medical treatments were not encrypted, and it was possible to find out some service provider ID numbers.
- In 2014, ‘de-identified’ data on 173 million taxi trips made in New York City was released under FOI. Within hours the ‘hashed’ driver and vehicle numbers were re-identified, and then the GPS data was matched with other publicly available data to identify specific trips taken by known individuals.
- In 2013, a University professor re-identified the names of more than 40% of a sample of ‘anonymous’ participants in a high-profile DNA study.
- A 2000 study linked public ‘anonymous’ health insurance data of public servants with electoral rolls (name, date of birth, sex, postcode) to identify the Massachusetts Governor’s diagnoses and prescriptions.

Further information

Data Analytics Centre (DAC) - Data protection and sharing principles

DAC is part of the Department of Customer Service which makes available a secure and well-governed central data platform for facilitating data-sharing and inter-departmental collaboration. Agencies are encouraged to consult the DAC on data sharing guidance for service delivery and projects https://data.nsw.gov.au/nsw-data-analytics-centre.

Agencies can access the NSW DAC Data Sharing Principles here: https://data.nsw.gov.au/data-sharing-principles

Office of the Australian Information Commissioner (OAIC) has published information about de-identification on its website. De-identification and the Privacy Act: https://www.oaic.gov.au/privacy/guidance-and-advice/de-identification-and-the-privacy-act/

De-identification decision-making framework: https://data61.csiro.au/en/Our-Research/Our-Work/Safety-and-Security/Privacy-Preservation/De-identification-Decision-Making-Framework

Australian Computer Society (ACS) Data Sharing Framework can be downloaded from the ACS website: https://www.acs.org.au/insightsandpublications/reports-publications/data-sharing-frameworks.html

The UK has published a comprehensive code on methods for de-identifying data, which provides some useful general guidance: https://ico.org.uk/media/1061/anonymisation-code.pdf

For more information

Contact the Information and Privacy Commission NSW (IPC):

Freecall:	1800 472 679
Email:	ipcinfo@ipc.nsw.gov.au
Website:	www.ipc.nsw.gov.au

The IPC can give general advice on rights and compliance under privacy and information access legislation, but cannot give legal advice. You should seek your own legal advice as required.