Applying the legislation

GUIDELINE *Information Privacy Act 2009*

Privacy and de-identified data

Personal information is information about *identifiable* individuals. All Queensland government agencies deal with personal information. In doing so, they must comply with the privacy principles in the *Information Privacy Act 2009* (Qld) **(IP Act**).

De-identification involves removing or altering information that identifies an individual or is reasonably likely to enable their identification. This may enable agencies to release information about individuals while still complying with the privacy principles.[1] De-identification can be technically complex and often requires specialist advice.

De-identified data is also at risk of 're-identification'. This often occurs when de-identified data is linked with other external information. Re-identification can reveal personal information and may breach the privacy principles. When agencies release de-identified data, they must adequately manage re-identification risk to protect the identity of individuals and their personal information.

Scope of this guideline

This guideline is not a step-by-step guide on managing privacy when de-identifying data. It is intended to provide high-level information on de-identification and some issues an agency should consider before undertaking a de-identification process.

References to comprehensive de-identification resources have been included for more detailed advice on the concepts and techniques set out in this guideline.

This guideline addresses de-identifying data and information containing words and numbers. This guideline does not cover de-identification of images.

De-identification techniques

There is no one 'right' way to de-identify data. There are many de-identification techniques that can protect privacy and ensure data is still useful for its intended purpose.

[1] The definition of personal information varies in different jurisdictions. Agencies should be aware that different legal considerations may apply when publishing information on the internet.

Selecting an effective de-identification technique, or a combination of techniques, requires a sound understanding of the data itself. Direct identifiers in data are likely to be obvious (such as name, address, driver licence number, telephone number). However, data can also contain other unique values that, while not personal information on their own, can quickly identify an individual when linked with external 'auxiliary information'.

Hint

De-identification is not just removing obvious personal information. Simply removing direct identifiers, like names and date of birth, is not always sufficient to adequately de-identify data and manage re-identification risk.

Some de-identification techniques include:

- **Suppression**—removing data that may identify individuals or which in combination with other information is reasonably likely to identify an individual.
- **Rounding**—grouping identifiers into categories or ranges. For example, age can be combined in ranges (25-35 years) rather than single years (27, 28). Extreme values can also be grouped in a range, such as an age value of '85+ years'.
- **Perturbation**—altering data that is likely to enable the identification of an individual in a small way, such that the aggregate information or information is not significantly affected but the original values cannot be known with certainty. For example, randomly adding or subtracting 1 to a person's year of birth.
- **Swapping**—swapping information that is likely to enable the identification of an individual for one person with the information for another person with similar characteristics to hide the uniqueness of some information.
- **Sampling**—when large numbers of records are available, it may be adequate to release a sample of records. This can create uncertainty that a person is included in the sample.
- **Generating synthetic information**—mixing up the elements of an dataset—or creating new values based on the original information—so the overall totals, values and patterns of the data are preserved but do not relate to any particular individual.
- **Encryption or 'hashing' of identifiers**—data is encrypted or obscured using a scheme that enables accurate analytics to be performed on it, while never revealing the encrypted raw data.

Agencies should seek expert advice to understand their data and determine the appropriate de-identification technique(s).

Balancing privacy and data utility

Agencies can face a privacy/utility trade-off when de-identifying data. It may be tempting to extensively de-identify data to lower re-identification risk, particularly where there are significant threats in the external environment. However, this can introduce new problems.

Effective de-identification may reduce data utility. In some cases, de-identification may render data useless, or potentially misleading. Agencies must balance this trade-off when applying de-identification techniques.

Tip

De-identification is one way to manage privacy risks when sharing data. Agencies can also implement technical and administrative controls to manage the 'who', 'what', 'where', and 'how' of accessing information.

Applying administrative safeguards and controls can reduce the risk of re-identification and better preserve the utility or richness of the data being released.

Examples of administrative controls and safeguards include:

- sharing only the data necessary to achieve the intended purpose
- specifying who is permitted to access the data
- allowing access only within a controlled environment
- arranging for the destruction or return of data on completion of the project; and
- using a data sharing agreement to limit use and disclosure of information, including a prohibition on any attempt at re-identification and specifying that all analytical outputs must be approved by the agency before they are published.

Re-identification

A re-identification event may breach the IP Act and disclose personal information about individuals. It also has the potential to undermine public trust in government and discourage other agencies from sharing information.

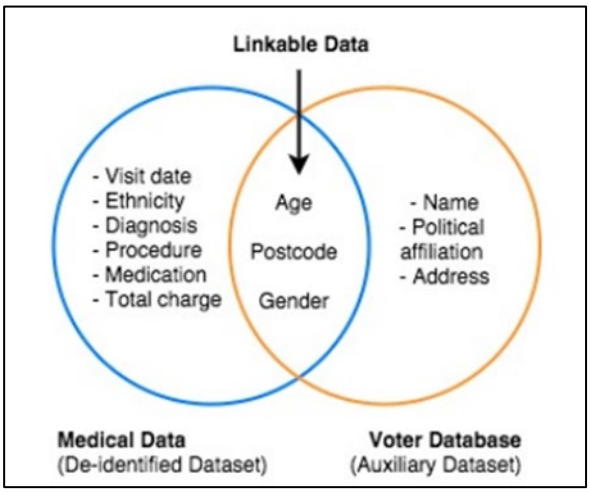
Re-identification often occurs when data is combined with external 'auxiliary information' to reveal information about an individual.

Some examples of auxiliary information include:

- other public datasets and information, including social media
- non-public datasets, for example, a business's customer database

- personally observed information, for example, overhearing a conversation or witnessing an event.

The following figure demonstrates how indirect identifiers (age, postcode, gender) can be linked with an auxiliary dataset containing personal information. In this example, re-identification could reveal an individual's demographic and medical information.



Source: Office of the Victorian Information Commissioner

Re-identification can also occur without auxiliary information, for example:

- **Inadequate de-identification**—where identifying information is inadvertently left in the data.
- **Pseudonym reversal**—if an algorithm with a key is used to assign pseudonyms, it can be possible to use the key to reverse the pseudonymisation process to reveal identities.
- **Inferential disclosure—**when personal information can be inferred with a high degree of confidence from statistical attributes of the data.[2]

Assessing re-identification risk

Before releasing de-identified data, agencies must assess whether the de-identification techniques they have chosen, and any safeguards and controls they applied to the release environment, adequately manage the risk of re-identification.

Like any risk, it is not possible to reduce re-identification risk to zero. However, privacy does not require that de-identification be absolute. The level of re-identification risk will vary with the sensitivity and intended use of the data. For

[2] Office of the Victorian Information Commissioner's De-identification Background Paper

example, an agency may tolerate a higher re-identification risk when sharing data with another agency than it would publishing information on its open data portal.

Hint

Agencies should fully understand their data and consider the wider risk environment. While it is not possible to foresee every possible re-identification scenario, agencies should conduct a detailed re-identification risk assessment prior to releasing de-identified data.

When assessing re-identification risk, it is not sufficient to simply ask 'how risky is the data?'. Agencies must also ask themselves 'how might a re-identification event occur?' and 'what new information could be learnt?'.

When looking at the data itself, agencies should understand the data format, whether the data has unique values, and who the information is about. This includes:

- **Format and structure of original information**—for example, is the data unit-level or aggregated?
- **Uniqueness**—does the data contain unique values that could be used to re-identify an individual? Does data become unique when combing multiple attributes (for example, age and postcode)?
- **Type and strength of de-identification technique(s) applied**—how technically complicated is re-identification?
- **Sensitivity**—for example, does the data contain information about vulnerable individuals?
- **Other safeguards and controls**—is the data protected by other access and usage controls, or is it published as open data?

When considering the wider data environment, agencies should think about risk scenarios, including the 'who, what and how' of a re-identification event. This might include:

- **Motivation and ability of attacker**—who is likely to attempt re-identification? Is an attacker likely to have the necessary technical capability?
- **Auxiliary information**—what additional information is needed to re-identify? How likely is an attacker to have this information?
- **Consequences of re-identification**—what new information could an attacker learn and how sensitive is this information?

At a minimum, applying a 'motivated intruder test'—assessing whether a reasonably competent motivated person with no specialised skills could succeed in re-identifying the information—is a good initial risk indicator.

Conducting this sort of assessment often requires specialist expertise, particularly if there needs to be a high degree of confidence that no individuals can be reasonably re-identified (for example, where information will be published in an open data environment).

Tip

Handling de-identified information may still carry certain privacy risks. It may be necessary to handle de-identified information in a way that would prevent a privacy breach.

For example, Agency A de-identifies information for use by Agency B: a privacy breach could occur if the de-identified information is made available in another environment, for example if Agency B inadvertently publishes it on its website and it can be re-identified by linking it with other information.

While the IP Act may not apply to data that is de-identified in a specific context, the same data could become personal information in a different context.

Managing re-identification risk over time

De-identification is not a fixed state. Like other risks, re-identification risks and their controls require ongoing monitoring and review. The risk of re-identification increases as technology develops and/or as more 'auxiliary information' is published or obtained by a person or entity.

Agencies should regularly monitor and review the risk of re-identification and, if necessary, take further steps to minimise the risk. In more extreme cases, they may consider removing or restricting access to the data.

Releasing de-identified data on public platforms

Queensland government agencies are encouraged to proactively release data on public platforms.[3] This supports the 'push model' and the proactive disclosure aims of the *Right to Information Act 2009* (Qld). Releasing data also supports transparent and accountable government.

Releasing data that contains, or is derived from, personal information requires rigorous privacy risk management. This could include information about an individual's gender, age, access to services and the location of individuals at a point in time.

Unlike sharing data in closed environments, it is not possible to control the access and use of public data. Rapid changes in technology and the increasing volume

[3] Queensland Government Open Data Policy Statement

of public information available make managing re-identification risk on public platforms more complicated.

Noting these risks, agencies should carefully consider the need to release de-identified data on public platforms. If they choose to release de-identified data, agencies must be confident re-identification risk is managed over the lifecycle of the data. Where re-identification risk cannot be safely managed, agencies should consider sharing data with relevant users in a controlled environment.

Note

Publishing de-identified data on public platforms can be high risk. The OIC strongly recommends agencies seek expert advice to fully understand these risks and rigorously de-identify data before publishing.

De-identification governance processes

De-identification is complex, with a range of factors to consider at each point in the process. A strong governance framework supports effective re-identification risk management. This includes:

- registers that capture all de-identified datasets and their custodians
- sufficient guidance on de-identification techniques and re-identification risk management, including a defined re-identification risk tolerance
- procedures to adequately capture, assess and treat re-identification risks
- systems to test re-identification risks management strategies effectively treat risks; and
- processes to monitor the external data environment and regularly review existing de-identified datasets for changes in re-identification risk.

Additional resources

Comprehensive resources on de-identification include the De-Identification Decision-Making Framework, produced jointly by the OAIC and CSIRO's Data61 and the United Kingdom Information Commissioner's Office's Anonymisation: managing data protection risk code of practice.

For additional information and assistance please refer to the OIC's privacy guidelines or contact the Enquiries Service on 07 3234 7373 or email enquiries@oic.qld.gov.au.

This guide is introductory only, and deals with issues in a general way. It is not legal advice. Additional factors may be relevant in specific circumstances. For detailed guidance, legal advice should be sought.

If you have any comments or suggestions on the content of this document, please submit them to feedback@oic.qld.gov.au.

Published 1 February 2019 and last updated 9 July 2020

Changes to legislation after the update date are not included in this document