Applying the legislation

GUIDELINE Information Privacy Act 2009

Privacy and De-identification

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).

If information is not about an identifiable individual it is not personal information, which means it presents no privacy concerns. De-identifying an individual’s information enables it to be shared or made publicly available while still complying with the privacy principles, protecting the privacy of the individual, and ensuring the information remains appropriate for its intended use.

Appendix A to this guideline is a quick guide to de-identification, with questions to walk you through the de-identification process.

What is de-identification?

De-identification involves removing or altering information that identifies an individual or is reasonably likely to enable their identification. To de-identify effectively, requires consideration of not only the information itself, but also the context or environment in which the information will be made available.

Scope of this guideline

This guideline is not a step-by-step guide on how to de-identify information. It is intended to provide high-level information on de-identification and the issues an agency should consider before undertaking a de-identification process.

De-identification is a complex topic. References to comprehensive de-identification resources have been included for agencies who want more detailed advice on the concepts and techniques set out in this guideline.

This guideline addresses de-identifying information that consists of words and/or numbers. The de-identification of images is not covered by guideline.

It is important to note that de-identification is a risk management exercise, not an exact science and that de-identification is a contentious contemporary privacy issue. Some commentators believe that, given the evolution of digital information capabilities, personal information can never be truly de-identified.

If agencies are undertaking complex de-identification exercises, for example when de-identifying sensitive information sets or where the de-identified information is to be made public, the Office of the Information Commissioner (OIC) recommends seeking specialist expertise.
When is an individual ‘reasonably identifiable’?

Personal information is any information about an individual whose identity is apparent or can reasonably be ascertained.\(^1\)

Whether information is about a ‘reasonably identifiable’ individual requires case-by-case consideration of factors such as:

- the nature and amount of information
- who will hold and have access to the information
- whether there are other available information sets containing identifiable information that could be used to cross-match or link the information to an individual; and
- the difficulty, motivation, time or cost of using that information to identify an individual.

De-identification has a practical component. It does not require an absolute surety that the individual cannot (or will not) be re-identified. It only requires that there be no reasonable likelihood that re-identification will occur.

The de-identification process

In general, the de-identification process involves:

- removing direct identifiers (such as name, address, driver licence number, telephone number, photograph or biometrics); and
- removing or altering the information to protect indirect or quasi-identifiers (such as date of birth, gender, profession, ethnic origin, marital status etc) that might permit identification when used in combination with other information.

This is followed by applying appropriate technical and administrative safeguards or controls, such as restrictions on who can access the information and specific physical/technical measures.

**Tip**

Making information publicly available or available to a wide audience may require considerable modification to the information, which may significantly lower its utility.

There is little point in making information available if the information does not actually represent whatever it is meant to represent. At worst, the information could lead to misleading conclusions which might have significant consequences if, for example, the information is used to influence policy or to make decisions.

Releasing the information into a controlled environment may allow the utility of the information to be better preserved. Working through the following three

\(^1\) See section 12 of the IP Act for the full definition.
points will help with decisions about choosing an appropriate means by which to release the information:

1. Clarify the reason for wanting to share or release your information.
2. Identify the groups who will access your information.
3. Establish how those accessing your information might want to use it.

De-identification techniques

There is no one ‘right’ way to de-identify information. There are a number of de-identification controls and techniques which can be used to both protect individuals’ privacy and ensure the information is still useful for its intended purpose. Which specific techniques should be used, and in what combination, depends on the environment in which the information will be released and may require expert advice.

Hint

Direct identifiers can generally be removed from the information relatively simply by either deleting or replacing them with a generic word or randomised number. However, the removal of direct identifiers is not always sufficient to appropriately de-identify the information.

Examples of de-identification techniques include:

- **Suppression**—removing information 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 and expressed in ranges (25-35 years) rather than single years (27, 28) or the more detailed date of birth. Extreme values above an upper limit or below a lower limit can be placed in an open-ended range such as an age value of ‘less than 15 years’ or ‘more than 80 years’.

- **Perturbation**—altering information 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. For example, a person from a particular town in Australia may speak a language that is unique in that town. Information about that individual’s spoken language could be swapped with the spoken language information for another individual with otherwise similar characteristics (based on age, gender, income or other characteristics as appropriate) in an area where the language is more commonly spoken.
• **Sampling**—when very large numbers of records are available, it may be adequate for analysis purposes to release a sample of records, selected through some stated randomised procedure. This creates uncertainty that any particular person is included in the information set.

• **Encryption or ‘hashing’ of identifiers**—information is encrypted or obscured using a scheme that enables accurate analytics to be performed on it, while never revealing the encrypted raw information.

• **Generating synthetic information**—mixing up the elements of an information set—or creating new values based on the original information—so that all of the overall totals, values and patterns of the set are preserved but do not relate to any particular individual.

### Choice of de-identification technique

Choosing an appropriate de-identification technique requires consideration of contextual factors, including:

- the nature of the information to be de-identified
- who will have access to the information, and what purpose this is expected to achieve
- whether the information contains unique or uncommon indirect or quasi-identifiers that could enable re-identification
- whether the information will be targeted for re-identification because of who or what it relates to
- whether there is other information or information available that could be matched up or used to re-identify the de-identified information; and
- what harm may result if the information or information is re-identified.

### Technical and administrative safeguards or controls

Technical and administrative safeguards and controls go to the ‘who’, ‘what’, ‘where’, and ‘how’ of accessing information. Applying safeguards and controls can reduce the risk of re-identification and better preserve the utility or richness of the information being released.

Examples of controls and safeguards include:

- including only the information necessary to achieve the intended purpose
- specifying who is permitted to access the information
- allowing access only within a controlled environment
- ensuring that those given access to the de-identified information cannot access the original information
- making arrangements for the destruction or return of the information on completion of the project
- enabling an analysis of information rather than providing access to it, for example, running an analysis of the information and providing the result rather than the raw information; and
- using an information sharing agreement or a memorandum of understanding 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.

**Assessing the risk of re-identification**

Before releasing de-identified information, it is important to assess whether the chosen de-identification techniques, and any safeguards and controls applied to the environment in which the information will be released, are appropriate to manage the risk of reidentification.

In many (if not all) instances where a de-identification process is undertaken, the risk of re-identification will never be totally eliminated. However, privacy does not require that de-identification be absolute. Privacy will be satisfied, and information will be suitably de-identified, where the risk of an individual being re-identified is very low.

Re-identification generally occurs through:

- **Poor de-identification**—where identifying information is inadvertently left in the information.
- **Data linkage**—it can be possible to re-identify individuals by linking de-identified information with an ‘auxiliary dataset’ that contains identifying information.
- **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**—this occurs when personal information can be inferred with a high degree of confidence from statistical attributes of the information.2

The following figure demonstrates how indirect identifiers (age, postcode, gender) can be linked with an auxiliary dataset containing personal information, enabling the two datasets to be linked and an individual’s identity to be revealed.

---

2 Office of the Victorian Information Commissioner’s De-identification Background Paper
Assessing the risk of re-identification involves considering various factors, including:

- content, value and structure of original information
- type and strength of de-identification technique(s) applied
- technical skill and resources of the ‘attacker’; and
- availability of other information that can be linked with the de-identified 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 indicator of the level of risk. Conducting this sort of assessment often requires specialist expertise and it can be helpful to engage an expert, particularly if there needs to be a high degree of confidence that no individuals are reasonably identifiable (for example, where information will be published in an open data environment).

**Considerations after the de-identified information has been released or shared**

De-identification is not a fixed state. The risk that de-identified information could be re-identified may increase as technology develops and/or as more information is published or obtained by an entity. This means agencies should regularly re-assess 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 originally de-identified information.

As a consequence of this expanding digital environment, and the increasing availability of linkable data, agencies should consider limiting their open data environments to:

- information not derived from personal information; or
- information that has been subject to an extremely robust de-identification process.

**Tip**

Handling de-identified information may still carry certain privacy risks and 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.

Agencies are encouraged to take a risk-management approach when handling de-identified information which acknowledges that 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.
De-identification governance processes

The above information has highlighted that de-identification can be complex, with a range of factors to consider at each point in the process. Having a strong governance framework in place will help to ensure that de-identification is carried out effectively. Robust de-identification governance includes activities such as:

- ongoing and regular re-identification risk assessments (to check that techniques and controls are still effective and appropriate at managing the risks involved)
- auditing recipients to ensure that they are complying with the conditions of any information sharing agreements or memorandum of understanding
- considering new information that becomes available, and whether any such information increases re-identification risk in the environment to which it was released; and
- ensuring that staff who undertake de-identification have adequate and up-to-date training, and/or ensuring appropriate external expertise is sought where appropriate.

Additional resources

The guideline is adapted from De-identification and the Privacy Act, published by the Office of the Australian Information Commissioner (OAIC).

Other comprehensive resources on the topic of 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 1 February 2019

Changes to legislation after the update date are not included in this document.
### Appendix A - Quick Guide to De-identification

**Questions to Consider When De-identifying**

<table>
<thead>
<tr>
<th>What do you know?</th>
<th>Understand the nature of your data, as well as the other data, people, infrastructure and governance associated with your data.</th>
</tr>
</thead>
</table>
| What are your legal responsibilities? | Know which laws apply to your dataset and what obligations they impose.  
This may include the IP Act and other legislation. |
| What is your data like? | Focus on the data type, features and properties. This involves the data subjects, variables, quality and age. This is important in assessing the re-identification risk. |
| What is the use case? | Know why you want to share your data, which groups will have access to the data, and how those groups might want to use the data:  
- What knowledge could the user of the data already have or be able to get?  
- Is it Open Data, semi-closed, or controlled release in a secure research environment?  
- Does it require the use of unit level data? Will aggregated data achieve the same result? |
| What are your ethical obligations? | Consider, for example, consent, transparency, stakeholder engagement and governance. |

---

3 Based on the work of Christine O’Keefe, Research Scientist – Data 61, CSIRO, Ten questions you should ask before sharing data about your customers, The Conversation, October 9, 2017 and the OAIC and CSIRO’s De-identification Decision-Making Framework.
| What is the risk that data will be re-identified? | • Does the data contain unique or uncommon characteristics (quasi-identifiers) that could enable re-identification?  
• Is there an increased risk the data will be re-identified because of who or what the data relates to? For example, high profile individuals or health data.  
• What other information or data exists in the data environment? How does it overlap with or connect to the data? Could other available data or information be matched up with or used to re-identify the de-identified data? |
|---|---|
| What processes will you need to go through to assess disclosure risk? | Establish plausible attack scenarios using risk assessment methods.
For example, someone trying to re-identify their neighbour in a local council dataset using characteristics they can easily observe, such as size of family, number of cars, and whether the home has reverse cycle air conditioning. |
| What is the potential harm that may result if the data is re-identified? | For example, is the information about domestic violence victims? |
| What are the relevant disclosure control processes? | This includes selecting the appropriate data sharing mechanism (such as open data or secure transfer to a single partner) and appropriate data modification methods, including possible reducing the amount of data under consideration. |
| Who are your stakeholders and how will you communicate with them? | Stakeholders could include data subjects, the general public, partner organisations, the media, funders |
Right to Information and Privacy Guidelines

| **What happens next, once you have shared or released the data?** | and special interest groups. Trust and credibility must be built. |
| **What will you do if things go wrong?** | This includes keeping a register of all the data you have shared or released. |
| | Have a plan to respond to a disclosure in the event one were to occur. Such measures include having a robust audit trail, a crisis management policy and adequately trained staff. |