Australian Government

Office of the Australian
Information Commissioner

De-identification and the Privacy Act

oaic.gov.au

March 2018

OAIC

Contents

Key points

- De-identification is a privacy-enhancing tool. When done well, it can help your entity meet its obligations under the Privacy Act and build trust in your data governance practices.
- Information that has undergone an appropriate and robust de-identification process is not personal information, and is therefore not subject to the *Privacy Act 1988* (Cth) (Privacy Act). Whether information is personal or de-identified will depend on the context. Information will be de-identified where the risk of an individual being re-identified in the data is very low in the relevant release context (or data access environment).[1] Put another way, information will be de-identified where there is no reasonable likelihood of re-identification occurring.
- De-identification involves two steps. The first is the removal of direct identifiers. The second is taking one or both of the following additional steps:
  - the removal or alteration of other information that could potentially be used to re-identify an individual, and/or
  - the use of controls and safeguards in the data access environment to prevent re-identification.
- This guide provides high-level guidance only. The OAIC recommends that entities also refer to the *De-Identification Decision-Making Framework*, produced jointly by the OAIC and CSIRO-Data61, which provides a comprehensive framework for approaching de-identification in accordance with the Privacy Act.
- The OAIC recommends that entities seek specialist expertise for more complex de-identification matters - for example when de-identifying rich or detailed datasets, where data may be shared publicly or with a wide audience, or where de-identification is carried out in the context of a multi-entity data sharing arrangement.[2]

Overview

Many entities[3] collect and retain information or data that includes personal information.[4] In doing so they must comply with the Australian Privacy Principles (APPs) in the *Privacy Act 1988* (Cth) (Privacy Act). The APPs regulate how entities collect, use, disclose and store personal information. In particular, APP 6 places limits on the use and disclosure (sharing or releasing) of personal information. However, where information has been appropriately de-identified, it is no longer

[1] The data access environment means the context that the data will be made available in. The data access environment is made up of four key components: other data, people, infrastructure, and governance structures. For further information see the OAIC and CSIRO's Data61 De-identification Decision Making Framework.

[2] Entities requiring more detailed information should seek specialist advice, and refer to the OAIC/CSIRO Data61 De-identification Decision Making Framework.

[3] The obligations in the Privacy Act apply to 'APP entities'. This term refers to most Australian and Norfolk Island Government agencies, all private sector and not-for-profit organisations with an annual turnover of more than $3 million, all private health service providers, and some small businesses.

[4] 'Personal information' is defined in s 6(1) of the Privacy Act, as information or an opinion about an identified individual or an individual who is reasonably identifiable. For more information, see the OAIC's guidance on the meaning of personal information.

personal information and can therefore be used or shared in ways that may not otherwise be permitted under the Privacy Act.

This resource provides general advice about de-identification, to assist APP entities to protect privacy when using or sharing information containing personal information. Guidance is provided on when de-identification may be appropriate, how to choose appropriate de-identification techniques, and how to assess the risk of re-identification. It is consistent with the approach taken in the *De-Identification Decision-Making Framework*.

The fundamental premise underpinning this guidance is that re-identification risk must be assessed contextually. To de-identify effectively, entities must consider not only the data itself, but also the environment the data will be released into. Both factors must be considered in order to effectively determine which techniques and controls are necessary to de-identify the data, while ensuring it remains appropriate for its intended use.

Importantly, entities should be aware that it is not always possible to draw a bright line between personal and de-identified information. De-identification is a risk management exercise, not an exact science. However, in all cases, for data to be considered 'de-identified', the risk of re-identification in the data access environment must be very low (no reasonable likelihood of re-identification).

To achieve this, data custodians need to balance data utility and the level of data modification required to achieve de-identification, while considering the most appropriate data access environment to facilitate this. For example, sharing data completely publicly or openly may necessitate significant alterations to the data itself, but lower data utility considerably. By contrast, a data access environment which enables a controlled release of data in a secure setting may provide greater protection, but it also places restrictions on who can access the data and how.

A note on terminology

Sometimes de-identification is used to refer to the removal of 'direct identifiers', such as name and address. In this guide, de-identification is used in a broader sense consistently with the meaning defined in the Privacy Act. It is therefore important to be aware that the removal of name, address or other direct identifiers alone may not result in de-identification for the purposes of the Privacy Act.

A number of different terms are used in Australia to describe processes similar to de-identification, for example 'anonymisation' and 'confidentialisation'. In particular, 'confidentialisation' is used by the Australian Bureau of Statistics (ABS).[5] Given the sometimes differing uses of terminology, it is important to check in any given scenario that the terminology being used is understood consistently by all parties.

[5] 'Confidentialisation' involves both the removal of direct identifiers, and then assessing and managing the risk of indirect identification occurring in the data. In practice, ensuring that information has been 'de-identified' for the purpose of s 6 of the Privacy Act will require entities to adopt a risk assessment approach similar to the one involved in confidentialisation. See, eg, the *Confidentiality Information Series* produced by the Australian Bureau of Statistics for more information.

Why de-identify?

This section briefly outlines the situations in which an entity may want or need to de-identify personal information, including under APP 11. There are a range of scenarios in which this might be necessary or desirable:

- when required by the APPs
- when an entity wants to share data or information, but this would not be permitted under the APPs
- for other risk-management purposes, or
- to build trust and meet community expectations around the handling of data.

It is important to remember that de-identification cannot eliminate all of the possible risks associated with a particular use of data. However, the OAIC encourages all entities to take a holistic approach to ensuring that their data use is reasonable and appropriate, compliant with the Privacy Act and considers ethical and social responsibilities.

When required by the APPs

The APPs require the de-identification of personal information in specified circumstances. For example, if an APP entity no longer needs personal information for any purpose for which it was collected, or any purpose for which it may be used or disclosed, the entity must take reasonable steps to destroy or de-identify the information (APP 11.2).

Other APPs that refer to de-identification are APP 4.3 (unsolicited personal information) and APP 6.4 (disclosure of personal information). See also APP 13.1 on correction of personal information. Further information about the APPs is available in the OAIC's *Australian Privacy Principle Guidelines*.

Note: The obligation to destroy or de-identify information that is no longer needed does not apply if the personal information is contained in a Commonwealth record.[6]

To enable the sharing or release of information

De-identification may allow an entity to share or release information in a way that would not otherwise be permitted under the APPs. For example:

- an agency holds sensitive information, which can only be used for a particular, specified purpose. However, another part of the agency wants to access that data and use it for an

[6] The obligation to destroy or de-identify does not apply to information that an entity is legally obliged to retain. Nor does it apply to personal information that is part of a Commonwealth record. This may be relevant both where the entity is an Australian Government agency, and where an entity has entered into a contract with an Australian Government agency. Generally, a Commonwealth record (as defined in s 6 of the Privacy Act) can only be destroyed or altered in accordance with s 24 of the *Archives Act 1983* (Cth) (Archives Act). The grounds on which this may be done include 'normal administrative practice' and destruction or alteration in accordance with an arrangement approved by the National Archives of Australia (often titled a 'Records Disposal Authority'). It is also an offence under s 26 of the Archives Act to alter a Commonwealth record that is over 15 years old. Advice about the Archives Act is available from the National Archives of Australia.

unrelated research or policy purpose. De-identification would allow that data to be shared for this secondary purpose.

- a medical research team creates a dataset about participants in a study. The dataset includes participant address information that could be used to identify or contact them. If the researchers want to share information from that dataset with researchers in another entity, the dataset should undergo a de-identification process prior to sharing.[7]
- a government agency wants to make data available for use by researchers and others outside that agency, to:
  - enable better public participation in government processes
  - inform policy and program development and design, or
  - drive innovation and economic growth by creating new opportunities for commercial enterprise.

The agency could achieve this by providing a de-identified version of the dataset via a secure access mechanism. This is discussed further below.

The process of de-identifying personal information can be considered a 'normal business practice' for the purposes of an entity's obligations under the Privacy Act,[8] and in particular APPs 3 and 6.[9]

Note: Australian Government agencies that undertake a de-identification process should be aware that they may need to retain an original, unaltered copy of the record, to comply with record-keeping obligations under other legislation.[10]

Other risk management purposes

There may be other circumstances in which de-identification is desirable to manage risk. For example:

- Where particular sections of an entity do not require access to all the information, but could make use of a de-identified version. For example, a real estate agency holds the personal

[7] In *Autism Aspergers Advocacy Australia and Department of Families, Housing, Community Services and Indigenous Affairs* [2012] AICmr 28, available on the OAIC's website, the Australian Privacy Commissioner found that de-identification can be used to protect an individual's privacy in response to a request under the *Freedom of Information Act 1982* (Cth). The Commissioner has also released the *Privacy Guidelines for the Medicare Benefits and Pharmaceutical Benefits Programs* under s 135AA of the *National Health Act 1953*, which establish when identifiable Medicare Benefits or Pharmaceutical Benefits claims information can be disclosed for the purposes of medical research.

[8] In this regard, de-identification can generally be considered incidental to the primary purpose of collection. This means that an entity's privacy policy does not need to refer to de-identification as one of the primary purposes of collection of the personal information – this is implied.

[9] Under APP 3, entities can only obtain personal information (whether from the data subject directly, or from any other entity) where this is reasonably necessary for, or directly related to, the entity's functions or activities. Further, under APP 6 entities should generally only use or share personal information for the same purpose(s) that it was collected for (the primary purpose), or where an exception applies - such as where the individual would reasonably expect the APP entity to use the information for that secondary purpose.

[10] See above, n 5. Before an agency de-identifies a record containing personal information it should consider its obligations under the Archives Act. An agency may be required to retain an original, unaltered copy of the record. This does not, however, prevent agencies from creating a de-identified version of the record to share or distribute in cases where privacy concerns would prevent disclosure of the original record.

information of clients, and wants to provide access to the database to agents in the firm so they can prepare a media release about recent property sales prices. The team requires access to information from the database, but not necessarily to information about the identity of individual clients. Consequently, the team could use a de-identified version of the database.

- Where de-identification would help protect information or data that an entity wants to keep confidential, such as sensitive business information.
- Where the entity wants to lessen the risk that personal information will be compromised when information is exposed to unauthorised access, use or distribution (or a data breach) — for example, where:
  - an intruder gains unauthorised access to information
  - an employee internally accesses information without authorisation, or
  - data or an information storage device is lost or stolen.[11]

Note: The Privacy Act also requires de-identification of certain types of information in some circumstances (namely tax file number information, credit information, and health information). These requirements are addressed briefly at the end of this guidance.

Build trust and meet community expectations

De-identification can also be valuable for promoting trust and meeting community expectations around data handling practices. Even if there is no specific legal requirement to do so, using appropriate de-identification techniques may alleviate public concerns about the handling of personal information. For example, the public may be concerned about potential privacy harms such as security risks, but also social or ethical harms such as discrimination, profiling or denial of a benefit or service.[12]

[11] The OAIC *Guide to information security* provides guidance on the reasonable steps entities are required to take under the Privacy Act to protect the personal information they hold.

[12] It is important to be aware that even where data has been appropriately de-identified, a risk of harm to individuals may still arise. For example, public transport data which contains information about trips that users have taken to a particular location (such as a medical treatment facility) could be hacked and made publicly available. The hacker could then incorrectly assume that a particular individual is in the dataset, leading to harm such as reputational damage or embarrassment, even though it was not possible to re-identify anyone from the relevant data. For these reasons, the OAIC encourages all entities to consider their ethical and social obligations as part of the broader management of their de-identification project, even where the materialisation of these risks would not necessarily result in a breach of the Privacy Act.

When is information 'de-identified' for the purposes of the Privacy Act?

Personal information is information which is 'about an identifiable individual, or an individual who is reasonably identifiable'.[13] 'De-identified' information is therefore information which has undergone a process of de-identification, and no longer falls within this definition.

When considering the meaning of de-identification, it may be useful to review the elements of the definition of personal information. For more information on this, please see the OAIC's *Guide to the meaning of personal information*.

When is a data subject 'reasonably identifiable'?

This is the key question that determines whether information is subject to the Privacy Act. There is no exact formula for assessing when information will be reasonably identifiable, and it can sometimes be difficult to draw a bright line between de-identified information and personal information. Whether information is about a 'reasonably' identifiable individual requires a contextual consideration of the particular circumstances of the case, including:

- the nature and amount of information
- who will hold and have access to the information
- the other information that is available to the person or people who will have access to the information, and
- the practicability of using that information to identify an individual.

Importantly, as part of considering the nature and amount of information, entities should consider what motivations there may be to attempt re-identification.

The inclusion of the term 'reasonably' in this phrase means that where it is possible to identify an individual, the next consideration is whether, objectively speaking, it is reasonable to expect that the subject of the information can or will be identified (having regard to the above factors). Even though it may be technically possible to identify an individual from information, if doing so is so impractical that there is almost no likelihood of it occurring, the information will not generally be regarded as personal information. There must be a reasonable likelihood that re-identification could occur for an individual to be reasonably identifiable.

Therefore, an individual will be reasonably identifiable (or, a de-identification process will *not* have been successful) where:

- it is technically possible for re-identification to occur (whether from the information itself, or in combination with other information that may be available in the data access environment), and
- there is a reasonable likelihood of re-identification occurring.

The Privacy Act does not require de-identification to remove the risk of re-identification entirely. Rather, those sharing or releasing data must mitigate the risk until it is very low. That is, until there

[13] This definition of 'de-identified' is given in s 6(1) of Privacy Act, and reflects the definition of 'personal definition' also contained in s 6(1).

is no reasonable likelihood of re-identification occurring. As part of this, the entity should consider all relevant risks that may impact on the likelihood of re-identification, including the risk of attribute disclosure,[14] and the risk of spontaneous recognition.[15] Entities should also consider the gravity of any harm that could arise from re-identification.[16]

What does a de-identification process involve?

De-identification is a process which involves the removal or alteration of personal identifiers, followed by the application of any additional techniques or controls required to remove, obscure, aggregate, alter and/or protect data in some way so that it is no longer about an identifiable (or reasonably identifiable) individual.[17]

In line with this, a de-identification process generally includes two steps. The first is the removal of direct identifiers, such as an individual's name, address or other directly identifying information. The second is taking one or both of the following additional steps:

- removing or altering other information that may allow an individual to be identified (for example, because of a rare characteristic of the individual or a combination of unique or remarkable characteristics that enable identification), AND/OR
- putting controls and safeguards in place in the data access environment, which will appropriately manage the risk of re-identification.

[14] While the risk of re-identification (or 'identity disclosure') is the principal consideration under the Privacy Act, other disclosures, for example 'attribute disclosure', where the user learns something about a data subject (without actually identifying them), must also be considered as part of this risk assessment. Attribute disclosure may not, in and of itself, constitute re-identification (and therefore, where attribute disclosure takes place, this does not necessarily mean that data was inadequately de-identified). However, where attribute disclosure might occur, this will likely have an impact on the risk of re-identification, and so entities must reduce the risk of both re-identification and attribute disclosure to an acceptably low level as part of the de-identification process. See the De-Identification Decision-Making Framework for further information.

[15] 'Spontaneous recognition' occurs where an individual is sufficiently unusual in a data collection, or the data user knows a sufficient number of an individual's attributes such that the user might be able to (unintentionally) identify the individual within the dataset. A common example might be recognising one's self in a data set, or recognising someone the user has a very close relationship with. This type of event can be difficult to prevent when working with rich or detailed datasets (and without making significant alterations to the data itself, which may reduce data utility). Therefore, it may be appropriate to manage any residual risk of spontaneous/unintentional re-identification through the use of environmental and/or user controls, which if appropriate and effective, would prevent re-identification occurring and therefore ensure that data is de-identified. See the De-Identification Decision-Making Framework for further information.

[16] Where re-identification could lead to serious harm, this may impact on the likelihood of re-identification (as people may be more motivated to invest time or resources into re-identifying sensitive data).

[17] For information on direct and indirect identifiers, see the De-Identification Decision-Making Framework.

Balancing data utility with the level of data modification required (and the choice of data access environment)

In determining the de-identification techniques that should be used to allow access to data, entities should consider all relevant contextual factors, including:

- the kind of information or data that is 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 characteristics (quasi-identifiers) that could enable re-identification
- whether the information or data will be targeted for re-identification because of who or what it relates to
- whether there is other information or data available that could be matched up or used to re-identify the de-identified data, and
- what harm may result if the information or data is re-identified.

Choosing appropriate techniques and an appropriate data access environment requires custodians to balance data utility and the level of data modification required to achieve de-identification, while considering the most appropriate data access environment which can facilitate this. There is sometimes an unavoidable trade-off here. Sharing data with a wider audience, (or indeed, completely openly), may require significant alterations to the data itself, which may significantly lower data utility.

In some cases the application of data modification techniques may reduce the data's utility. Nevertheless, this may be necessary in the choice of data access environment to minimise the risk of disclosing personal or confidential information. However, entities should note that the use of controls and safeguards in the data access environment, such as restrictions on who can access the information and specific physical/IT measures, , may allow the utility of data to be better preserved, compared with the use of data modification techniques.

Importantly, entities should be aware that if information is made publically available, control is effectively lost over that dataset. In future, new and more detailed data may become available that could be matched with this data leading to potential re-identification. Further technological advances will also be made, which could in turn increase the likelihood that information could be re-identified. Therefore, the future risk of re-identification must also be assessed as part of deciding on an appropriate mechanism for the sharing or releasing of data.

For these reasons, open data environments are generally only appropriate for information that is either not derived from personal information, or information that has been through an extremely robust de-identification process (inevitably focussed on data treatment) that ensures - with a very high degree of confidence - that no individuals are reasonably identifiable.[18]

[18] For policy guidance on the publication of open data, see also the Department of Prime Minister & Cabinet's *Process for publishing sensitive unit record level public data as open data* (December 2016): available at www.pmc.gov.au.

For further information on identification risk factors and methods for assessing identification risks, see the CSIRO Data61/OAIC resource The *De-Identification Decision-Making Framework*.

Choice of de-identification techniques

There is no one 'right' way to de-identify data. De-identification techniques should be carefully chosen, based on a risk assessment, to ensure that personal information is protected and that the information will still be useful for its intended purpose after the de-identification process.

As outlined above, removing or modifying personal identifiers such as a person's name and address is an essential component of de-identification. Other de-identification techniques that may be considered include further modifying the data itself, and/or applying appropriate controls and safeguards in the data access environment (note that there can be some overlap between the two).

The following examples are provided merely by way of illustration to demonstrate the two broad categories of techniques that may be used. Which specific techniques should be used, and in what combination, is a question that depends on the context of the data share or release, and may require advice from an expert (especially where data will be shared with a wide audience or publicly).

This guide is not intended to provide specialist advice on technical matters or de-identification standards, such as 'k-anonymity' or 'differential privacy'. The following techniques are given as illustrative examples only. For more information, see *the De-Identification Decision-Making Framework*.

Data modification or data reduction techniques

Examples of such techniques include:

- **Sampling** — providing access to only a fraction of the total existing records or data, thereby creating uncertainty that any particular person is even included in the dataset.
- **Choice of variables** — removing quasi-identifiers (for example, significant dates, profession, income) that are unique to an individual, or which in combination with other information are reasonably likely to identify an individual.
- **Rounding** — combining information or data that is likely to enable identification of an individual into categories. For example, age may be combined and expressed in ranges (25-35 years) rather than single years (27, 28). Extreme values above an upper limit or below a lower limit may 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 data is not significantly affected — a 'tolerable error' — but the original values cannot be known with certainty.
- **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.

- **Manufacturing synthetic data** — creating new values generated from original data so that the overall totals, values and patterns are preserved, but do not relate to any particular individual.[19] For example, a synthetic dataset could be used to test a program that detects fraud. The synthetic data could be built to enable a test environment that replicates patterns from an authentic dataset of normal use and fraud but does not allow individuals from the original data to be identified. This allows systems to be tested with data in a realistic way, but poses less risk of re-identification.
- **Encryption or 'hashing' of identifiers** — techniques that will obscure the original identifier, rather than remove it altogether, usually for the purposes of linking different datasets together (but without sharing the information in an identified form).

Applying controls and safeguards in the data access environment

These controls go to the 'who', 'what', 'where' and 'how' questions regarding access to data. These controls can be more effective at reducing the risk of re-identification than modifying the data itself, and better preserve the utility or richness of the data.

Examples of controls and safeguards which can help to manage and minimise the risk of re-identification include:

- Limiting access to information, for example, allowing other organisations or researchers to view the data rather than providing a copy.
- Allowing access via a secure mechanism or controlled environment, such as a data lab.
- Enabling an analysis of data rather than providing access to it, for example, running an analysis of the data and providing the result rather than the raw data.
- Requiring the data or information receiver to sign a contract limiting the use and distribution of the information or data, and enforcing the terms of that contract. For example, the contract could include an assurance that the receiver will not attempt to re-identify the information (and will destroy any information re-identified unintentionally, and notify the data owner, etc).

Further detailed advice and examples of different de-identification techniques are discussed in the *De-Identification Decision-Making Framework*.[20]

[19] United Kingdom Information Commissioner's Office, *Anonymisation: managing data protection risk code of practice*, published 2012, United Kingdom Information Commissioner's Office, Wilmslow, Appendix 2, p 53, available on the ICO's website.

[20] See also the *Confidentiality Information Series* produced by the Australian Bureau of Statistics.

Confirming that de-identification techniques are appropriate

Before sharing or releasing de-identified information, you should confirm whether the de-identification techniques chosen are appropriate to manage the risk of re-identification.

Entities may need to engage an expert to undertake a statistical or scientific assessment of the information to ensure the risk of re-identification is very low in the data access environment. There are a number of ways to assess disclosure risk, including Data Analytical Risk Assessment and Penetration testing.

Regardless of the method chosen to assess disclosure risk, whether the risk of re-identification has been reduced to an acceptably low level is a highly contextual question. Relevant factors to consider when determining whether information has been de-identified effectively could include the difficulty, cost, practicality and likelihood of re-identification. The *De-Identification Decision-Making Framework* sets out a framework for assessing and managing the risk of re-identification with reference to a specific data environment.

As a minimum, an entity must apply the 'motivated intruder' test - this test considers whether a reasonably competent, motivated person with no specialist skills (who would have access to the data) would be able to identify the data or information (the specific motivation of the intruder is not relevant). It assumes that the motivated intruder would have access to resources such as the internet and all public documents, and would make reasonable enquiries to gain more information.[21]

Depending on the context, in addition it may be appropriate to consider re-identification 'in the round'. That is, assessing whether any entity or member of the public (including experts and people with specific knowledge or a high level of skills) could identify any individual from the data or information being disclosed, either on its own or in combination with other available information or data.[22] Conducting this sort of assessment will require specialist expertise, and if an entity does not have this capability in-house, external assistance should be sought.

As publically released data can be accessed by anyone in the world (including experts and people with specific knowledge or a high level of skill), entities must consider the risk of re-identification 'in the round' when considering whether data has been de-identified effectively for public release.

Considerations *after* data sharing or release has occurred

De-identified information may still carry some risk of re-identification, particularly where it could be matched with other information which becomes available in future.[23] The risk of re-

[21] United Kingdom Information Commissioner's Office, above n 15, p 22.

[22] Ibid, p 19.

[23] See, eg, Anna Cavoukian and Khaled El Emam, *Dispelling the myths surrounding de-identification: anonymisation remains a strong tool for protecting privacy*, (June 2011), available for free download on the Information and Privacy Commissioner of Ontario's website, p 4; and Ira S Rubinstein and Woodrow Hartzog, *Anonymization and Risk* (2015). New York University Public Law and Legal Theory Working Papers. Paper 530.

identification may therefore shift as technologies develop and a greater amount of information is published or obtained by an entity.

As a result, entities should regularly re-assess the risk of re-identification and, if necessary, take further steps to minimise the risk. This may include:

- assessing whether a higher level of de-identification is required for information, and
- assessing whether the release of further de-identified information (or technological changes) could potentially facilitate re-identification of already-published information.

As changes could occur in future (regarding technology or data availability), this is one of the main reasons why open data environments are generally only appropriate for information that is either not derived from personal information, or data that has been through an extremely robust de-identification process.

Remember: a ‘de-identified’ status may be context-dependent

It is important to remember that de-identification is not a fixed or end-state. The same information may be personal information in one situation, but de-identified information in another. To use a very simple example to illustrate this point, a person’s licence number (in and of itself, and in the absence of any other information) may not be personal information when in the hands of a random member of the public, for example a school teacher. However, the same number is likely to be personal information if held by an employee of the relevant motor vehicle registry, or any other person who has access to a relevant database and can look up the number and see other identifying details associated with the licence number.

The same is true for data which has undergone a de-identification process. Imagine an entity has undertaken a de-identification process, but they still retain a copy of the original dataset. This may enable them to re-identify the data subjects in the de-identified dataset if they wished to do so. So, the dataset may be personal information when handled by that custodian, but may be de-identified when handled by a different entity, since the data access environment is different.

Further, information may be de-identified for some parts of an entity, but remain personal information for others. Imagine that an entity decides to undertake a de-identification process on a dataset, to enable in-house research to be conducted on that data (in ways that may not otherwise be permitted if the data was subject to the Privacy Act). As for the above scenario, the custodian undertaking the de-identification will likely retain a copy of the original dataset which would enable them to re-identify the data subjects in the de-identified dataset if they wished to do so. If the custodian wants to ensure that this particular use of the dataset is de-identified and therefore outside the scope of the Privacy Act, additional techniques or controls would need to be put in place (see discussion above).

Do privacy obligations still apply to de-identified data?

The above discussion demonstrates that de-identified information can still carry certain privacy risks, if it is still possible for re-identification to occur. For example, such information may still carry privacy risks where the de-identification techniques used involved the use of controls applied in the data access environment (which, if removed, could increase the likelihood of re-identification). This can sometimes be difficult to conceptualise, so an example is helpful.

Imagine that an entity has de-identified information for internal use, in part through the application of controls and safeguards in the data access environment. However, if that information is then subject to a data breach, for example through inadvertent publication on the entity's website, in this open access data environment the information could become personal information again (for example, because it could be matched with other data sources). The APPs would then apply to this information, and this event could therefore amount to a breach of APP 11 (depending on the entity's security practices). By way of another example, if the entity disclosed the information to another entity without properly de-identifying it for that data access environment, this could amount to a breach of APP 6 (and/or APP 8, if the disclosure was to an overseas entity).

How should an entity manage such a risk, while still using de-identified data in the ways that it is legally permitted to use it? The OAIC recommends that entities take a risk-management approach when handling de-identified data which acknowledges that while the APPs may not apply to data that is de-identified in one specific context, the same data could become personal information in a different context.

Some APPs therefore continue to be relevant to the handling of de-identified information, specifically, APPs 6, 8 and 11, as these are the APPs which may apply if the data was to be transferred to another environment. These APPs therefore remain relevant to the handling of de-identified information - not as a matter of law (at that time), but as a matter of risk management (because if the information was to be transferred to another environment, re-identification may become possible and the APPs may therefore apply again).

Entities should therefore handle de-identified data in a way that would prevent any breaches of the APPs occurring. While this requires some extra thought and planning, it should not prevent the entity from using de-identified information to achieve its organisational goals.

APP 6 — Use and disclosure

APP 6 states that personal information can be used or disclosed (shared/released) only for the same purpose for which it was obtained, unless an exception applies.

APP 6 will not generally be relevant in relation to uses of de-identified information, provided that the de-identified status of the data doesn't change in the context where the use takes place. This will usually be the case where the information is confined within the context of the original entity. After all, the whole point of the de-identification process may have been to enable different uses of the data in-house.

However, APP 6 will remain relevant in relation to the sharing or release of information, if the status of the information would change to being personal information in the hands of another entity.[24]

[24] Entities will therefore need to consider whether the information would be identifiable or reasonably identifiable in the hands of the other entity. Otherwise, the APPs will apply to the sharing/release of that data.

APP 8 — Overseas transfers

APP 8 prevents the sharing or release of personal information with an entity outside Australia unless certain steps have first been taken.

APP 8 may remain relevant to de-identified data, because it applies to situations where data is shared or released to an overseas entity (if the de-identified status of the information may change when released out of the entity's control).

APP 11 — Security

APP 11 states that entities must take reasonable steps to protect personal information from 'misuse, interference and loss', and 'unauthorised access, modification or disclosure'. APP 11 is probably the most important APP to keep in mind when handling de-identified data, and the data breach example given above illustrates this point. As for all valuable data, de-identified data should, as a matter of risk management, be protected and stored securely to prevent unauthorised access. This is particularly so in light of the obligations which now apply to entities as a result of the commencement of the *Privacy Amendment (Notifiable Data Breaches) Act 2016* (Cth).[25]

Entities need robust de-identification governance processes in place

The above information has highlighted that de-identification can be complex, with a range of factors to consider at each point in the process. To ensure effective de-identification practices, entities should be mindful of their general responsibilities under APP 1.2.[26]

APP 1.2 states that entities must take reasonable steps to implement practices, procedure and systems that will ensure compliance with the APPs. Having a strong data governance framework in place will help to ensure that de-identification is carried out effectively and in particular, will help to ensure that no breaches of APP 6 or APP 11 will occur when undertaking de-identification (or through the subsequent use of data that has undergone such a process).

Good data governance procedures should apply at all stages of the process. Procedures should be in place for both before and after de-identification has occurred, as well as during. Even after a de-identified data set has been shared or released, as outlined above the entity's responsibilities in relation to that data may continue in some circumstances.

Robust de-identification governance practices may include activities such as:

- ongoing and regular re-identification risk assessments (to check that methods used are still effective and appropriate at managing the risks involved)
- auditing data recipients to ensure that they are complying with the conditions of any data sharing agreements

[25] For more information on the obligations that apply, see the OAIC's Notifiable Data Breaches Scheme web page.

[26] APP entities that are Australian Government agencies also have specific obligations under APP 1.2 due to the *Privacy (Australian Government Agencies — Governance) APP Code 2017*.

- considering new information that becomes available, and whether any such information increases re-identification risk in a particular data access environment
- building and ensuring ongoing transparency around the entity's de-identification practices
- ensuring that your entity knows who is accountable for de-identification internally
- putting in place a plan to deal with any re-identification attacks or events (this could be part of your entity's broader data breach response plan)
- ensuring that in-house staff who undertake de-identification have adequate and up-to-date training, and/or ensuring appropriate external expertise is sought where appropriate.

De-identification in specific contexts

Tax file numbers

Entities should also be aware of the obligation imposed by the *Privacy (Tax File Number) Rule 2015* issued under s 17 of the Privacy Act. The Guidelines require an entity to take reasonable steps to securely destroy or permanently de-identify tax file number information that is no longer required by law to be retained, or is no longer necessary for a purpose under taxation law, personal assistance law or superannuation law.[27]

Credit Related Research Rule

De-identification obligations for credit reporting bodies and credit providers also apply in relation to credit-related personal information.[28] Credit reporting bodies must also comply with s 20M of the Privacy Act, which prevents the use and disclosure of de-identified credit reporting information except when that use and disclosure is for the purpose of conducting research in accordance with the *Privacy (Credit Related Research) Rule 2014*.

Health information for research for public health or safety purposes

Entities cannot collect health information about individuals for one of the research or public health or safety purposes permitted under s 16B(2) of the Privacy Act if de-identified information would serve the same purpose (s 16B(2)(b)). If de-identified information would not serve the same purpose (and if other conditions imposed in s 16B(2) have been met) the entity can only collect the information in accordance with guidelines approved by the Information Commissioner under s 95A about use of health information for research or public health or safety purposes.

[27] See the OAIC's *Privacy agency resource 5: The Privacy (Tax File Number) Rule 2015 and the protection of tax file number information*.

[28] See the OAIC's website for more information on the credit-related research rule, and for more information on the operation of Part IIIA of the Privacy Act more generally, which regulates credit reporting in Australia.

Quick de-identification checklist

Is your information de-identified for the purposes of the Privacy Act?

Step 1:

□ Have you removed or otherwise altered 'direct identifiers' — such as names and addresses?

Step 2:

□ Have you done either or both of the following as required?

- Removed or otherwise altered other information ('quasi-identifiers') that may lead to re-identification in the data access environment? AND/OR
- Put in place any environmental controls that may be necessary to manage the risk of re-identification in the data access environment?

Note: Whether information is personal or de-identified will depend on the context. Information will be de-identified where the risk of an individual being re-identified in the data is very low in the relevant release context (or data access environment).[29] Put another way, information will be de-identified where there is no reasonable likelihood of re-identification occurring.

Further resources

More information about an entity's obligations under the Privacy Act is available at www.oaic.gov.au.

The OAIC and Data61 joint resource on de-identification, *The De-Identification Decision-Making Framework* is available for download from the CSIRO Data61 website.

The OAIC *Guide to Data Analytics and the Australian Privacy Principles* provides guidance about how the Privacy Act applies to data analytics activities and how data analytics can be conducted while protecting personal information.

[29] The data access environment means the context that the data will be made available in. The data access environment is made up of four key components: other data, people, infrastructure, and governance structures. For further information see the De-identification Decision Making Framework.