De-identification Guideline

Issued by the Chief Data Officer on 21 February 2018 under section 33 of the *Victorian Data Sharing Act 2017*

VICTORIA State Government
Premier and Cabinet

Contents

1. About this guideline

Section 18 of the *Victorian Data Sharing Act 2017* (the Act) requires the Chief Data Officer (CDO) and data analytics bodies to take reasonable steps to ensure that data received from data sharing bodies and designated bodies under the Act no longer relates to an identifiable individual or an individual who can be reasonably identified, before using that data for the purposes of data analytics work.

This guideline is issued under section 33 of the Act. It sets out the techniques and considerations the CDO and data analytics bodies must have regard to in determining what are ‘reasonable steps’ under section 18. It is also relevant to the requirement to ensure that the results of data analytics work no longer relates to an identifiable individual or an individual who can be reasonably identified under section 19.

The guideline applies only to the CDO and data analytics bodies which **receive** identifiable data under the Act. It does **not** apply to data sharing bodies or designated bodies which provide identifiable data. If you are unsure if your organisation is a data analytics body, data sharing body or designed body under the Act, please read the Guidance for Departments and Agencies.

2. About the *Victorian Data Sharing Act 2017*

2.1 Overview of the Act

The Victorian Government collects a lot of data when serving the community. The *Victorian Data Sharing Act* (the Act) promotes data sharing across government by:

- creating a clear framework for sharing and using data for policy making, service planning and design
- establishing the Chief Data Officer (CDO) who leads the Victorian Centre for Data Insights (VCDI) in working to transform how government uses data.

2.2 Requirements around the handling of identifiable data

The Act provides a range of protections and safeguards to ensure that data is used in the right way, including:

- requiring that all data must only be used for informing policy making, service planning and design (section 5)
- annual reporting and notification to privacy regulators (section 24 and Part 6), and
- new offences for unauthorised access, use or disclosure (Part 5).

The Act also has specific requirements around the handling of identifiable data received by the CDO and data analytics bodies, which align with 3 key phases of the analytics lifecycle:

1. After receiving identifiable data under the Act, they must only use the data for the purpose of data integration (section 17)
2. After integrating the data, they must take reasonable steps to ensure that the data no longer relates to an identifiable individual or an individual who can be reasonably identified, before performing data analytics work (section 18), and
3. Before disclosing the results of data analytics, they must ensure that the results no longer relate to an identifiable individual or an individual who can be reasonably identified (section 19).

This is shown in Figure 1 as follows:

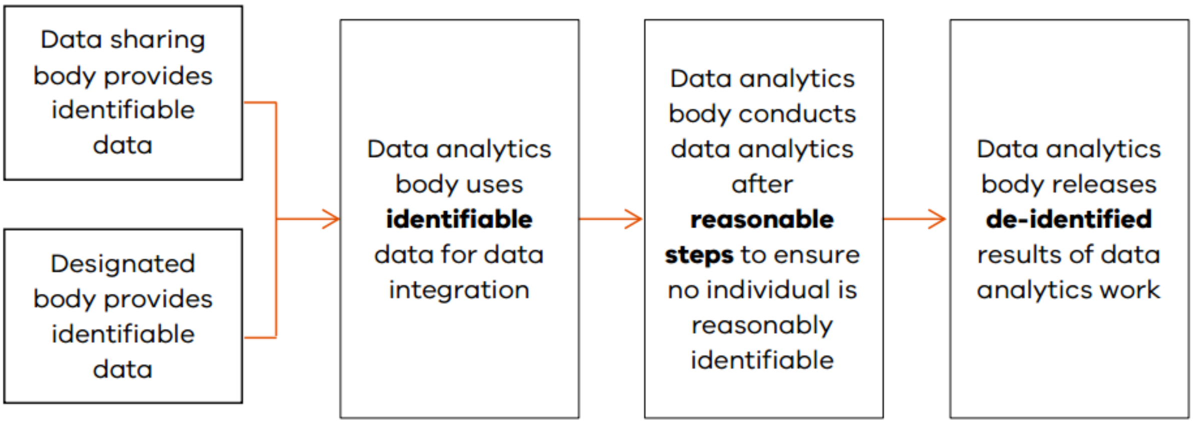
Data sharing body provides identifiable data
Designated body provides identifiable data
Data analytics body uses **identifiable** data for data integration
Data analytics body conducts data analytics after **reasonable steps** to ensure no individual is reasonably identifiable
Data analytics body releases **de-identified** results of data analytics work

Figure 1: Data handling requirements under the Act

To help decide whether reasonable steps have been taken, section 18(2) requires the CDO and data analytics bodies to have regard to:

- the de-identification techniques applied to treat the data
- the technical and administrative safeguards and protections implemented in the data analytics environment to protect the privacy of individuals, and
- any other considerations specified in guidelines issued by the CDO.

This guideline sets out both the de-identification techniques, as well as the technical and administrative safeguards and other considerations to have regard to under section 18 of the Act. It is also relevant to understanding the obligations under section 19.

For more information on how the Act applies to data analytics bodies, read the Guidance for Departments and Agencies.

3. De-identification in context - within the data analytics lifecycle

3.1 Overview of data analytics lifecycle

The high level phases in a data analytics lifecycle include the following:

- Data Collection
- Data Ingestion/ETL/Cleansing
- Data Integration/Linkage
- Data Analytics/Modelling
- Data Release/Publish/Visualisation.1

This section highlights the key phases of the data analytics lifecycle where some form of 'de-identification' is required under the Act, specifically:

- Data Integration phase: where identifiable data can be used (section 17)
- Data Analytics phase: where reasonable steps have to be taken to 'de-identify' the data before data analytics work can take place (section 18), and
- Data Release phase: where the results of data analytics must be 'de-identified' before being disclosed (section 19).

The flow of data through the analytics lifecycle as it aligns with the requirements under the Act can be shown in Figure 2 as follows:

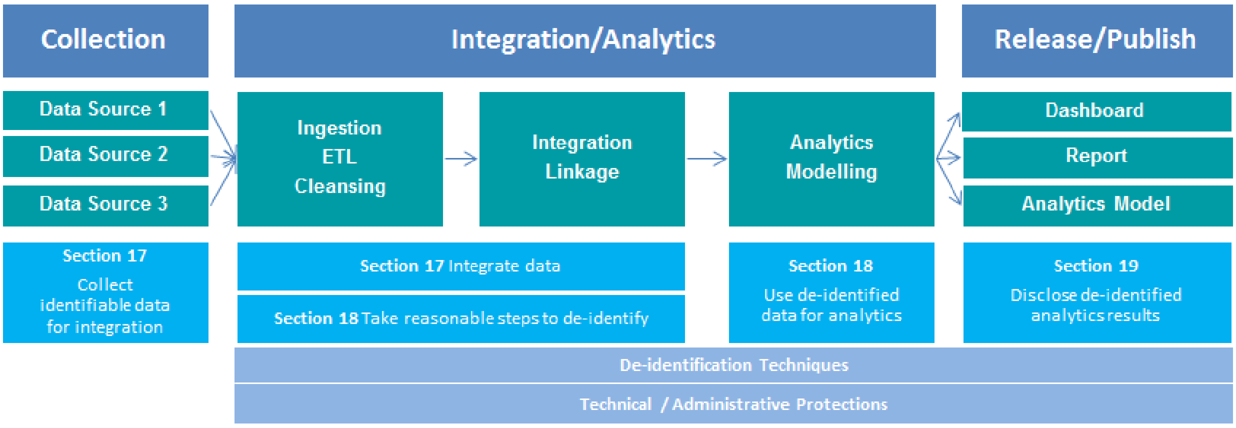
Collection
Integration/Analytics
Release/Publish
Data Source 1
Data Source 2
Data Source 3
Ingestion
ETL
Cleansing
Integration
Linkage
Analytics
Modelling
Dashboard
Report
Analytics Model
Section 17
Collect identifiable data for integration
Section 17 Integrate data
Section 18 Take reasonable steps to de-identify
Section 18
Use de-identified data for analytics
Section 19
Disclose de-identified analytics results
De-identification Techniques
Technical / Administrative Protections

Figure 2: Data Analytics Lifecycle as aligned with the Act

[1] **Note**: These phases do not represent the full data lifecycle as represented in various other data lifecycle models, which typically also include data retention, archival and destruction phases that are not relevant for this guideline.

3.2 Key issues at each phase

Key issues typically considered in each phase of the data analytics lifecycle include:

3.2.1 Collection

- Identifiable data must only be received and used for the purpose of data integration under the Act.
- The scope of identifiable data received must be restricted by the collection minimisation principle, where only data that is necessary to achieve the agreed analytics project purpose is collected.
- Identifiable data for collection must be encrypted and password protected.
- Identifiable data must only be collected by an authorised member of a data integration team (e.g. a data engineering role) who has the required security vetting, technical expertise, and relevant data privacy and security training.
- Data content and identifiers must be collected separately from the data provider. Data linkage is enabled via unique client ID's contained in both Content and Identifiers.

3.2.2 Ingestion / ETL / Cleansing

- An authorised member of a data integration team ingests the raw data from the Collection phase and performs the required Extract/Transform/Load, normalisation, cleansing etc. including creation of linkage keys, and removal of identifiers.
- Raw data is transformed into internal and external database schema that will be accessed by a separate data analytics team (view/read only).

3.2.3 Integration / Linkage

- The Act requires that identifiable data can only be used to conduct data integration.
- Separate data linkage and integration teams perform the required linking and integration of 'de-identified' data using the linkage keys created in the Ingestion/ETL/Cleansing phase.

3.2.4 Analytics / Modelling

- The Act requires a data integration team to take reasonable steps to ensure that the data no longer relates to an identifiable individual or an individual who can be reasonably identified before providing the integrated, 'de-identified' data to a data analytics team for analytics and modelling.
- A data analytics team has view/read access only to the data.
- The process of iterative analytics will necessitate ongoing re-assessment of re-identification risks as either:
    - further 'de-identified' data is integrated with the existing analytics data set, or
    - further analysts are provided with access to the analytics environment.
- Ongoing assessment of re-identification risk must be guided in the above scenarios by project team members that have intimate knowledge of and insight into the project purpose, data, analysts, and access environment.
- Finding a re-identification risk must lead to implementation of further controls and protections to mitigate the risk, e.g. by applying a different de-identification technique, storing and accessing the data

in a more hardened environment, aggregating data to a higher level cohort, restricting access to a reduced number of analysts, or to analysts with a higher security vetting.

3.2.5 Release / Publish

- The results of analytics work must only contain 'de-identified data' under the Act.
- Results must be aggregated and 'de-identified' before being shared or released to users that have been authorised within the project governance model.
- The level of data treatment required at the results stage will be shaped by an assessment of the sensitivity of the data, the users who will get access, and how the analytics outputs align with the overall project purpose.

4. De-identification techniques and technologies

4.1 Meaning of 'de-identification'

'De-identification' involves removing or altering data that identifies an individual or is reasonably likely to do so. In most cases the de-identification process involves:

1. removing direct identifiers (such as name, address, or Medicare number), and
2. removing or altering the data to protect indirect or quasi-identifiers (such as data of birth, gender, profession, rare condition etc. that might permit identification when used in combination with other data).

4.2 Overview of techniques and technologies

There are many different techniques and technologies that can enable privacy preserving analytics appropriate for different user groups.

These techniques and technologies can be grouped by function as follows:

1. **Masking personal identifiers**: remove or replace fields that may identify individuals, such as names, addresses, telephone numbers etc.
2. **Generalisation and grouping**: group data at a granularity that obscures unit records; including aggregation of data as well as more advanced techniques.
3. **Perturbation**: change raw data or results to protect unit records, e.g. the 'TableBuilder' product from the Australian Bureau of Statistics (ABS) uses this technique to protect remote tabular queries to census data. A very significant new group of privacy protection methods that satisfy differential privacy largely fit within this topic.
4. **Synthetic data generation**: generate a new data set of unit records that is "similar" to raw data, e.g. the US Census Bureau uses this technique to allow analytics over the Longitudinal Business Data (LBD).
5. **Encrypted computation**: data is encrypted using a scheme that enables accurate analytics to be performed on it, while never revealing the encrypted raw data. As the results of the computation may be disclosive, it may be necessary to apply other protection mechanisms to the results. This is a new development and is rapidly evolving.

Two additional technology groups are relevant in maximising the protection of sharing identifiable data:

- **Privacy preserving record linkage**: link data from two different sources without disclosing the personal information used to do the linkage to any external organisation.
- **Re-identification risk assessment**: techniques to determine how much residual risk of re-identification there is in a data set after it has been protected by a privacy protecting method.

No categorisation into groups for such a complicated domain is complete or unambiguous. However, these functional distinctions are useful for the design of a data analytics system based on clear privacy principles.

4.2.1 Masking personal identifiers

The simplest method to protect privacy is to remove fields that are considered personally identifying. Masking may involve suppressing entire fields or just at-risk data values.

Suppression of particular data values will in general leave holes in the data that can be difficult to analyse properly. In the US, the Health Insurance Portability and Accountability Act specifies 18 types of variables as indicative of the presence of identifying information and considers a data set ‘de-identified’ if they are not present. This standard is increasingly becoming the minimum requested standard worldwide for management of personal medical information in both clinical and research settings.

The residual re-identification risks associated with masking personal identifiers relate to the fact that the masked data can still be identifying, particularly in combination with other data sets. Data that may not be personally identifying in the context of a particular database can become so when joined with data in another data set.

It is therefore imperative that further assessment of re-identification risks occurs, and that additional controls are considered when other data sets are joined.

Pseudonymisation

Pseudonymisation is a kind of masking where personal information is replaced with a pseudonym, a specially crafted value that can be used to identify unit records, but which does not itself contain personal information.

A pseudonym can act as a key to link unit records between data sets and, if the mapping between direct identifiers and pseudonyms is preserved or can be reversed, the data can be optionally re-identified later in a controlled environment. Thus, in principle, a pseudonym provides some privacy while also permitting the linkage of different data sets that use the same pseudonymisation function, something which is impossible after masking by suppression.

Pseudonomisation techniques should be carefully considered and implemented, as many commonly used techniques only pseudonymise directly identifying information, while leaving quasi-identifiers in raw form.

As an example, the use of Statistical Linkage Keys doesn’t always ensure that personal information is hidden and often introduces ambiguities in matching due to their poor uniqueness properties. More advanced pseudonymisation methods such as “Cryptographic Long-term Keys” also have limitations.

Even when the pseudonymisation process is itself robust, the use of the same pseudonym across multiple data sets and the availability of quasi-identifiers could leave pseudonymised data at risk of re-identification through linkage.

4.2.2 Generalisation and grouping

Data generalisation and grouping (referred to commonly as ‘grouping’) occur when data is expressed in summary form by grouping related values into categories or ranges. This can reduce disclosure risks by removing unit level identifiers and turning atypical records, which generally are most at risk, into typical records. Essentially, grouping trades accuracy or ‘resolution’ for privacy, since any analysis on the grouped data cannot be more specific than what the grouping permits.

For grouping to work effectively, the groups must be defined by someone with relevant domain knowledge, which can be a significant expense.

Grouping can suffer from some of the same re-identification risks such as masking, when joining several data sets results in the possibility of re-identifying data that is not re-identifiable in isolation. Grouping does mitigate this problem to a certain extent by necessitating more joined data before re-identification becomes possible than plain masking does.

k-anonymity

A release of data is said to have the k-anonymity property if the information for each person contained is common with at least k-1 individuals whose information also appears in the release. There are two methods of reducing the granularity of a data set so that it satisfies the k-anonymity property:

- **Suppression**: where sensitive values are removed or replaced with ‘placeholder’ symbols e.g. replacing name and religion values with a star ‘*’, and
- **Generalisation**: where individual attribute values are replaced with a broader category e.g. replacing precise ages with one of a fixed set of age ranges, age 25 becoming ‘between 20 and 30’.

l-diversity

l-diversity was proposed as an extension to k-anonymity to address the limitation where protecting identities to the level of k-individuals is not equivalent to protecting the corresponding sensitive values that were generalised or suppressed, especially when the sensitive values within a group exhibit homogeneity. The l-diversity model promotes intra-group diversity for sensitive values in the anonymisation mechanism.

4.2.3 Perturbation

Perturbative methods for disclosure control usually involve swapping or permuting values, or adding random amounts of noise. Swapping data values for selected records can discourage attackers from matching, since matches may be based on incorrect data. For example, one might switch the values of age, race, and sex for at-risk records with those for other records.

Numerical data can be protected by adding some randomly selected amount of noise (e.g. a random draw from a normal distribution with mean equal to zero). Adding noise to values can reduce the possibilities of accurate matching on the perturbed data, and distort the values of sensitive variables.

Both of these perturbative methods are used extensively by government agencies. Other perturbation methods include rounding values, resampling, micro-aggregation, post-randomisation, and macro-agglomeration.

There is still a re-identification risk associated with inadequately perturbed data sets e.g. Netflix published a ‘de-identified’ data set of users’ movie ratings, which was subsequently re-identified by matching the data with Internet Movie Database movie ratings.

Differential privacy

The current state of the art standard for controlling re-identification risk is called differential privacy. It can greatly reduce the disclosure risk issues in k-anonymity, l-diversity and their extensions. An algorithm is said to be differentially private if the operation of that algorithm on the data set produces ‘essentially the same’ answers regardless of the presence or absence of any particular unit record.

The main drawback of using differentially private algorithms for data analysis is that, like the other perturbation methods, it does not preserve the accuracy of the algorithms applied to the original unit records, and in fact will deviate more from the results based on the original unit record data in proportion to the very privacy guarantees that it provides.

4.2.4 Synthetic data generation

Synthetic data replaces original data values at high risk of disclosure with values simulated from probability distributions. These distributions are designed to reproduce as many of the relationships in the original data as possible. There are two approaches to creating synthetic data:

- **Partially synthetic data** comprises the units originally surveyed with some subset of collected values replaced with simulated values, and
- **Fully synthetic data** involves replacing an entire data set (rather than just a subset) with synthesised replacement data.

There are a variety of methods for generating synthesised data, each representing a trade-off between utility and re-identification risk. Methods for generating synthesised data can generally be classed according to what kinds of statistical properties are preserved from the original data, and these properties determine which analyses provide reliable results and which do not.

4.2.5 Encrypted computation

The significant limitation of the techniques above is either that they present a real risk of re-identification, or they significantly reduce the value and insights that could be derived from the data being analysed. New federated privacy-preserving analytics methods are being developed that combine distributed machine learning with homomorphic encryption or secure multiparty computation to provide the ability to perform machine learning across multiple data sets without any of the data leaving its secure source i.e. without the need for organisations to disclose any unit records in their raw form.

The primary drawback is that the source data is hidden from the analyst, which can be frustrating for analysts who are used to interacting directly with data. The second is in the sophistication of the systems, which are usually based on advanced mathematics. This has the dual consequences that the technology involved can be difficult to explain and justify to non-experts, and the implementations usually require much more computational resources than the equivalent analyses on ‘de-identified’ data discussed above.

Re-identification risks are also different to the techniques above. Since the computation performed has all the accuracy that it would if it were performed on the raw data, when the final result of the analysis is decrypted it can in certain circumstances contain traces of some of the data that was used to produce it. This is an active area of research and examples of such ‘information leakage’ are difficult to find in practice, but in principle this is something that needs to be considered.

5. Other technical and administrative safeguards and protections

Re-identification risks cannot be mitigated by de-identification techniques and technologies alone, as no de-identification technique or technology that exists today can be considered foolproof, and all 'de-identified' data carries some residual re-identification risk.

Effective mitigation of re-identification risk requires implementation of a wider set of administrative controls and technical protections on top of the techniques outlined in Section 4 above.

The most effective combination of controls and protections must be determined by balancing the intended uses and benefits of the data against the potential re-identification risks.

5.1 Key safeguards

De-identification techniques must be implemented in combination with other risk mitigation tools and techniques, including:

5.1.1 Collection Minimisation

- Collection of data for integration/analytics projects must only include data that is necessary to achieve the agreed project purpose.

5.1.2 Data Risk Indicators

- Ingestion of data that has been collected for integration/analytics must include a set of filters to identify 'red flag' data types, e.g. tax file numbers, biometric information, credit card numbers etc.

5.1.3 Data Separation Protocols

- Data Separation Protocols must be implemented to ensure separation between data integration and analytics teams, and separation within the integration team between linkage and content integration teams (including physical and logical separation).

- Access to identifiable data in the integration environment must be restricted to an authorised data integration role who will perform data integration and apply the appropriate de-identification techniques and technologies.

- Data analytics roles must not have access to any identifiable data and will only get access to 'de-identified' data in the analytics environment.

- Approved authorised external users may obtain access to aggregated data elements as required for a specific project. Access must be granted within a controlled environment with restricted user credentials.

5.1.4 Access Controls

- The baseline for data access is that access must only be granted to authorised users within a secure environment.

- Data access must be provided on a project by project basis, based on agreed and authorised project roles within a Role Based Access Control framework, and be aligned with the required levels of 'de-identification' for each phase of the data analytics lifecycle.

- Data access must be managed in accordance with an Identity Lifecycle Framework to ensure that user access is de-provisioned at the end of a project, a change in role, or a user leaving the organisation.

5.1.5 Access Mechanisms

- Access to data can be granted through many different mechanisms, based on the following access types:

    - **Controlled direct access** – the analyst can see the fields of the data set for the purpose of analysis, but all interactions with the data are monitored and recorded. Any outputs from the data are reviewed or checked to determine if they reveal source data. Examples of this are secure virtual data analysis environments, such as Secure Unified Research Environment.

    - **Mediated access** – the analyst can generate analytical results, but cannot see the actual fields within the data set (i.e. the raw data set). The generation of count tables from the Australian Census by the 'TableBuilder' facility (overseen by the ABS) is an example of one of the many ways of enabling this form of access.

    - **Open access** – release of the data for use in an analyst's own environment where the analyst may perform any data analysis or manipulation without constraint. Open data (released to the public) is one common form of open access, but may also include data released within a single organisation for use.

- More fine-grained distinctions enforced by access controls and regulatory mechanisms are necessary in practice (e.g. systems administrators, etc.), but for explanatory purposes, these three groups are sufficient.

- More recent technologies involving analytics over encrypted data allow mediated access to a federated data source. In this case, as the data is encrypted, it is possible to make stronger guarantees about what is disclosed and to whom during analysis, and it is possible to generate analytics results without disclosing anything but the result of the analysis to any other group.

- Where data from multiple sources is used for analytics, a unified repository of the data is not constructed, and no data is exposed between the sources. These techniques are rapidly being developed and exploited in the intelligence community where secrecy is paramount, and also commercially in a variety of ICT companies where business confidentiality is a strong requirement. These approaches can be considered for the most sensitive types of data analysis.

5.1.6 Auditing

- Data access must be monitored and audited to ensure compliance with the RBAC and Identity Lifecycle Frameworks.

- Audit logs must be kept to show when data was received and who has access.

- Data analytics team coding, modelling, and all incremental changes must be captured in audit logs for traceability.

6. Decision on which de-identification techniques and other controls to apply

In deciding which is the most appropriate de-identification technique, or other technical or administrative safeguard to apply to data to meet the requirements to 'de-identify' under sections 18 or 19 of the Act, the CDO and data analytics bodies must have regard to matters including the following:

6.1 Assessing the level of re-identification risk

Contextual evaluation of re-identification risk must include consideration of:

- the project purpose
- disclosure risks in the data itself
- the people who will access the data to conduct integration or analytics
- how data sets are joined
- sensitivity levels of data sets
- the data access environment (including physical, technical, and procedural access controls)
- how disclosive the analytics results are
- the people who the data analytics results will be shared with or published to, and
- levels of trust and control that can be placed on data users.

A key part of the evaluation process for integration/analytics projects includes collaboration with subject matter experts in project partners (e.g. data custodians, information managers) to leverage their insights, particularly around:

- data quality
- metadata
- data sensitivity, privacy, secrecy and confidentiality
- data context
- data lineage, and
- data ownership.

A number of these considerations are discussed in more detail below.

6.1.1 Data linkage and integration

Typically, data from different sources must be linked or integrated using a common reference. There are three major types of common reference used, with different levels of associated re-identification risks:

- **Person**: records from each source pertain to an overlapping set of people, linked together by comparing the identifying information held about that person by each source. As this process generally involves the use of personal identifiers and carries the highest level of privacy sensitivity, particular care must be taken to secure the information and ensure that it is compliant with privacy laws and regulations that apply.

- **Location**: when data has been aggregated at a postcode level, for instance, it is easy to join that data to other data aggregated the same way. This typically has little privacy sensitivity, as sufficiently aggregated data is impossible to re-identify. When the level of detail of the spatial data increases, it can cause privacy concerns due to the ability to link individuals to specific locations, which makes it ‘personal information’.
- **Business**: data from different sources can be joined using a common identifier such as ABN. This typically does not cause privacy issues (except in the case of sole traders), however it does raise concerns about confidentiality, where a company may not wish its data to be exposed to a competitor. In this way, protecting data about single individuals and single companies has overlapping issues in common.

6.1.2 Data sensitivity

Data can have a variety of sensitivity levels, including sensitivity due to the data pertaining to a personal matter for an individual that they would prefer to keep private (e.g. health data), or due to the data pertaining to a commercially or organisationally important matter that a group prefers to keep confidential (e.g. company performance data).

When it comes to personal data, there are a variety of data types and their relative sensitivity is often dependent on the person and context. Some examples of data types that can be sensitive and therefore carry higher re-identification risks in different contexts are:

- Health or biometric data (often considered amongst the most sensitive data)
- Relationship data
- Data about beliefs, memberships or affiliations
- Law enforcement data
- Financial data
- Location data, and
- Communication data.

6.1.3 Whether unit record data or aggregated data will be used or disclosed

- Re-identification risks are higher in ‘de-identified’ unit record data (that could potentially be linked with other data sets) than in aggregated data sets (where individual unit records have been aggregated into broader cohorts).
- The choice of de-identification technique and other safeguards must be guided by the level of risk associated with unit record or aggregated data.
- Disclosure of analytics results to a broader audience outside the analytics environment will require a higher level of data aggregation than unit record level and cohorts not smaller than 5.

6.1.4 Who will access the data for which purpose

- Analysts accessing data in the analytics environment must be trusted users with the appropriate security vetting, technical capability, and privacy and security training. This will enable ‘de-identified’ unit record data to be provided for analysis within the analytics environment.
- Authorised, external researchers may in certain circumstances be provided with access to ‘de-identified’ data. Access must be provided to only those data elements that are required for the particular authorised research project, within a secure access environment, and depending on the nature of the project and security vetting of the researcher, may include access to ‘de-identified’ unit record data.

6.1.5 Controls and end-user trust

As re-identification risk depends on the end-user of the data, end-users must be considered in three groups, with different levels of expectation and control applied to their behaviour, and different types of recourse, should a problem occur:

- **High control users** – users who hold a higher level of trust, due to strong contractual and procedural controls e.g. employees
- **Medium control users** – users who hold some trust, due to some contractual and procedural controls, but not direct control – e.g. employees of other organisations, third party researchers, and
- **Uncontrolled users** – users who should only have access to the data through mechanisms that are known to be immune to re-identification and other attacks. e.g. members of the public.

6.1.6 Models of data collaboration

- There are two major models of data collaboration where data from more than one organisation needs to be analysed jointly:
    - **Centralised models** – data is collected and joined in a single organisation prior to analysis. For analyst access the required data subset is extracted and provided from the complete data set.
    - **Federated models** – data remains at the source institution and the formation of linked data for analysis only happens at the time when the data is required.
- From a privacy perspective federated models are the preferred option, as only the data that is required for analysis is generated, and the overall joint data set is never instantiated in a single organisation.
- A centralised model is more convenient from the perspective of management simplicity, ease of data integration, and control.

6.2 Whether data is being 'de-identified' for analytics under section 18 or for disclosure under section 19

- Data used for analytics under section 18 must achieve the right balance between the data being granular enough to produce useful analytics insights, while being aggregated enough to satisfy the 'reasonable steps to ensure no individual can be reasonably identified' requirement. This requires a consideration of all of the de-identification techniques and other controls applied in the restricted access environment overall.
- The more that data is aggregated, the less useful it becomes for detailed analysis, and the appropriate de-identification technique must be chosen on this basis.
- For example, including 'de-identified' unit record data may be more useful for detailed analysis, but carries a higher risk of re-identification, which means additional administrative controls and technical protections must be implemented in the analytics environment to reduce this risk.
- 'De-identified' data for disclosure under section 19 must be aggregated to a higher level to reduce the risks of re-identification as the data moves out of the more restricted access environment.
- Analytics results will in general be accessed by a wider audience than the analysts in the restricted analytics environment, e.g. in the form of dashboards or reports, which requires a reduction of re-identification risk in the data itself through aggregation to higher level cohorts instead of providing 'de-identified' unit record data.

6.3 Balancing re-identification risk with public benefit

- The acceptable level of re-identification risk within a particular data analytics project must be assessed against the objectives and potential public benefits of that project and whether the risk to individuals are necessary and proportionate to those objectives and benefits.

- This evaluation is required to identify the most effective, proportionate combination of de-identification techniques, and other technical and administrative safeguards, based on the level of risk assessment and the level of residual risk that is appropriate in that context. Implementation of these techniques and safeguards must strike the right balance between achieving useful data analytics insights and mitigating re-identification risks, including assessment of necessity and proportionality.

- In assessing re-identification risks versus potential public benefits, it is important to consider community expectations and the attitudes of the individuals or groups to whom the information relates.

7. Conclusion

The approach to de-identification in this guideline is designed primarily to assist with assessing whether the 'reasonable steps to ensure no individual can be reasonably identified' requirement under section 18 is satisfied.

It is important to recognise that de-identification is not a singular end state, but rather an exercise in proportionate risk management. Many of the common techniques used for de-identification are susceptible to re-identification. As such, 'de-identified' data represents a spectrum of re-identification risks, depending on both the methods used to de-identify the data, and the capabilities and assets of any person seeking to re-identify the data.

Within this context, this guideline is not aimed at providing a prescriptive, template based, tick-a-box guide to de-identification, but rather a set of potential techniques and contextual factors that must be considered in assessing 'reasonable steps' under the Act.

In determining the level of risk that will be acceptable for a particular analytics project, the CDO and data analytics bodies must consider proportionality and necessity as key factors to help determine the balance between privacy protection and useful analytics.

This rights-based balancing approach considers the objectives and potential public benefits of the project and assesses whether the risk to individuals (via re-identification or otherwise, given the various controls that will be in place) is necessary and proportionate to that objective. If there is a clear public benefit to an analytics project under this Act, then the risk of potential harm to individuals must be proportionate and necessary in light of that benefit.

A number of factors are making it increasingly difficult to evaluate the reasonable likelihood of re-identification in a data set:

- increasing complexity of large data sets
- exponential availability of other data sets both within and outside government
- increasing number of analytics tools available to a growing number of people with digital and data capabilities
- lack of widely accepted standards to test the effectiveness of de-identification, and
- lack of clarification in existing legal frameworks and regulatory guidelines around what is considered to be an acceptable level of risk with regards to de-identification in data analytics projects.

As a matter of policy, it must therefore be determined what de-identification techniques and level of access and associated controls are appropriate for the sensitivity of the data and the types of users. Harmonisation of this approach across and within different levels of government would be beneficial, particularly from the viewpoints of governance, training, and communication with the community.

Given the lack of clear, accepted standards for measuring the effectiveness of de-identification processes, the CDO and data analytics bodies must make a policy decision about the levels of residual risk that are appropriate for which level of sensitivity and for which audience. This guideline provides some techniques and factors that are relevant to that policy decision.

8. Resources

This guideline focuses on de-identification as understood within the requirements of the *Victorian Data Sharing Act 2017*.

Numerous de-identification guidelines have been published by other jurisdictions and regulators, which provide further insight into the issues to be aware of.

While the CDO and data analytics bodies do not have to have regard to these other publications, they may find further context and detail helpful:

1. De-Identification of Personal Information – National Institute of Standards and Technology, U.S. Department of Commerce
2. De-identification Decision-Making Framework – Office of the Australian Information Commissioner
3. De-identification of data and information – Office of the Australian Information Commissioner
4. De-identification Guide – Australian National Data Service
5. De-identification Background Paper – Commissioner for Privacy and Data Protection
6. Privacy by design in big data – European Union Agency for Network and Information Security
7. Anonymisation: managing data protection risk code of practice – UK Information Commissioner's Office.