ASEAN Guide on Data Anonymisation

January 2025

Contents

EXECUTIVE SUMMARY

The **ASEAN Guide on Data Anonymisation** (this "**Guide**") is a technical and application-oriented introductory guide to anonymisation of personal data.

Part 1: Introduction

Part 1 of this Guide introduces the Guide's purpose and scope. Specifically, the purpose of this Guide is to provide information and guidance on basic data anonymisation that may be referenced by policymakers, regulators as well as industry organisations within countries who are members of the Association of Southeast Asian Nations ("**ASEAN**"). As member states are increasingly adopting data protection laws, this Guide may be particularly useful as a baseline for adaptation to their specific jurisdictional contexts. To this end, it sets out a general introduction to the anonymisation process and some common anonymisation techniques.

Data anonymisation is a risk-based process of converting personal data into data that can no longer be used to identify an individual, either alone or in combination with other information, by applying relevant techniques and in combination with governance measures. Whether a set of data can be considered no longer able to identify an individual would depend on the level of re-identification risks and the applicable data protection laws. While data anonymisation is not necessarily a specific legal requirement under many data protection laws in ASEAN, practising data anonymisation can assist in the protection of personal data, facilitate compliance with applicable data protection laws and provide additional benefits (e.g., safe sharing and collaboration using data from individuals).

Part 2: Key Concepts and Terminology

Part 2 of this Guide discusses key concepts and terminology at an introductory level, which can serve as a useful reference and promote harmonisation in data anonymisation practices across ASEAN jurisdictions. For example, it sets out the definition of a data attribute and how it may be categorised as a direct identifier, indirect identifier, or target attribute before the anonymisation process. Similarly, it explains identifiability and related concepts, which facilitates effective categorisation of data attributes, application of anonymisation techniques and risk assessments. It also describes typical scenarios (also known as use cases) for anonymisation such as internal and external data sharing, to illustrate the outcomes of anonymisation.

The annexes to this Guide provide a more detailed and technical explanation of various concepts as follows:

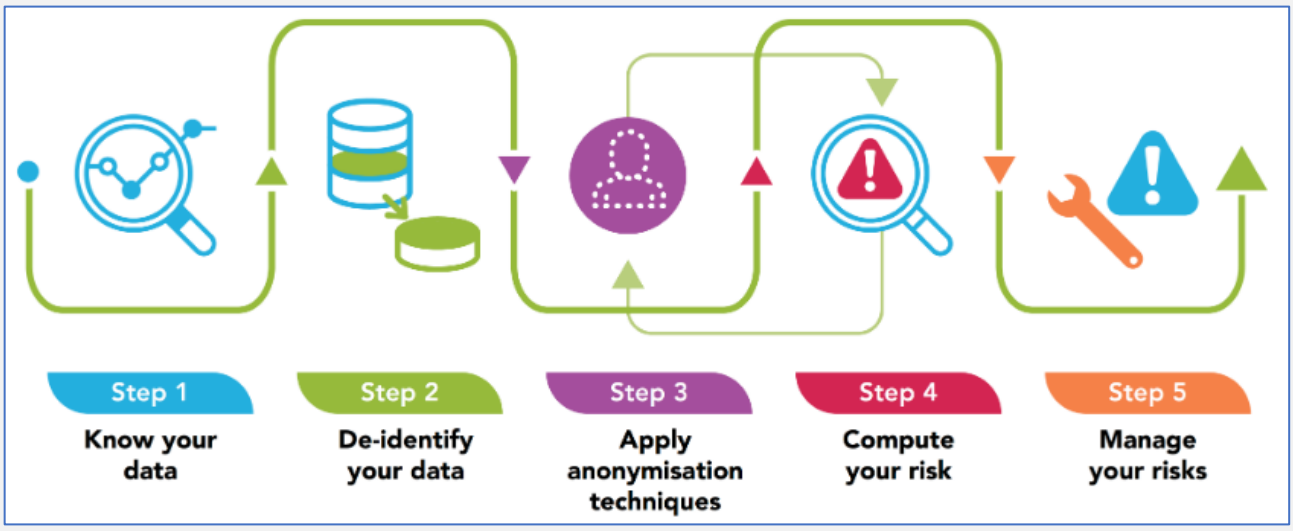
- Annex A: Basic Data Anonymisation Techniques
- Annex B: An Overview on *K*-Anonymity
- Annex C: Common Misunderstandings in Anonymisation
- Annex D: Anonymisation Tools

Part 3: The Anonymisation Process

Part 3 of this Guide provides an overview of the nature of anonymisation techniques in general, briefly summarises good practices for documentation, and sets out key anonymisation steps that can be adopted as part of the anonymisation process. The necessity of tailoring these steps to suit specific requirements, and/or repeating steps to better achieve anonymisation, depends on factors such as the use case and complexity of the data.

Anonymisation Steps

The anonymisation steps in this Guide are summarised in the following diagram[1]:



For avoidance of doubt, the steps 'Apply anonymisation techniques' and 'Compute your risk' (steps 3 and 4 above) can be an iterative process (hence represented in a loop).

STEP 1

Step 1 (know your data) involves understanding the suitability of data for anonymisation, and the appropriateness of anonymisation for the intended use case. There are various

[1] Diagram reproduced with permission from the Singapore Personal Data Protection Commission's Guide to Basic Anonymisation.

factors to consider, such as the nature of use and extent of disclosure. Data minimisation should be practised to exclude any data attributes which are not needed for the use case, and to limit the data to a sample of records rather than the full dataset (where possible).

STEP 2

Step 2 (de-identify your data) involves the removal of direct identifiers from the data and, optionally, using reversible pseudonymisation where there is need to be able to link each record in the (anonymised) dataset back to a unique individual and/or back to the original database.

STEP 3

Step 3 (apply anonymisation techniques) involves the application of anonymisation techniques to indirect identifiers in the de-identified dataset, so that they cannot be easily combined with other datasets that may contain additional information to re-identify individuals.

STEP 4

Step 4 (compute your risks) involves an established risk threshold for anonymisation and the application of procedures to determine whether a sufficient anonymisation level has been achieved. If the risk threshold has not been met, Step 3 (apply anonymisation techniques) should be repeated. A final risk assessment should be conducted and residual risks will need to be reviewed, as this would affect the additional risk management measures / controls that need to be applied in Step 5 below. This is especially important in cases where the final anonymisation level is insufficient to satisfy the legal threshold (i.e. relevant data protection requirements).

STEP 5

Step 5 (manage your risks) involves the imposition of controls / measures in relation to the anonymised data, to further reduce the risks of re-identification of the data. Such measures are usually contractual, administrative and/or technical in nature.

1) INTRODUCTION

PART 1: INTRODUCTION

Purpose of this Guide

The purpose of this Guide is to provide information and guidance on basic data anonymisation concepts and techniques. It is aimed primarily at governments and industry organisations that process personal data and are located in countries who are members of ASEAN, and those working within such organisations. It will also benefit those working in fields of risk assessment and compliance who may need to appreciate and understand the capabilities and limitations of anonymisation techniques in the context of their specific domain.

Data anonymisation may not necessarily be a specific requirement under various countries' data protection laws. However, anonymised data is generally not considered personal data and thus, not subject to data protection laws. Besides that, anonymising personal data would also enable organisations to enjoy the practical benefits summarised at paragraph 1.2. below.

Anonymisation is a risk-based process of converting personal data into data that can no longer be used to identify an individual, either alone or in combination with other information, by applying relevant techniques and in combination with governance measures. Whether a set of data can be considered no longer able to identify an individual would depend on the level of re-identification risks and the applicable data protection laws. The specific type and number of anonymisation techniques as well as governance controls to apply to achieve anonymisation will depend on the sensitivity of the data itself, the intended use case for the anonymised data, and the assessed risks and potential attacks regarding such data.

A proper risk assessment helps to determine the amount of resources that ought to be invested for data anonymisation to strike the appropriate balance between the utility / usefulness and anonymity of the data. In short, anonymisation is a risk-based process which requires understanding the requirements of the intended use case and assessing the risks involved.

Benefits of Anonymisation

Engaging in anonymisation of personal data has several key benefits. These include:

(a) Building trust in organisations' data protection practices;

(b) Enabling the safe use of data while preserving the data's utility and individuals' privacy during analysis and research, which may be carried out with partners through the sharing of anonymised data;

(c) Promoting data sharing and collaboration as anonymised data can be shared with third parties and across jurisdictions safely and without infringing individuals' privacy;

(d) Demonstrating good governance over data and increasing consumers' confidence that their personal data is protected when data is shared amongst businesses and across borders;

(e) Enhancing individuals' privacy and safeguards against data misuse and exploitation, especially when used in combination with governance measures / controls to minimise unauthorised access to data; and

(f) reducing the impact or harm to individuals in the event of a data breach, including identity theft.

Scope of this Guide

This Guide provides a general introduction to the anonymisation process and some common anonymisation techniques[2]. These anonymisation techniques are suitable for data where each record within the data pertains to and represents a single individual. Additionally, the anonymisation process set out in this Guide assumes that the data which anonymisation techniques are applied to are complete and accurate or have been pre-processed so that they are sufficiently complete and accurate for anonymisation. As pre-processing data, sometimes referred to as data cleansing, is a major topic on its own, it is outside the scope of this Guide.

This Guide focuses on tabular and similarly structured data, which is typically stored in Excel sheets, SQL databases, JSON format, CSV format, etc., as these are the most commonly used format to store and process datasets.

Data Protection Landscape in ASEAN

Across ASEAN, member states are increasingly adopting data protection laws. At present, Singapore, Malaysia, Thailand, the Philippines, Indonesia and Vietnam have an existing overarching data protection law[3]. In addition, as of December 2024, Brunei

[2] For further information and resources, please refer to international standards such as ISO/IEC 20889:2018 on privacy enhancing data de-identification terminology and classification of techniques and ISO/IEC 27559:2022 on information security, cybersecurity and privacy protection - privacy enhancing data de-identification framework.

[3] See Singapore's Personal Data Protection Act 2012; Malaysia's Personal Data Protection Act 2010; Thailand's Personal Data Protection Act B.E. 2562 (2019); the Philippines' Republic Act No. 10173 – Data Privacy Act of 2012; Indonesia's Law No. 27 of 2022 on Personal Data Protection; and Vietnam's Decree No. 13/2023/ND-CP on the Protection of Personal Data.

Darussalam and Cambodia are in the process of enacting their own data protection laws[4].

Based on a survey conducted across the ASEAN member states, about half of the ASEAN member states have laws, regulations, guidelines[5], or standards relating to data anonymisation and a corresponding number of ASEAN member states have observed that it is common (and practicable) for private or government organisations to perform data anonymisation in their jurisdictions. While there were indications that simpler anonymisation techniques such as character masking and de-identification were primarily adopted, more sophisticated anonymisation techniques were also sometimes utilised.

Important Note: This Guide is primarily a 'technical and application-oriented' introduction to the common concepts around anonymisation in the context of personal data protection laws. Data protection laws vary across the ASEAN member states, and the legal definition and treatment of 'anonymised data' and other concepts introduced in this Guide may also differ across jurisdictions. Nevertheless, this Guide aims to set out a risk-based approach to anonymisation that can serve as a useful reference across ASEAN (which can then be adapted for each jurisdiction's specific requirements).

[4] See, for instance, Brunei Darussalam's Authority for Info-communications Technology Industry's website (accessible at: https://aiti.gov.bn/regulatory/pdp/), and the Ministry of Post and Telecommunication of Cambodia's public announcement dated 4 November 2022 (accessible at: https://opendevelopmentcambodia.net/announcements/press-release-on-the-progress-of-digital-policies-and-regulations-in-the-digital-sector-in-cambodia/).

[5] See Singapore's Personal Data Protection Commission's Guide to Basic Anonymisation, accessible at: https://www.pdpc.gov.sg/help-and-resources/2018/01/basic-anonymisation.

2) TERMINOLOGY AND KEY CONCEPTS

PART 2: TERMINOLOGY AND KEY CONCEPTS

2.1 Key Terms

The concept of anonymisation is fairly new to many organisations. Hence, various key terms sometimes bear a different meaning when used by organisations, as compared to their specific meaning under different data protection laws. For the purposes of this Guide, the following table provides the definitions of key terms used in this Guide[6]:

Term	Definition / Explanation of Concept
Personal data	Generally, this refers to data about an individual who can be identified from that data alone or in combination with other information to which an organisation has or is likely to have access to.
Non-personal data	This refers to data that does not relate to an individual.
De-identified data	This generally refers to data from which direct identifiers (see below for the definition of "direct identifiers") have been completely removed, voided (set to "null") or overwritten.
Data attribute	This refers to features / characteristics of a dataset, *e.g.,* customer names, products purchased and so on. Hence, data attributes are the inputs in an anonymisation process.
Anonymisation	This refers to a risk-based process of converting personal data into data that can no longer be used to identify an individual, either alone or in combination with other information, by applying relevant techniques and in combination with governance measures. Whether a set of data can be considered no longer able to identify an individual would depend on the level of re-identification risks and the applicable data protection laws (see also definition of "anonymised data" below).

[6] Note that these are not legal definitions and are intended only to provide guidance on the terms used in this Guide. The terms in the table may have variations in their specific legal definitions across different jurisdictions.

Term	Definition / Explanation of Concept
Anonymised data	This refers to data to which anonymisation techniques have been applied (if necessary, in combination with governance measures) to achieve a low level of re-identification risk, so as to meet a particular legal and/or industry-accepted (e.g., risk-based) standard. Generally, anonymised data is not considered personal data under a jurisdiction's data protection laws. Whether or not data is sufficiently anonymised would depend on the applicable laws. Hence, organisations should refer to the regulatory guidance on anonymisation standards in their respective jurisdictions (if any), to ensure compliance with relevant data protection legal requirements.

2.2 Identifiers and Target Attributes

It is important to understand how data attributes, which are inputs for the anonymisation process, are categorised before anonymisation is performed. This facilitates a proper execution of risk assessments and achievement of desired outcomes. Data attributes are usually categorised as follows:

(a) Direct identifier: A direct identifier (also referred to as "unique identifier"[7]) is usually seen as a 'high risk' attribute. These are data attributes that are unique to an individual and can be used to identify the individual. Because a person may be identifiable from a single direct identifier, all direct identifiers need to be removed as part of the anonymisation process.

(b) Indirect identifier: An indirect identifier (also referred to as "quasi-identifier") is usually seen as a 'medium risk' attribute. These are data attributes that are not unique to an individual but can potentially identify an individual when combined with other indirect identifiers. Many data anonymisation techniques primarily focus on the treatment of indirect identifiers in order to achieve a sufficient level of anonymisation.

(c) Target attribute: A target attribute often contains the main utility of the dataset (i.e., they are pieces of useful information associated with the individual). It is usually seen as a 'low risk' attribute in terms of its potential to re-identify the relevant individual as it is usually information that is not publicly or easily

[7] Note that although these terms are sometimes used interchangeably, a unique identifier is not always a direct identifier because sometimes a pseudonym, record identifier or foreign key can be unique but not identifying.

accessible to others. Nevertheless, such attributes may be sensitive and may result in high potential for adverse effect to the individual if disclosed.

The appropriate categorisation for any given data attribute depends on the broader context in which the data attribute is located. For example, data attributes that would ordinarily be indirect identifiers in larger datasets could become direct identifiers in smaller datasets (e.g., information about a small group of people, each person being of a different age). Hence, the categorisation of data attributes is not always a trivial process and often requires some deliberation.

Some examples for a typical categorisation of data attributes are listed below. For avoidance of doubt, these are not intended to serve as a legal definition or classification under any of the laws of the ASEAN member states.

Direct Identifiers	Indirect Identifiers	Target Attributes
• Account number • Birth certificate number • Email address • Full Name • Mobile phone number • National identification number • Passport number • Social media username • Biometric data	• Address • Postal code / Postcode • Age • Date of birth • Sex / Gender • Marital status • Race • Company name • Job title • Vehicle license plate number / vehicle registration number • Internet Protocol address • Weight / Height • Geolocation	• Financial transactions • Retail purchases • Salary • Credit rating • Insurance policy • Medical diagnosis • Vaccination status

2.3 Identification, De-identification and Re-identification

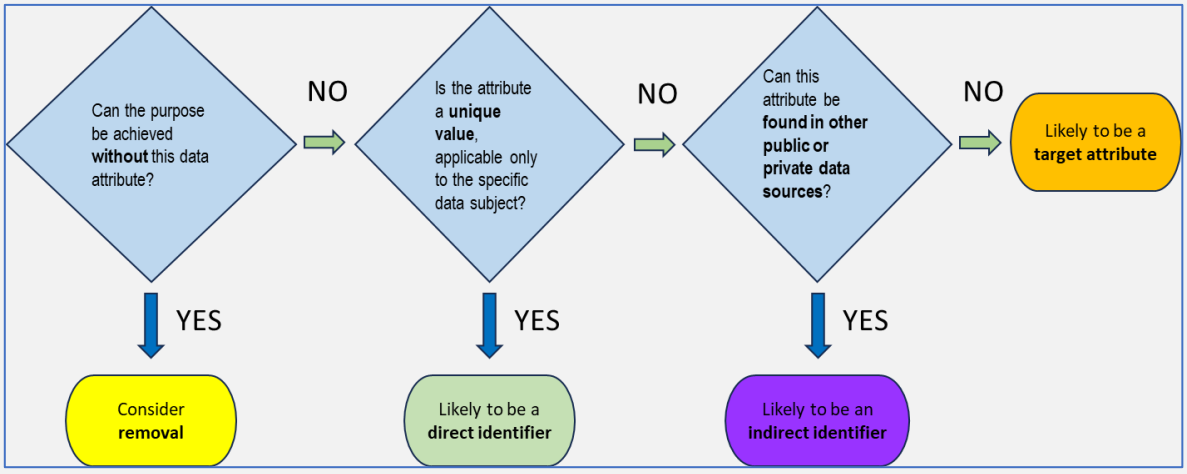
To properly place attributes or identifiers of a given dataset into one of the three categories above, it is important to understand the process of "identification", "de-identification" and "re-identification", as well as what it means for an individual to be "identifiable" from a dataset. These terms can be understood as follows:

(a) Identifying, Identifiable: As an action, "identifying" and "identification" refers to a process of establishing one or more individuals' identity from the data. When evaluating a dataset, "identifying characteristics" refers to the information content contained in the dataset which is sufficient to establish the identity of one or more individuals. Hence, an individual is "identifiable" from data if it contains identifying characteristics pertaining to the individual.

(b) De-identification: De-identification usually refers to a complete removal, voiding (setting to "null") or overwriting of direct identifiers in the dataset. This does not necessarily result in complete anonymisation of the data – individuals may be identified from indirect identifiers when combined with other information.

(c) Re-identification: This term is commonly used to refer to the identification of an individual from a dataset that was previously de-identified or anonymised. It can sometimes involve the reversal of previous steps taken to perform de-identification or anonymisation, or the combination of various datasets to obtain identifying characteristics (as described above).

A general approach to determine the respective attribute type (e.g., direct identifier, indirect identifier, and target attribute) in the absence of a specific list from the relevant data protection authority ("DPA") can be gleaned from the chart below. Organisations may wish to consider establishing and following a similar approach to sort data attributes into their respective attribute types.



2.4 Typical Scenarios for Anonymisation

Anonymisation typically involves removal of direct identifiers and modification of indirect identifiers. Target attributes are usually left unchanged, except where the purpose is to create fictitious data.

To illustrate the outcomes of anonymisation, the following examples of use cases describe common scenarios (i.e., use cases) and set out common considerations during anonymisation when dealing with the same data for different purposes. Guidelines for the process by which data can be anonymised (after determining the relevant use case) are described below at **Part 3: The Anonymisation Process.**

It should be noted that the examples below are for illustration only. When carrying out their own anonymisation exercises, organisations will need to assess the appropriate balance between the data utility and depth of anonymisation (in terms of the techniques and controls applied) required for each of their specific use cases, taking into account the amount and types of data involved, specific risks and potential attacks within the use cases, and the applicable laws in each ASEAN member state.

Internal data sharing (low risk)

Example	De-identified customer data shared between the research & development department and the products department for analysis and in-house development of new goods and services.
Description	Only direct identifiers (e.g., names and customer IDs) are removed from the dataset while indirect identifiers (e.g., age, gender, address) and target attributes are left unaltered to support the intended use case. The de-identified data is still personal data as individuals are likely to be re-identifiable from the other attributes in the data. Hence, even though the data is only shared within the organisation, it is still advisable in such cases to practice data minimisation (i.e., removing any indirect identifiers and/or target attributes which are not needed for the use case). This will provide an additional layer of protection to the de-identified data.

Internal data sharing (high risk)

Example	Anonymised data on the spending habits and demographics of high net-worth customers shared with in-house loyalty teams to create differentiated customer value propositions.
Description	Anonymised data (using the appropriate anonymisation technique(s) to treat both direct and indirect identifiers) should be shared instead of only de-identified data in cases where: • the internal data sharing does not require detailed personal data (e.g., for trend analysis); and/or • the data involved is more sensitive and/or granular in nature (e.g., financial information).

External data sharing

Example	Anonymised customer data shared between an in-house marketing team and external marketing partner for analysis of customer profiles and development of marketing campaigns.
Description	In such cases, the datasets are shared with an authorised external party for business collaboration purposes. Hence, appropriate anonymisation techniques can be applied to the datasets to help organisations better comply with data protection requirements.

Long-term / archival data retention

Example	Retention of anonymised data (where the legally permissible retention period in relation to the personal data has passed) for the purpose of data analysis and historical analysis of customer trends.
Description	Anonymisation techniques can be used to convert personal data to non-personal data. This allows the organisations to legally retain the resultant data as useful business records for long-term data analysis when there is a retention limitation obligation applicable to the original personal data. Take note that this use case is different from the others as: a) Since such data is to be retained beyond the legally permissible period for retention of personal data, no copies (whether original or otherwise) of the data, or sub-sets of the data, should contain personal data. b) In contrast, the other use cases typically involve organisations retaining both the anonymised and original personal data (assuming that the legally permissible retention period for the personal data has not been exceeded). c) The organisation should ensure that the anonymised data will not be re-identifiable, as this use case demands stronger (and irreversible) anonymisation techniques to be applied in the context where the legally permissible retention period has passed. If the data is anonymised, but the organisation has the ability to reverse the anonymisation, this would potentially result in non-compliance with the retention limitation obligation.

3) THE ANONYMISATION PROCESS

PART 3: THE ANONYMISATION PROCESS

3.1 Overview of Anonymisation Techniques

The anonymisation process described in this Guide consists of several key steps. Before considering the steps in detail, it will be helpful to first have a general understanding of the nature of anonymisation techniques.

Anonymisation techniques are applied at Step 3 of the anonymisation process (described below). They consist of various methods to remove identifying characteristics from personal data. Different anonymisation techniques have different characteristics and modify the data in different ways (see Annex A for further details on common anonymisation techniques). Moreover, several anonymisation techniques can be used in combination on a single data attribute.

The appropriateness of a technique depends on the categorisation and the characteristics of the data in question. For instance, certain techniques (e.g., character masking) can be more appropriate for direct identifiers. On the other hand, techniques such as aggregation can be better suited for indirect identifiers. Another characteristic to consider is whether the attribute value is a continuous value (e.g., height = 1.61 m) or a discrete value (e.g., "yes" or "no") because certain techniques (e.g., data perturbation) may be more suitable for continuous values.

As explained in greater detail below, the choice of anonymisation techniques also depends on the intended use case for the resultant data. If the use case requires more granularity / details in the data or to retain the data format, then certain techniques, such as aggregation or masking, may not be appropriate as the details and data format may not be retained.

Anonymisation techniques also modify data in significantly different ways. Some modify only part of an attribute (e.g., character masking); some replace the value of an attribute across multiple records (e.g., aggregation or generalisation); some replace the value of an attribute with an unrelated but unique value (e.g., pseudonymisation); and some remove the attribute entirely (e.g., attribute suppression).

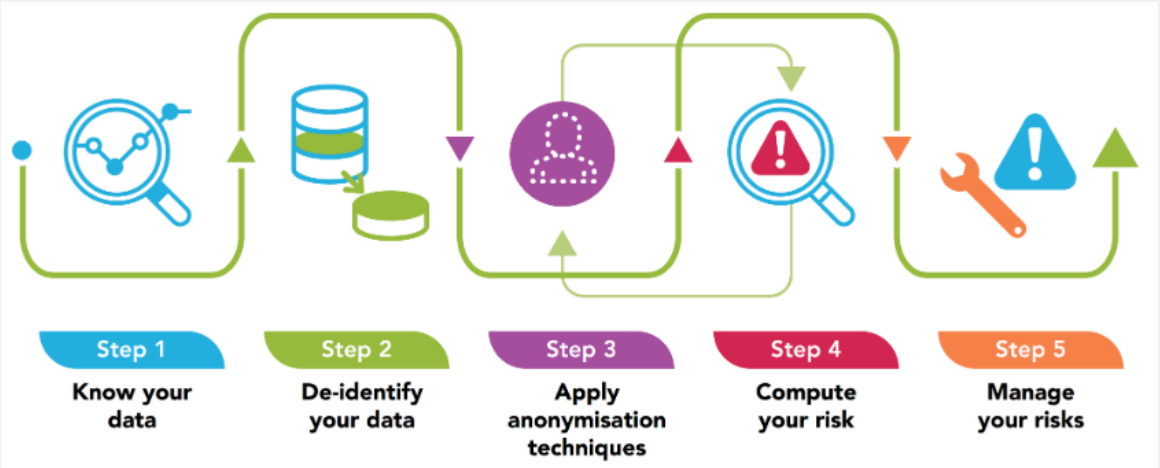
3.2 Key Anonymisation Steps

Overview of steps

To manage the multifaceted anonymisation process, a step-by-step process can be adopted. Although this Guide explains each of the steps in the anonymisation process, organisations may have to tailor these steps to fit their specific requirements in different cases. Additionally, certain steps may sometimes need to be repeated multiple

times during an anonymisation exercise. The necessity of this depends on factors such as the use case and complexity of the data.

The paragraphs below set out an overview of the various steps. Further detailed explanations of these steps, especially in terms of anonymisation techniques and *k*-anonymity, can be found in the Annexes to this Guide.



For avoidance of doubt, the steps 'Apply anonymisation techniques' and 'Compute your risk' (steps 3 and 4 above) can be an iterative process (hence represented in a loop).

Good practices for documentation

As part of any anonymisation project, it is a good practice to document (a) the risk assessment process, (b) the details of the anonymisation approach and the techniques chosen, and (c) the parameters and their justifications. After these steps are taken, a final approval should be obtained from qualified and authorised personnel. This will provide records (including implementation details) which can guide anonymisation in future projects, as well as facilitate future reviews, maintenance and improvement efforts, or even audits of the current practices. Documentation will also facilitate responses to any investigations or queries from authorities and may be necessary for the purposes of complying with applicable data protection laws relating to requirements to maintain records of data processing. As such documentation could potentially enable re-identification of the data, they should be safeguarded from disclosure to unauthorised parties.

Step 1: Know your data

Understand your data and use case

When determining whether to anonymise data before using or disclosing it, organisations should note that not all data can be effectively or meaningfully anonymised. Importantly, the decision to anonymise data and the extent to which it is anonymised depends primarily on the suitability of the affected data.

'Knowing one's data' at the start is therefore important, such that the suitability of the data for anonymisation can be assessed effectively. For example, a local primary school will be aware that its students' data will contain a limited age range and most students will be of the same nationality, while an international website portal can have a wide variety of personal data relating to users of different ages and nationalities. Anonymisation techniques will be easier to apply in the case of the local primary school because of the lower variability of identifying characteristics in the dataset. In contrast, more sophisticated techniques and/or greater effort will likely be needed to anonymise the international websites' user data.

The above being said, before opting for data anonymisation, organisations should also consider whether anonymisation might be inappropriate for their intended use case. In particular, use cases requiring a greater level of detail / granularity may render anonymisation techniques unsuitable due to the loss of granularity of data that would occur when the techniques are applied to the data.

In summary, organisations should establish the use case for the data and the suitability of anonymisation at the start of the anonymisation process, so that more appropriate anonymisation techniques will be chosen. In this regard, the following factors should be taken into consideration:

(a) Nature and uniqueness of data: The nature of the data itself will affect the extent to which direct and indirect identifiers need to be removed or altered so that the dataset can no longer be used to identify individuals.

(b) Nature of use and extent of disclosure: The intended nature of use and extent of disclosure (also called the release model) of the anonymised data will affect the risks. The release model affects (i) how much additional information (apart from the disclosed data) would be needed for the anonymised data to be re-identified (which in turn affects the assessment of how much of such additional information, if any, should be given to the recipient) and (ii) the assessment of what mitigating actions need to be taken to prevent re-identification.

(c) Potential impact on individuals: Organisations should always consider any potential adverse impact on the data subjects if they were to be re-identified from the data subsequently (*e.g.*, due to an attacker taking steps to ascertain their identities). This is especially important if the dataset involves sensitive information, such as health records and financial information. In such a scenario, the organisation should consider whether it would be appropriate for such data to be used or shared. This factor should be considered regardless of whether the organisation ascertains that the risk of re-identification is low (see below Step 4 (Compute your risk)).

(d) Information loss and utility: Organisations should always ensure that the anonymisation is carried out specifically for the intended use case (*i.e.*, the purpose of the data). Generally, as the level of anonymisation increases, the utility of the dataset decreases. As such, the organisation must determine how to balance the trade-off between the required utility of the data (which will depend on the intended use case) and the risk of insufficient anonymisation (which could lead to re-identification). To better manage the risks of insufficient anonymisation and safeguard against re-identification risks, organisations should apply additional protection measures during the anonymisation process, as well as when the data is shared or disclosed (see Step 5 (Manage your risks)).

Practise data minimisation

Once a proper understanding of the data and use case is established, data minimisation should be carried out before proceeding to Step 2 (de-identification). Data minimisation involves the exclusion of any data attributes (including direct, indirect and target attributes) which are not needed for the use case. Generally, organisations should also consider if the dataset to be anonymised can be limited to a sample of records rather than the full dataset.

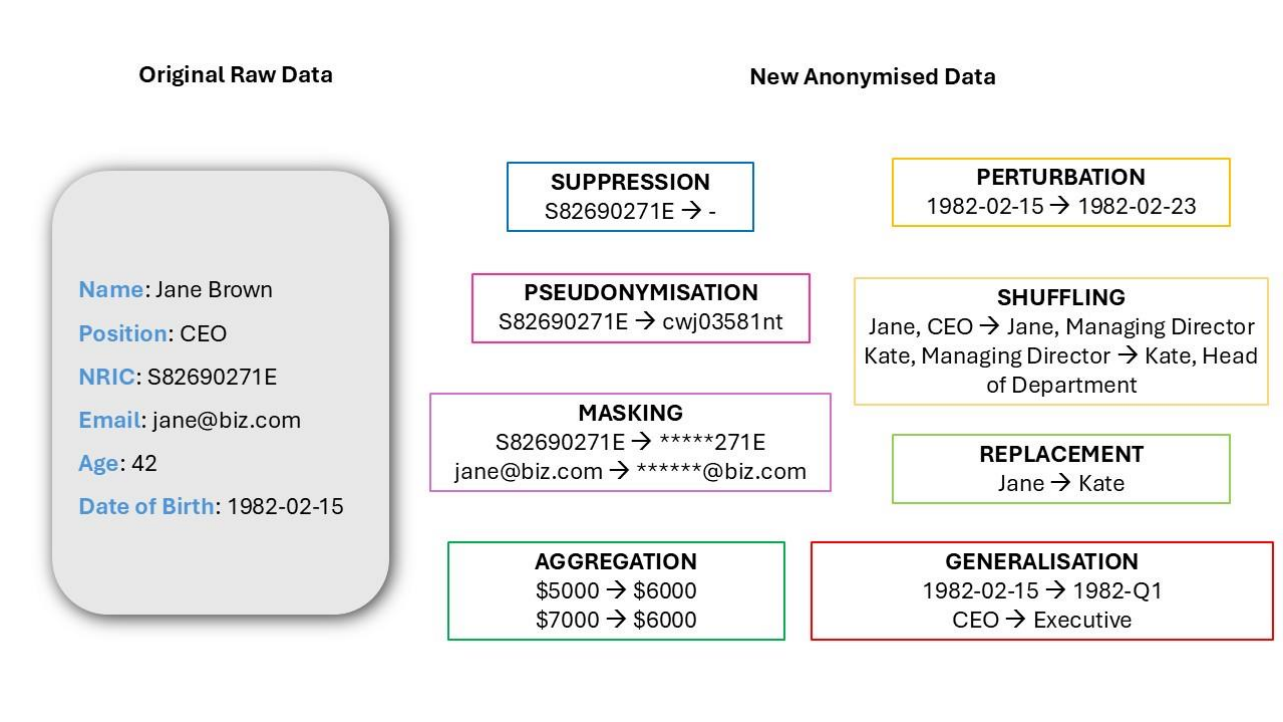
Step 2: De-identify your data

As mentioned above, even after carrying out data minimisation, the dataset may contain direct identifiers. In such cases, the next step to achieve anonymisation of the dataset is de-identification of the data by removing all direct identifiers.

If there is a need to be able to link each record in the dataset back to a unique individual and/or back to the original database, organisations can use reversible pseudonymisation. This involves creating a unique pseudonym, such as a string of numbers (see Annex A for further details on pseudonymisation). The method of generating and assigning pseudonyms to the records in the dataset should not allow a third party to guess or deduce the original direct identifiers from the assigned pseudonyms.

For organisations that wish to retain the ability to link the de-identified data back to the original record, the details of the assignment of pseudonyms (i.e., the mapping of the pseudonyms) should be stored securely, to prevent any unauthorised persons from performing re-identification.

Step 3: Apply anonymisation techniques



In this step, anonymisation techniques are applied to the indirect identifiers so that they cannot be easily combined with other datasets that may contain additional information to re-identify individuals.

A possible grouping for a set of sample records might be:

Example 1: Classification of Data Attributes in an Employee Data Record

StaffID	Name	Department	Gender	Date of birth	Start date of service	Employment type
39192	Sandy Thomas	Research & Development	F	08/10/1971	02/03/1997	Part-time
37030	Paula Swenson	Engineering	F	15/05/1976	08/03/2015	Full-time
22722	Bosco Wood	Engineering	M	31/12/1973	30/07/1991	Full-time
28760	Stef Stone	Engineering	F	24/12/1970	18/03/2010	Part-time
13902	Jake Norma	Human Resources	M	15/07/1973	28/05/2012	Part-time

Direct Identifiers **Indirect Identifiers** **Target Attributes**

Example 2: Classification of Data Attributes in a Customer Data Record

CustomerID	Name	Gender	Date of birth	Postal code	Occupation	Income	Education	Marital status
56833	Jenny Jefferson	F	05/08/1975	570150	Data scientist	$13,000	Masters	Widowed
50271	Peter G	M	14/12/1973	787589	University lecturer	$12,000	Doctorate	Married
53041	Tim Lake	F	02/03/1985	408600	Researcher	$7,000	Doctorate	Divorced
17290	Remy Ray	M	27/03/1968	570150	Database administrator	$8,000	Bachelor	Married
52388	Walter Paul	M	25/06/1967	199588	Architect	$10,000	Masters	Single

Direct Identifiers **Indirect Identifiers** **Target Attribute** **Indirect Identifiers**

Different anonymisation techniques and the applied parameters will affect the precision and thus the utility of the resultant data. During the entire anonymisation process, a single attribute may require the application of more than one option or technique to achieve the pre-determined anonymisation level. For example, a date of birth may first be generalised into a numerical age (in years), and if it turns out to be insufficient during the risk assessment stage, further aggregation into 5-year periods may be needed. If that is still insufficient, the data may be further generalised into a category like minor, adult, and senior.

Outlier records or attributes values (i.e., those values that are unique and cannot reasonably be generalised or grouped with other records) that are resistant to suitable anonymisation may have to be removed or voided if possible.

Organisations may refer to Annex A for the basic anonymisation techniques. The techniques listed in Annex A are not intended to be authoritative or exhaustive. It is important that due diligence is applied in selecting the appropriate technique (or combination of several techniques) for specific use cases, taking into account the appropriate balance between utility and anonymity required for the data in each use case.

Step 4: Compute your risks

The risk threshold for the anonymised data should be decided upfront by organisations for more objectivity. This is similar to a Data Protection Impact Assessment ("**DPIA**") which is most effective before a project is initiated and committed, and where implementation details are governed by the outcome of the DPIA. Likewise, during an anonymisation process, the type and depth of the anonymisation techniques should be guided by the risk threshold.

In this regard, it is important to establish the relevant terminology when establishing the types of risks and attacks that might apply to anonymised data:

(a) Reversibility: Data anonymisation aims to be 'irreversible' such that it would not be feasible to recreate parts of the original data. However, there may be cases where an organisation applying anonymisation intentionally retains the ability to recreate (or at least trace back to) the original data from the anonymised data. For instance, in the context of outsourcing data analytics on health data, or contact tracing, if the analysis finds (new) health related indicators, it would be beneficial that the organisation could link the respective subset of the anonymised data (typically the target attributes) back to the individual(s), so that they can be notified.

Reversibility, however, must be limited to the organisation which had the original data. In this regard, the organisation must establish controls such that only authorised personnel can reverse the data, for example via access to the mapping between pseudonyms and the original direct identifier.

(b) Singling out: When a unique record can be determined to relate to a specific individual, then this 'singles out' this one record from the rest. Outlier record, i.e., record with very unique data attributes as compared with the other records, are generally more susceptible to singling out. Singling out may not always imply identification, as the record itself may not have sufficient information to identify a specific individual. However, in the context of linking attacks, where, for instance, the attacker already knows a group of people (a typical example would be a group of VIPs), singling some records pertaining to that group can lead to inference attacks and other disclosures.

(c) Attribute disclosure: This refers to determining that an attribute described in the dataset belongs to a specific individual, even if the particular record cannot be distinguished with a high level of confidence. For instance, a dataset containing anonymised client records of an insurance broker shows that all her clients above the age of 60 have purchased a life insurance policy from XYZ Company. If it becomes known that a particular individual is above 60 years old and is a client of the insurance broker, it can be deduced that the individual had purchased a life insurance policy from XYZ Company. This is notwithstanding the fact that the particular individual's record cannot be distinguished from others in the insurance broker's dataset.

(d) Inference disclosure: This refers to drawing an inference about an individual even if he/she is not in the dataset through the statistical properties of the dataset, with a high level of confidence. For example, if a dataset released by a nutritional scientist reveals that 90% of the test subjects who are female have a certain

health condition at the end of the experiment, an inference may be drawn regarding an individual who is not in the scientist's dataset. Although no one was identified in this case, new information, which is potentially personal data about someone was disclosed. This type of attack is more relevant to rare or sensitive data attributes.

Additionally, there are at least 4 key factors that ought to be considered when performing a risk assessment of possible risks and attacks to the anonymised data:

(a) Recipient's attack capabilities and motivations: As far as practicably possible, the ability and the motivation of the data recipient to re-identify individuals from the data should be accounted for in the risk assessment (see e.g., below regarding the Motivated Intruder Test).

(b) Recipient's knowledge of techniques used: The risk of re-identification will usually be higher if the data recipient is aware of the anonymisation techniques applied to the data.

(c) Public versus special knowledge: Even after the risks have been established and appropriate safeguards put in place to minimise the risks, there remains the risk of re-identification of the anonymised data by persons with special (including prior) knowledge of a particular individual. This risk arises even if ordinary members of the public or an organisation would not have such knowledge. For instance, a patient's doctor who is reading a medical journal containing datasets that incidentally contains his patient's data might be able to recognise the patient's medical profile from the data used (even after such data was assessed to be sufficiently anonymised). If such a risk is identified, additional steps might need to be taken to further anonymise the data.

(d) Legal inhibitions: Some jurisdictions may have explicit clauses in their data protection laws which provide that attacks and identification attempts on anonymised data (and/or their additional measures) are offences. Practically, this can assist to dissuade potential attackers and hence the risk of attacks would be somewhat lower. Nevertheless, organisations in these jurisdictions should not rely solely on such legal inhibitions, as they will likely still be legally responsible to ensure sufficient anonymisation. Hence, some form of anonymisation techniques will usually still need to be applied.

The above factors should be taken into account when applying the procedures for risk assessment. The following section briefly introduces the "Motivated Intruder Test", which is an example of a risk measure commonly adopted by organisations.

The Motivated Intruder Test, as adapted from the UK Information Commissioner's Office's *Anonymisation: Managing Data Protection Risk Code of Practice*, is a helpful baseline test to assess the re-identification risks for the data.

The test considers whether a reasonably competent intruder[8] would be able to identify individuals from the anonymised data (possibly in combination with other data) *if* motivated to attempt this. The intruder is taken to:

(a) Not have or apply specialist knowledge;

(b) Not gain access to data via specialist equipment;

(c) Have access to common resources (e.g., libraries, the Internet, and publicly available information); and

(d) Use common investigative techniques (e.g., enquiring with people who may have additional knowledge of the data subject's identity).

The test further takes into account the strength of different motivation(s) (e.g., financial benefits, causing public mischief, political purposes) and resource(s) of the intruder.

'Risk' is commonly expressed in qualitative terms like low/medium/high, or in quantifiable terms of probability / likelihood of an event. The risk threshold is an indicator for the maximum acceptable risk. In particular, *k*-anonymity is popular as a risk threshold because it is a quantitative and objective measure of linkability and potentially, the re-identifiability of anonymised data (see Annex B for more details on *k-anonymity*).

Establishing the *k*-value as a risk threshold is typically a policy decision, depending on whether the relevant DPA has issued specific guidance in this regard or subjects organisations within the jurisdiction to risk-based reviews. As an example, Singapore's Personal Data Protection Commission ("**PDPC**") recommends a *k*-anonymity value of 5 or more for external data sharing. Where specific guidance from the relevant DPA is not available, the *k*-anonymity value is typically decided internally by organisations based on risk appetite and results from DPIAs.

For the purpose of risk assessments, measures like *k*-anonymity treat target attributes equally (e.g., they are left unchanged). *k*-anonymity is typically relied on where the risk assessment is almost exclusively based on a linking attack. If other risks or attacks are relevant and critical for the data, additional risk measures may be needed, such as *l*-diversity and *t*-completeness. As such, while *k*-anonymity is considered a good

[8] "Intruder" in this context is not limited to adversaries, but includes the intended recipient, as the test is part of the risk assessment concerning the appropriate level of anonymisation carried out on the dataset.

minimum baseline measure to have, it may not always be sufficient on its own (see Annex B for further details).

The repeated application of techniques (i.e., Step 3) and determination of anonymisation level (i.e., Step 4) continues until 'enough' records meet the threshold. What is 'enough' might be decided based on the utility requirements, which balance further loss of granularity against further exclusion of entire records (or even attributes).

In the context of *k*-anonymisation, the risk, expressed via a *k*-value, is taken as the probability of 1/*k* that (re-)identification will be successful. In the case of $k = 5$, this means a 20% chance of a specific record, within a group of 5 identical records, being successfully re-identified to an individual by an attacker.

Once the dataset has been generalised and trimmed such that all records and attributes fit the given threshold(s), a final risk assessment should follow.

Residual risks will need to be reviewed in the context of how the data is shared or used, for example, where the same data is disclosed to different parties, or where the same party receives updated versions of data.

This final risk assessment will serve as assurance that the actual data meets the intended anonymisation level and help to determine the depth and rigor of additional controls, which may have to be imposed on the receiving party.

This step also sets the baseline for future regular re-assessments, which should be conducted as an ongoing process to ensure the continued effectiveness of anonymisation applied to personal data. New data may become available or new attacks may be found, which could render the disclosed dataset vulnerable. Importantly, the robustness of the actual implementation still needs to be regularly confirmed against specific attacks (old and new). It would also be prudent to involve an independent party for this step.

In situations where there is a higher assessed risk on a dataset, this may require the application of additional / multiple anonymisation techniques to the relevant data attributes.

Step 5: Manage your risks

As a final step, additional measures can be put in place to safeguard the anonymised data. In particular, contractual, administrative or technical controls can be imposed.

In determining the appropriate controls to be imposed on the anonymised data, organisations should primarily consider whether there are any risks of re-identification

on the anonymised data and in accordance with guidance on best practices for protection by the relevant DPA in each of the ASEAN member states. In the absence of any specific guidance by the regulator, it is nevertheless good practice to put in controls to safeguard data against any unauthorised access. Organisations may take reference from PDPC's Guide to Basic Data Anonymisation for examples of controls to manage the re-identification and disclosure risk of anonymised data.

In addition to anonymisation, organisations will benefit from keeping abreast of developments in other areas, including tapping on the rapidly growing area of privacy enhancing technologies (also known as PETs) to enhance the protection of personal data[9].

[9] PETs are techniques that allow the processing, analysis and extraction of insights from data without revealing the underlying personal or commercially sensitive data. Some PETs are considered quite different from anonymisation (e.g. zero knowledge proof), while others tend to be grouped closer to the domain of anonymisation although they are not anonymisation techniques (e.g. differential privacy). For more information on PETs, see, for example, OECD, "*Emerging Privacy Enhancing Technologies – Current regulatory and policy approaches*", OECD Digital Economy Papers, March 2023, No. 351.

ANNEXES

ANNEX A: Basic Data Anonymisation Techniques

This section provides an overview of basic anonymisation techniques with their commonly used options and variations[10]. Each technique is briefly described and illustrated through the hypothetical scenario of "BestBooks" below. The list of techniques is not exhaustive; real life applications usually require more bespoke modifications. Most of the techniques explained below modify the values at the attribute (e.g., cells in a column) level, whereas some modify the record (e.g., row) level and others completely do away with individual values. As suggested by the BestBooks example, there is no single best way in which sequence attributes and techniques should be addressed and it may be that a few rounds of anonymisation is required with different techniques to achieve the intended result.

As data can be managed in many formats and platforms, no coding examples are provided on how to achieve the modifications; free anonymisation tools are also available online (see Annex D).

However, while the technical aspect of these techniques is not complicated, choosing the right level of anonymisation and assessing the respective risks is a core component of the anonymisation process from beginning to end, as explained in the main part of this Guide.

A bookshop ("**BestBooks**") wants to analyse its data to identify various business improvements it can make, such as a suitable extension of its offerings, a better strategy to stock up on their books, and create a recommender system in conjunction with better route planning for deliveries.

BestBooks has recently hired a data analyst to explore these areas of interest. The initial discussion with the analyst identified a group of attributes for consideration. However, the CEO, in collaboration with the data protection officer ("**DPO**"), decides that the new analyst should not have access to all data and instructs the team to anonymise selected attributes. First, the DPO determines the attribute types and sets the target risk threshold as $k = 3$[11], seeing this as an internal sharing scenario where the analyst is an employee and is obligated under the company's policy not to attempt any re-identification of the data.

Together with the analyst, they then explore what utility they would like to achieve and explore the appropriate anonymisation techniques to apply, resulting in the following:

[10] See also the European Union's Article 29 Data Protection Working Party Opinion 05/2014 on Anonymisation Techniques for an analysis on the effectiveness and limits of existing anonymisation techniques against the European Union's legal background of data protection.

[11] The value $k = 3$ as well as the categorisation of attributes are not meant to be representative for a similar use case, these are for illustrative purposes only in the examples used in this Guide.

Category of Attribute	Direct	Direct	Indirect	Target	Target	Target
Attribute Type	Customer ID	Name	Postal Code	Preferred Delivery time	Membership type	Education
Anonymisation Technique	RandomPseudo	Suppress	Mask and swap	Generalise to 2 hour	Swap	*Not modified*

Category of Attribute	Direct	Target	Target	Target	Direct
Attribute Type	Address	Discount%	Order Date	Title	OrderID
Anonymisation Technique	Remove	Aggregation average	*Not modified*	*Not modified*	Suppress

A sample of the original data is available in Appendix 1, and the dataset after anonymisation techniques have been applied can be found in Appendix 2.

Note: For the purposes of demonstrating the anonymisation techniques used in the illustrations, only excerpts of the dataset from Appendices 1 and 2 are reproduced.

Record Suppression

Description	Remove (or void) an entire record/row. In contrast to most other techniques, which affect a single attribute across (typically all) records, this technique affects a single, but entire record.
When to use it	Remove outlier records which are unique or do not meet other criteria, such as *k*-anonymity, or which would skew aggregation operations too much across several attributes. Outliers can lead to easier re-identification.
How to use it	Delete (or void) the entire record/ row. Suppression must be permanent and not just a 'hide row' function. Similarly, "redaction" is not sufficient if the underlying data remains accessible.
Other tips	• May impact the dataset in terms of statistics such as average and median. • Differs from de-identification, as it removes or voids entire records, whereas de-identification removes selected attributes across all records.

Illustration

In the initial phase, BestBooks notes that each row contains one specific order per customer. Multiple entries per customer require special considerations, as this allows for the inferring of additional information. BestBooks considers that for some attributes such as *Discount%*, the multiple entries per customer could be combined into a single record by using aggregation to average the value (see section on Aggregation below).

After applying such aggregation, record suppression could then be used to remove the other multiple records. However, BestBooks cannot proceed in this way, as that would mean other attributes like *Title* would have to be suppressed as a consequence. This would therefore impact the utility too much. As there are also no obvious outliers in the data, at this stage, record suppression is not applied and the multiple records are kept, but handled separately, e.g., BestBooks opts to aggregate the *Discount%* attribute for similar customers.

Attribute Suppression

Description	Remove (or void) an entire attribute.
When to use it	As part of the data minimisation process, or to remove direct identifiers, or to remove attributes where application of anonymisation techniques affects overall utility too much.
How to use it	Delete (or void) the entire attribute/ column. Suppression must be permanent and not just a 'hide column' function. Similarly, "redaction" is not sufficient if the underlying data remains accessible.
Other tips	• For direct identifiers, attribute suppression is also known as de-identification. • Where data is exported from other sources like a database, it is a good practice not to export any unnecessary attributes from the original source instead of suppressing them later on. • Suppression is more effective than masking the entire attribute values, as simple masking may be prone to reveal implicit information like length of specific values, which can be revealing e.g., to distinguish "Paul" from "Jeremy Chong Wei Jie".

Illustration

BestBooks maintains the original data in an SQL database, but the anonymised data will be passed as an Excel sheet to the analyst. Therefore, BestBooks ensures that the attributes *Name*, *Address*, and *OrderID* are not exported. The only direct identifier exported is the *CustomerID*, which later will be replaced with pseudonyms.

Character Masking

Description	Change values by replacing carefully chosen parts with a typically consistent symbol (e.g., "*" or "x"). Masking is typically applied only to some characters in the attribute and not all.
When to use it	When the value does not need to be interpreted as a whole by the system and hiding / replacing part of it provides the extent of anonymity required.
How to use it	Depending on the nature of the attribute, replace the appropriate characters with a chosen symbol. Depending on the attribute type, replace a fixed number of characters (e.g., for credit card numbers) or a variable number of characters (e.g., for email address).
Other tips	• May need to take into account whether the length of the original data provides information about the original value. • Subject matter knowledge is critical for partial masking to ensure that the right characters are masked. Special consideration may also apply to checksums within the data; sometimes, a checksum may be used to recover (other parts of) the masked data. • Complete masking is similar to value suppression, unless the length of the masked data is of some relevance. • The scenario of masking data in such a way that the individual is meant to recognise their own data is a special case, as it is not an objective of data anonymisation. One example of this is the publishing of lucky draw results, where the names and partially masked national identification numbers of lucky draw winners are published for the individuals to recognise themselves as winners. Generally, anonymised data should not be recognisable even to the data subject themselves.

Illustration

BestBooks aims to optimise its delivery routes by combining deliveries to similar locations and within similar preferred delivery times. As the analyst does not need to know the exact address of each customer, masking to the *Postal Code* in its right-hand side digits is applied, as those determine the specific building. After some discussion, BestBooks decides that for the targeted granularity, the first 3 digits suffice, and applies character masking to the last 3 digits.

After partial masking:

CustomerID	Postal Code
10114	100***
10227	180***
11096	161***
11096	161***
11096	161***
11358	133***
11358	133***
11633	141***
12145	122***
13990	133***

Pseudonymisation

Description	Replace identifying values with made-up values. It is also referred to as coding or tokenisation. Pseudonyms can be: • irreversible when the original values are disposed of properly and the generation of the pseudonyms is random and non-repeatable, or • reversible when the original values are securely kept but can be retrieved and linked back to the pseudonym, should the need arise, or when the generation is not random. Persistent pseudonyms allow linking by using the same pseudonym values to represent the same individual across different datasets. Different pseudonyms may be used to represent the same individual in different datasets to prevent linking of the different datasets.
When to use it	Values need to be uniquely distinguished and no character or any other implied information about the direct identifiers of the original attribute is kept.
How to use it	Replace the value with generated, made-up values. One way to do this is to pre-generate a list of made-up values and randomly select from this list to replace each of the original values. The made-up values should be unique and the original values should not be guessable or computable from the pseudonyms.

Other tips	• When allocating pseudonyms, ensure not to re-use pseudonyms that have already been utilised in the same dataset, especially when they are randomly generated. Also, avoid using the exact same pseudonym generator over several attributes without a change (e.g., at least use a different random seed). • Persistent pseudonyms usually provide better utility by maintaining referential integrity across datasets. • For reversible pseudonyms, the identity mapping table cannot be shared with the recipient; it should be securely kept and can only be used by the organisation where it is necessary to re-identify the individual(s). • If encryption or a hash function is used to pseudonymise a value, the encryption key or hash algorithm and salt value for the hash must be securely protected from unauthorised access. The security of any key used must be ensured like with any other type of encryption or reversible process, and regular review of the method of encryption (e.g., algorithm and key length) and hash function is required. • In some cases, pseudonyms may need to follow the structure or data type of the original value (e.g., for pseudonyms to be usable in software applications). In such cases, special pseudonym generators may be needed to create synthetic datasets or in some cases, so-called "format preserving encryption" can be considered, which creates pseudonyms that have the same format as the original data. • In some cases, it may also be more prudent to change the header/attribute name. • Pseudonyms are usually not included in risk level assessments, as they are intended to be unique, and would, for example, render *k*-anonymity as good as impossible for any $k > 1$. • Sometimes, instead of replacing some direct identifier with pseudonyms, a new column or attribute may be generated for a record-related pseudonym.

Illustration

BestBooks has not opted for aggregation across the multiple records per customer. To avoid easy linkage even by a casual viewer, all direct identifiers are removed, but for the final risk assessment and tests, the DPO prefers to maintain internal linkability.

BestBooks assigns each record a random pseudonym (even to the multiple entries, so that they appear different) and only the DPO has access to the linking table (which identifies the record and customer). Pseudonymisation on record level is done by using a six-character long pseudonym consisting of uppercase alpha-numeric values.

Identity mapping table:

Pseudonym	CustomerID	OrderID
FCH3CO	10114	133620
YAI6YG	10227	141633
TR65O7	11096	105973
XJ8WT4	11096	161096
WMCF3X	11096	122145
MZZMXN	11358	104885
9ZXG5L	11358	138408
BHN6OE	11633	189800
I8B5V1	12145	177613
BXUNOO	13990	181315

Generalisation

Description	Reduce the precision of values. Examples include converting a person's birth date to an age in years, an age into an age range, or a precise location into a less precise location via truncation.
When to use it	Values that can be generalised and still be precise enough for the intended purpose.
How to use it	Design appropriate data categories and rules for generalising data. Consider suppressing any records that still stand out after the generalisation (i.e., age above 99 years).

Other tips	• Generalisation can be achieved by single value replacements (like birth date by age) or by a range (like age by range of 5 years). For ranges:
	  ○ choose an appropriate value range. A value range that is too large may mean significant loss in data utility, while a value range that is too small may mean that the data is hardly modified and therefore, still easy to re-identify. If *k*-anonymity is used, the *k* value chosen may affect the possible data ranges.
	  ○ consider flexible ranges, especially for the first and the last range, as they may permit a larger range to accommodate the typically lower number of records at these ends; this is often referred to as top/bottom coding.

Illustration

BestBooks has categorised the *Preferred Delivery time* attribute as a target attribute. However, due to having multiple records for several customers, this attribute may dilute the obfuscation by the other techniques. Also, for delivery planning, the data analyst does not need the exact information provided by the customers, so BestBooks decides to generalise the *Preferred Delivery time* to 2-hour slots.

After generalisation:

CustomerID	Preferred Delivery time
10114	16:00
10227	20:00
11096	10:00
11096	10:00
11096	10:00
11358	20:00
11358	20:00
11633	14:00
12145	20:00
13990	14:00

Swapping

Description	Rearrange values within individual attributes but across records such that they generally do not remain within the original records. This technique is also referred to as shuffling and permutation. Shuffling can also apply to records to break any sequential information.
When to use it	Subsequent analysis only needs to look at aggregated data. Analysis is at the intra-attribute level; there is no need for analysis of relationships between attributes at the record-level.
How to use it	For each value in the attribute, swap or reassign the values to other records in the dataset.
Other tips	• Assess the need to ensure or verify after swapping that in fact no value ends up in the same position (or replaces an identical value elsewhere). • Consider whether swapping may be limited to certain rows or values, e.g., using rank-based swapping. • Ensure that the swapping order is not reversible and not reproducible.

Illustration

BestBooks assigns each customer a *Membership type*. As customers can be upgraded and downgraded over the years, the Membership type is not a critical identifier for a customer. Nonetheless, BestBooks decides to swap the *Membership type* attribute, because a single customer might have multiple records which all share the same *Membership type*. Also, the team remembers that the multiple records for some customers require them to swap the *Postal Codes*. As a final step, BestBooks shuffles the entire dataset on record bases, so as to ensure multiple records for the same customer do not remain grouped together.

After swapping separately Postal code and Membership:

CustomerID	Postal Code	Membership type
10114	177***	Basic
10227	161***	N/A
11096	177***	Silver
11096	141***	Silver
11096	144***	Silver
11358	104***	Gold
11358	138***	Gold
11633	133***	Platinum
12145	146***	Basic
13990	161***	Silver

Perturbation

Description	Modify the values from the original dataset to be slightly different (typically in a non-systematic way).
When to use it	Values where slight (and random) changes in values are acceptable for the attribute. This technique might not be useful where strict data accuracy is crucial.
How to use it	It depends on the exact data perturbation technique used. These include rounding and adding random noise. The example in this section shows a percentage-based change.
Other tips	• The degree of perturbation should be proportionate to the range of values of the attribute. If the base is too small, the anonymisation effect will be weaker; on the other hand, if the base is too large, the end values will be too different from the original and utility of the dataset will likely be reduced.

	• Where computation is performed on attribute values that have been perturbed before, the resulting computed value may experience perturbation to an even larger extent. • Perturbation can be in fixed ranges or relative to the value. Fixed ranges tend to distort smaller values more, whereas relative values (typically set in percentage points) tend to distort larger values more. • Truncation (instead of rounding) would usually be considered generalisation, as it is a more systematic reduction in precision than a perturbation, which alters the value. • Similar to swapping, assess whether some unchanged values are acceptable. • Perturbation can be an alternative to range-based generalisation when distinct values are required for processing. • Perturbation is typically applied to distinct numeric values but can extend to composite values like IP addresses e.g., by perturbing only the last segment. • When rounding values up or down, perturbation and generalisation often achieve the same outcome; perturbation typically uses random noise or other non-systematic methods.

Illustration

BestBooks collects the *Date of Birth* of its members for promotions. However, for the current scenario, the exact date is not needed; rather, the age group to which a customer belongs is more relevant. BestBooks therefore decides to use only the year of birth, which is a form of generalisation, and converts the year to an age value.

BestBooks further decides that the exact age is not critical and could be used to link multiple records in the dataset. Thus, to counter this risk without real loss of utility, the age is then randomly perturbed by a 25% margin with rounding down. BestBooks also changes the header to Age. At this point, it becomes obvious that some dates might not be accurate, as ages below 10 are not plausible (see the highlighted rows below). Accordingly, the DPO needs to decide whether fully masking the age for those few values is more useful than considering the entire record as outliers or if statistical properties are not critical, change the records to reasonable but fake values.

After generalisation:

CustomerID	Age

After final perturbation:

CustomerID	Age

10114	19
10227	25
11096	48
11096	48
11096	48
11358	13
11358	13
11633	34
12145	14
13990	5

10114	20
10227	20
11096	52
11096	50
11096	36
11358	14
11358	10
11633	40
12145	15
13990	10

Aggregation

Description	Convert values across several records to summarised values. This typically removes the list of individual records, but it can also be used as a form of generalisation to replace values within records.
When to use it	Individual records are not required and aggregated data is sufficient for the purpose.
How to use it	Typical methods include using totals or averages, etc. It may also be also useful to discuss with the data recipient about the expected utility and find a suitable compromise.
Other tips	• Where applicable, watch out for groups having too few records after performing aggregation as it could be easy for someone with some additional knowledge to identify the data subject. • Sometimes aggregation may need to be applied in combination with suppression. Some attributes may need to be removed, as they contain details that cannot be aggregated and new attributes may need to be added (e.g., to contain the newly computed aggregate values). • Aggregation typically creates completely new data structures / datasets.

Illustration

BestBooks had decided early on that removing the multiple records per customer would affect the utility of the data too much. As with other attributes, BestBooks obfuscates the exact *Discount%* value by replacing the respective records of a single customer with the average of the discounts given for that customer. The average value is rounded up/down, otherwise it would be obvious which records have been aggregated. BestBooks notes that some records remain unchanged as the average is the same as the original value but decides

against further perturbation. As with all other observations and decisions, these are recorded in the project and risk assessment documentation.

After local aggregation:

CustomerID	Discount%
10114	9
10227	10
11096	3
11096	3
11096	3
11358	5
11358	5
11633	1
12145	8
13990	7

Summary of the Illustration Outcome

After applying the anonymisation techniques described above, the BestBooks team finds that the resulting dataset maintains a high level of utility. However, the team realises that the target threshold of $k = 3$ has not yet been achieved. While this can be done with the application of further generalisation, there would be a dilution of the information in the dataset below a useful level (the required utility).

As the $k = 3$ threshold cannot be met for the entire dataset with sufficient utility, the BestBooks team considers additional measures that may be applied to achieve a similar outcome. First, the team decides to break the data into smaller, partially overlapping datasets based on the type of data analysis it intends to perform on the respective dataset. This can achieve the required *k*-threshold for the individual and smaller datasets with sufficient utility maintained. In addition, to mitigate the risk that the smaller datasets may be combined, the team implements further internal risk control measures in the form of an internal governance rule to disallow staff and analysts working on one or more of the smaller datasets from combining them (or allowing them to be combined) without appropriate management authorisation.

Appendix 1: Sample Original Data for the Illustration

CustomerID	Name	Postal Code	Preferred Delivery time	Membership type	Education	Date of Birth	Address	Discount%	Order Date	Title	OrderID
10114	Zhe Sy Ming	100114	16:00	Basic	Humanities & Social Sciences	29/5/2005	53 Bishan Circle	9	11/2/2023	To Kill a Mockingbird by Harper Lee	133620
10227	Diyana Eric Yu	180227	20:15	N/A	Engineering Sciences	7/3/1999	87 Clementi Park Road #06-03	10	18/5/2023	1984 by George Orwell	141633
11096	Jingwei Wei Kevin	161096	11:30	Silver	Law	24/7/1976	171 Eastwood Crescent	0	8/4/2023	The Great Gatsby by F. Scott Fitzgerald	105973
11096	Jingwei Wei Kevin	161096	11:30	Silver	Law	24/7/1976	171 Eastwood Crescent	8	24/7/2023	Pride and Prejudice by Jane Austen	161096
11096	Jingwei Wei Kevin	161096	11:30	Silver	Law	24/7/1976	171 Eastwood Crescent	0	10/6/2024	The Catcher in the Rye by J.D. Salinger	122145
11358	Mei Jianwei Lee	133620	21:45	Gold	Business & Administration	11/2/2011	1 Bukit Panjang View	2	29/2/2024	Harry Potter and the Sorcerer's Stone by J.K. Rowling	104885
11358	Mei Jianwei Lee	133620	21:45	Gold	Business & Administration	11/2/2011	1 Bukit Panjang View	7	15/7/2024	The Hobbit by J.R.R. Tolkien	138408
11633	Xinying Chan Tingting	141633	14:30	Platinum	Fine & Applied Arts	18/5/1990	91 Novena Terrace #04-02	1	14/2/2024	The Da Vinci Code by Dan Brown	189800
12145	Lee alia Andrew	122145	20:15	Basic	Business & Administration	24/7/2010	181 Tiong Bahru Drive	8	29/5/2024	The Hunger Games by Suzanne Collins	177613
13990	Linlin Md Mei	133990	15:00	Silver	Natural & Mathematical Sciences	20/1/2019	128 Braddell View	7	11/4/2024	One Hundred Years of Solitude by Gabriel García Márquez	181315
14130	Shufen Wei pllee	144130	10:15	Gold	Business & Administration	11/10/1976	51 Little India Avenue 10	0	20/1/2024	Moby-Dick by Herman Melville	100114
14885	Lis Chen Li	104885	19:15	Basic	Engineering Sciences	10/6/1975	101 Bedok Drive #12-11	5	2/8/2024	The Lord of the Rings by J.R.R. Tolkien	146472
15973	Weiling Ming Jie	105973	19:15	N/A	Humanities & Social Sciences	8/4/2021	40 Tiong Bahru View	6	7/3/2024	The Alchemist by Paulo Coelho	133990
16472	Wei Eric	146472	14:00	Gold	Business & Administration	11/4/1977	32 Macpherson Avenue 7	0	1/8/2024	Brave New World by Aldous Huxley	101358
17613	Lee Yong Wei	177613	15:30	Platinum	Education	15/7/1975	6 Yishun Hill	0	15/1/2024	The Book Thief by Markus Zusak	180227
17613	Lee Yong Wei	177613	15:30	Platinum	Education	15/7/1975	6 Yishun Hill	0	17/1/2024	The Shining by Stephen King	118127

CustomerID	Name	Postal Code	Preferred Delivery time	Membership type	Education	Date of Birth	Address	Discount%	Order Date	Title	OrderID
17846	Han Ming Wu	117846	9:45	Platinum	Engineering Sciences	21/8/1979	31 Golden Mile Hill #11-02	1	10/6/2024	Jane Eyre by Charlotte Brontë	118589
18127	Junjie Wei Peng	118127	17:45	Basic	Business & Administration	1/8/2022	117 Choa Chu Kang Crescent	6	13/8/2024	The Road by Cormac McCarthy	117846
18408	Wei Wei	138408	18:00	Silver	Health Sciences	24/11/2018	63 Dairy Farm View	7	3/3/2024	Little Women by Louisa May Alcott	144130
19800	Han Ming Chun	189800	20:45	Basic	Business & Administration	19/11/1982	94 Springleaf Lane	4	16/2/2024	Gone with the Wind by Margaret Mitchell	181615

Appendix 2: After Application of Anonymisation Techniques

The following tables summarise the anonymisation techniques applied to the dataset and the data after anonymisation, but before the final record-based swapping. The second table contains the *CustomerID* only for information; that attribute would not be included in the data for the analyst.

Category of Attribute	Direct	Direct	Indirect	Target	Target	Target	Direct	Target	Target	Target	Direct
Attribute Type	Customer ID	Name	Postal Code	Preferred Delivery time	Membership type	Education	Address	Discount%	Order Date	Title	OrderID
Anonymisation Technique	RandomPseudo	Suppress	Mask and swap	Generalise to 2 hour	Swap	*Not modified*	Remove	Aggregation average	*Not modified*	*Not modified*	Suppress

Pseudonym	**CustomerID**	**Postal Code**	**Preferred Delivery time**	**Membership type**	**Education**	**Age**	**Discount%**	**Order Date**	**Title**
FCH3C0	10114	177***	16:00	Basic	Humanities & Social Sciences	20	9	11/2/2023	To Kill a Mockingbird by Harper Lee
YAI6YG	10227	161***	20:00	N/A	Engineering Sciences	20	10	18/5/2023	1984 by George Orwell
TR6507	11096	177***	10:00	Silver	Law	52	3	8/4/2023	The Great Gatsby by F. Scott Fitzgerald
XJ8WT4	11096	141***	10:00	Silver	Law	50	3	24/7/2023	Pride and Prejudice by Jane Austen
WMCF3X	11096	144***	10:00	Silver	Law	36	3	10/6/2024	The Catcher in the Rye by J.D. Salinger
MZZMXN	11358	104***	20:00	Gold	Business & Administration	14	5	29/2/2024	Harry Potter and the Sorcerer's Stone by J.K. Rowling
9ZXG5L	11358	138***	20:00	Gold	Business & Administration	10	5	15/7/2024	The Hobbit by J.R.R. Tolkien
BHN60E	11633	133***	14:00	Platinum	Fine & Applied Arts	40	1	14/2/2024	The Da Vinci Code by Dan Brown
I8B5V1	12145	146***	20:00	Basic	Business & Administration	15	8	29/5/2024	The Hunger Games by Suzanne Collins

Pseudonym	CustomerID	Postal Code	Preferred Delivery time	Membership type	Education	Age	Discount%	Order Date	Title
BXUN0O	13990	161***	14:00	Silver	Natural & Mathematical Sciences	6	7	11/4/2024	One Hundred Years of Solitude by Gabriel García Márquez
O5D1KF	14130	117***	10:00	Gold	Business & Administration	46	0	20/1/2024	Moby-Dick by Herman Melville
NZZFOQ	14885	105***	18:00	Basic	Engineering Sciences	37	5	2/8/2024	The Lord of the Rings by J.R.R. Tolkien
37MKXC	15973	133***	18:00	N/A	Humanities & Social Sciences	2	6	7/3/2024	The Alchemist by Paulo Coelho
T7VMOH	16472	122***	14:00	Gold	Business & Administration	49	0	1/8/2024	Brave New World by Aldous Huxley
1MD1XP	17613	189***	14:00	Platinum	Education	55	0	15/1/2024	The Book Thief by Markus Zusak
VYIE42	17613	100***	14:00	Platinum	Education	54	0	17/1/2024	The Shining by Stephen King
YB95J0	17846	118***	08:00	Platinum	Engineering Sciences	39	1	10/6/2024	Jane Eyre by Charlotte Brontë
5C02VH	18127	161***	16:00	Basic	Business & Administration	1	6	13/8/2024	The Road by Cormac McCarthy
D48H0P	18408	180***	18:00	Silver	Health Sciences	6	7	3/3/2024	Little Women by Louisa May Alcott
4YC5IM	19800	189***	20:00	Basic	Business & Administration	41	4	16/2/2024	Gone with the Wind by Margaret Mitchell

ANNEX B: An Overview on *K*-anonymity, *L*-diversity and *T*-closeness

K-anonymity

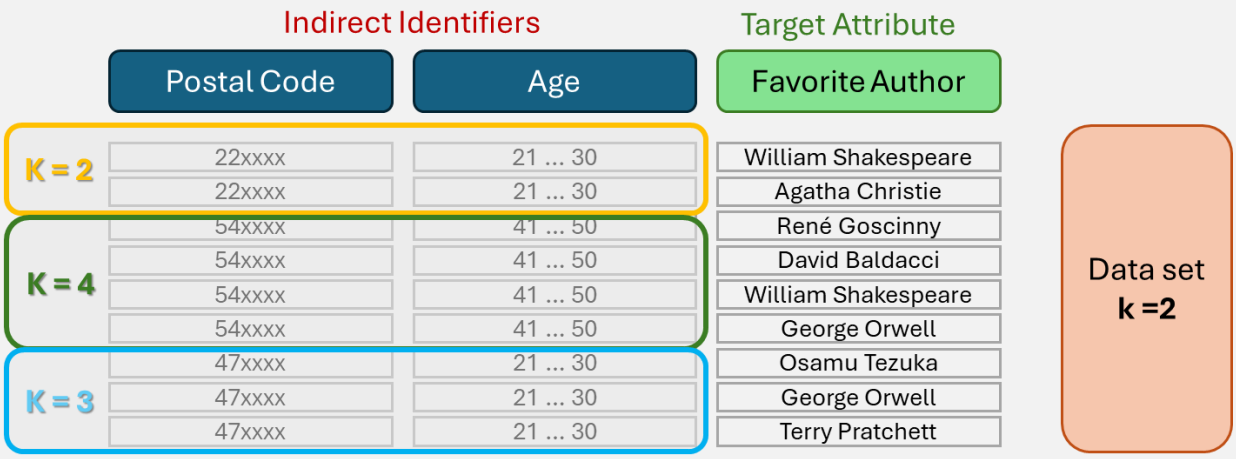
K-anonymity, when taken as a measure, is a simple, efficiently calculated and objective number. In essence, it is a simple counting mechanism, which is agnostic of the actual content of the data, be it before, during, or after anonymisation. *K*-anonymity has two scopes: each row will have a certain *k*-value (or *k*-anonymity), but only the lowest of all *k*-values across all rows will establish the final *k*-anonymity of the entire dataset.

To calculate the current *k*-anonymity of a dataset, the indirect identifiers of each row are considered as a single, ordered unit. The number of all rows in the anonymised dataset that have exactly the same unit establishes the *k*-value for that unit (and thus for all those rows/records). The group of records sharing the same unit is often called an equivalence-group or equivalence class. The *k*-anonymity of the entire dataset, however, is a single number, which is the minimum of all these *k*-values.

As a simple illustration, one can think of *k*-anonymity as 'hiding in the crowd'. Assuming the only facts known to the police chasing a thief is that the thief was male and had a beard. The thief then managed to escape into a large crowd. The more bearded males that exist in the crowd, the harder and less likely it will be for the police to *reliably* identify, single-out, and point out the thief within the crowd. Other factors may still give the thief away, but those are factors not considered relevant in the counting for *k*-values.

K-anonymity focuses exclusively on indirect identifiers, based on the assumption that all direct identifiers have been removed (or possibly pseudonymised) and that target attributes are essentially non-identifying enough to pose any serious risk for re-identification attempts. However, these assumptions have subsequently been shown to be less reliable in the increasingly digitalised world than during the time of *k*-anonymity's creation, and so other complementary measures like *l*-diversity and *t*-closeness have been created for target attributes. Nevertheless, *k*-anonymity remains a simple and easy to understand measure that can be used to support basic anonymisation, together with other risk mitigation controls.

For example, in the image below, the final *k*-anonymity for the dataset equals to 2, as that is the size of the smallest group of identical units, even though the majority of the individual records have a different *k*-value, namely $k = 3$ and $k = 4$. Also note that *Age* taken by itself would achieve $k = 4$, while *Postal Code* taken by itself remains at $k = 2$.



A higher *k*-value indicates a lower risk, but typically the utility may become lower and more records may need to be suppressed. A *k*-anonymity of *k* = 1 means that (at least) one record is unique in its 'unit' of indirect identifiers. Given the underlying definition of indirect identifiers (potentially being able to identify an individual when taken together), such a unique unit is basically equivalent to being a direct identifier. Accordingly, the dataset will be susceptible to singling out attacks, even if anonymisation techniques have been applied.

Due to the simple counting approach, certain datasets may sometimes automatically display groups of records with identical units of indirect identifiers, and thus even without applying anonymisation techniques, those groups already start with a *k*-value above 1. This effect largely depends on the data distribution within the attributes (e.g., possible combinations across the unit). As most techniques are typically applied on the entire attribute and not only on individual or smaller groups of rows, the original data will often still change from its original value (and then may or may not fall into a larger *k*-value group).

Nevertheless, it is still prudent to check that changes have indeed been made to all indirect identifiers, because even if a group of them 'technically' fulfils the anonymity associated with the *k*-threshold (and thus singling out and other attacks may not seem to work), there is still a risk of unauthorised disclosure of 'real' personal data.

It is important to reiterate that *k*-anonymity is primarily a measure of resistance against linking attacks based on indirect identifiers to find a specific record. Once a dataset has been found to have at least *k* = 2, it means an attacker cannot establish with certainty which of those 2 records match the linking attack, and that applies for all records. However, the attacker still has a 50% chance of 'picking' the correct one and that is generally considered too high a risk for the purposes of data protection, even though anonymisation techniques may have been applied.

When *k*-anonymity is taken as a threshold, it is the outcome of a risk assessment process and to be established *before* applying anonymisation techniques. The details

of that risk assessment and how the risk and impact factors are ultimately mapped to a specific *k*-threshold value is outside the scope of this Guide. Note, however, that unless specified by regulators, this mapping also depends on the release model and the distributions of the attributes.

While *k*-anonymity is a means to ensure that the anonymisation techniques applied achieve the desired threshold against linking and singling-out attacks, other re-identification and inference attacks may still be possible and require additional anonymisation of target attributes. To counter those attacks, extensions to *k*-anonymity such as *l*-diversity and *t*-closeness may be considered.

L-diversity and T-closeness

As it is beyond the scope of this Guide to provide an in-depth discussion on *l*-diversity and *t*-closeness, these concepts are only covered briefly below[12].

L-diversity extends *k*-anonymity by taking the statistical distribution of the values within a target attribute into consideration. *L*-diversity is achieved for an individual equivalence class (see above for an explanation of equivalence classes) if there are at least "*l*" "well-represented" values for a target attribute within the class. There are different ways to establish that a target attribute is well-represented; the simplest method is to ensure that there are at least "*l*" *distinct* values in each class. A dataset, which has achieved *k*-anonymity, is also said to have *l*-diversity if every equivalence class in the dataset has *l*-diversity. For *l*-diversity across multiple target attributes, additional considerations are required as even if each attribute may be *l*-diverse in itself, when they are considered in combination, they may not remain well-represented as a group anymore.

To illustrate, reference may be taken from the diagram above in relation to *k*-anonymity. It is both 2-anonymous and also 2-diverse as each equivalence class contains at least 2 different values in the target attribute. If any of the 3 classes in the diagram had less than 2 different favourite authors, the dataset would no longer be 2-diverse. On the other hand, the dataset would remain 2-diverse even if in the $k = 3$ class, the second record contained "Terry Pratchett", instead of "George Orwell", because that class still has 2 different values.

The key attacks addressed by *l*-diversity are homogeneity and background knowledge attacks. While *k*-anonymity reduces the ability of an attacker to match a record to an

[12] For more information, see for example, A. Machanavajjhala, D. Kifer, J. Gehrke and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity". ACM Trans. Knowl. Discov. Data 1, 1 (March 2007), 3–es, https://doi.org/10.1145/1217299.1217302; and N. Li, T. Li and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity". 2007 IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey, 2007, pp. 106-115, doi: 10.1109/ICDE.2007.367856.

individual, the dataset remains vulnerable to such attacks which may disclose sensitive information or narrow down the attacker's search for the respective equivalence class.

In essence, *l*-diversity requires sufficient variation which limits the occurrence of identical values within equivalence classes (i.e., it requires target attributes to be well-represented within each equivalence class), or more technically, *l*-diversity reduces / limits the risk of attribute disclosure. In contrast, *k*-anonymity reduces the basic risk of re-identification (i.e., identity disclosure).

T-closeness is a further extension and refinement of *l*-diversity. Even though the target attribute within an equivalence class may appear well-represented, the values may still be skewed within equivalence classes when compared to the entire dataset (or by extension, the entire population). *T*-closeness checks that for each class, the statistical distribution of the values (of the target attributes) remain close (within a certain percentage, provided by the *t*-value) to the statistical distribution (of the target attributes) across all records.

T-closeness ensures that the target attributes (which may be sensitive attributes) are distributed similarly within groups and the overall population, making it difficult for attackers to identify the target attributes of individuals or re-identify individuals through group membership or background knowledge. Similar to *l*-diversity, a dataset which has achieved *k*-anonymity is also said to have *t*-closeness only if each equivalence class achieves *t*-closeness.

ANNEX C: Common Misunderstandings in Anonymisation

Insufficient obfuscation of alumni record

A university collects personal data from its alumni to provide career related statistics on its website. Peter is an alumnus of the university and the university's alumni database contains the following record of him with his respective attributes:

Name	Gender	Date of Birth	Occupation / Company	Date Graduated
Peter Lee	Male	1 July 1997	Privacy Engineer / DPIA Firm	1 August 2022

Peter decided to withdraw his consent for data to be collected and processed by the university. The university has determined that it should cease to retain Peter's full personal data in its alumni database in compliance with its retention limitation obligation. Instead of deleting the entire record, the university's DPO decided to anonymise this record so that the data can be used for internal analysis and research purposes. As full name is the only attribute explicitly classified as personal data in the university's data protection policy, the DPO deleted the name "Peter Lee" from the record, which is changed to:

Name	Gender	Date of Birth	Occupation / Company	Date Graduated
Null	Male	1 July 1997	Privacy Engineer / DPIA Firm	1 August 2022

However, Peter's name may still be 'reconstructed' from the modified record by combining existing data with other data that may be easily available from web searches, such as Peter's personal and business social media accounts. As such, additional anonymisation techniques need to be applied to the other attributes in the record, even though the data is not shared with any external party.

The above example shows how someone lacking in experience or familiarity with anonymisation and the various attribute types, may not anonymise data sufficiently. Even organisations which are more familiar with the intricacies and pitfalls of anonymisation can still face issues when a more motivated attacker analyses and 'de-anonymises' the data with a new approach and/or known attacks. An often-quoted example is the Netflix case: In 2006, Netflix published an anonymised dataset comprising several million movie rankings of about 500,000 customers. The intent was to crowdsource for a better recommendation system, and so the anonymised dataset was publicly released. The issue was not that the data itself was simply reverse engineered. Instead, it was possible to cross-reference the data for certain users, who

also entered movie rankings using the same identifier in another ranking system, in this case the Internet Movie Database. This meant that some users in the Netflix dataset were compromised.

It is not expected that the entire set of individuals' personal data would be compromised in such attacks. There is no unanimity nor any simple way to define 'how many' such records or individual identities would have to be compromised, or how 'easy' a certain compromise was to achieve.

ANNEX D: Anonymisation Tools

Software tools (both commercial and open-source) can be employed to assist in implementing anonymisation techniques, but they should not be used without understanding how they work. Some anonymisation tools that are available in the market are listed in the table below. The list below is neither a recommendation nor endorsement by ASEAN members. Organisations should exercise due diligence and ensure that the appropriate tools are used for their respective purposes.

Tool name	Description
PDPC Singapore Anonymisation tool	A free basic data anonymisation tool to transform simple datasets by applying anonymisation techniques in Excel sheets. (for information) https://www.pdpc.gov.sg/help-and-resources/2018/01/basic-anonymisation (for downloading the tool, fill in the required form) https://form.gov.sg/62981e766cf13d001200f4bc
Amnesia	Amnesia anonymisation tool is a software used locally to anonymise personal and sensitive data. It currently supports *k*-anonymity and *km*-anonymity guarantees. https://amnesia.openaire.eu/
Arcad DOT-Anonymizer	DOT-Anonymizer is a tool that maintains the confidentiality of test data by concealing personal information. It works by anonymising personal data while preserving its format and type. https://www.arcadsoftware.com/dot/data-masking/dot-anonymizer/
ARGUS	ARGUS stands for "Anti Re-identification General Utility System". The tool uses a wide range of different statistical anonymisation methods such as global recoding (grouping of categories), local suppression, randomisation, adding noise, microaggregation, top- and bottom coding. It can also be used to generate synthetic data. https://research.cbs.nl/casc/mu.htm

Tool name	Description
ARX	ARX is an open-source software for anonymising sensitive personal data. https://arx.deidentifier.org
Eclipse	Eclipse is a suite of tools from Privacy Analytics that facilitates anonymisation of health data. https://privacy-analytics.com/eclipse-software/
sdcMicro	sdcMicro is used to generate anonymised microdata such as public and scientific use files. It supports different risk estimation methods. https://cran.r-project.org/web/packages/sdcMicro/index.html
UTD Anonymisation Toolbox	UT Dallas Data Security and Privacy Lab compiled various anonymisation techniques into a toolbox for public use. https://labs.utdallas.edu/dspl/software/anonymization-toolbox/