GUIDE TO BASIC DATA ANONYMISATION TECHNIQUES

Published 25 January 2018

TABLE OF CONTENTS

PART 1: OVERVIEW

1 Introduction

1.1. The collection, use and disclosure of individuals' personal data by organisations in Singapore is governed by the Personal Data Protection Act 2012 (the "**PDPA**"). The Personal Data Protection Commission ("**PDPC**") was established to enforce the PDPA and promote awareness of protection of personal data in Singapore.

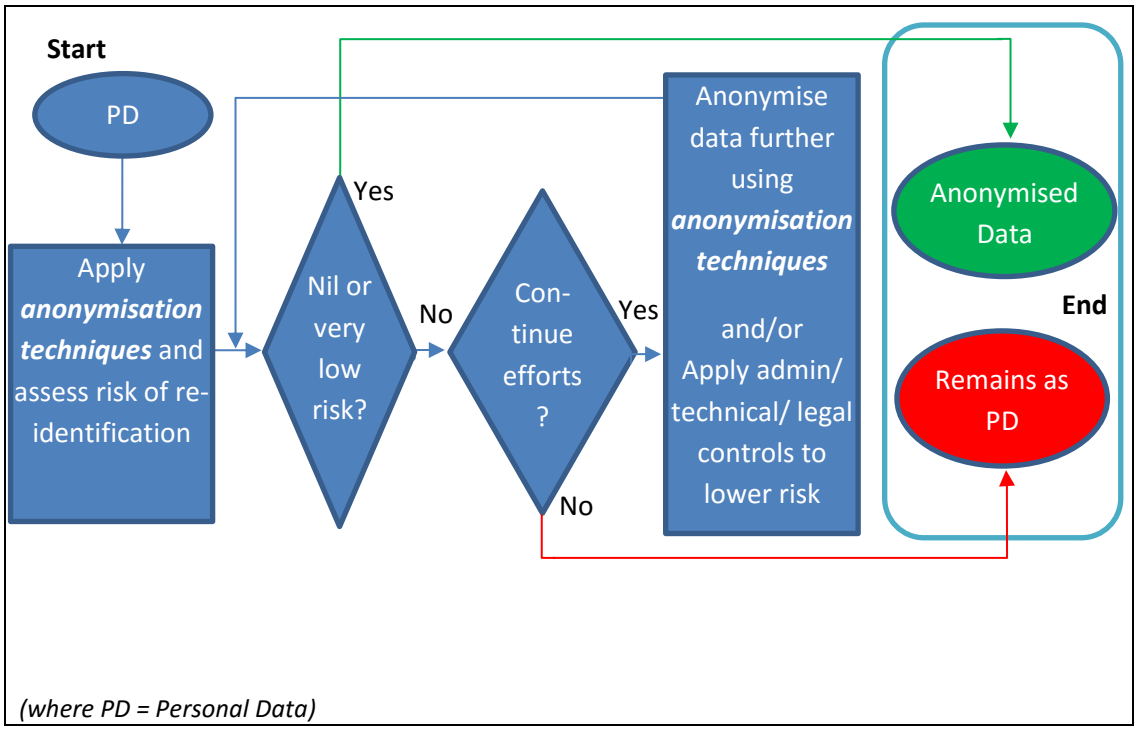
2 Purpose and Scope of This Guide

2.1. This Guide seeks to provide a general introduction to the technical aspects of anonymisation[1]. It should be read together with Chapter 3 (Anonymisation) of the PDPC's Advisory Guidelines on the PDPA for Selected Topics ("Advisory Guidelines"), which sets out PDPC's interpretation and considerations for determining what constitutes "anonymisation" under the PDPA.

2.2. The basic concepts and techniques discussed in this Guide make reference to the terms "data anonymisation", and "anonymised data". "Data anonymisation" refers to the conversion of personal data into "anonymised data" by applying a range of "anonymisation techniques". "Anonymised data", for the purposes of this Guide, refers to data that has undergone transformation by anonymisation techniques in combination with assessment of the risk of re-identification. Typically, the process of data anonymisation would be "irreversible" and the recipient of the anonymised dataset would not be able to recreate the original data. However, there may be cases where the organisation applying the anonymisation retains the ability to recreate the original data from the anonymised data; in such cases, the anonymisation process is "reversible".

2.3. In this Guide, the terms "data anonymisation" and "anonymised data" are intended to be understood generically and aligned to the technical literature on this topic. They are not intended to be understood in the same way as the terms used in the Advisory Guidelines, nor give determinative legal effect to the data that has undergone transformation by anonymisation techniques. The following diagram provides a pictorial summary of the data anonymisation concept in the Advisory Guidelines:

[1] To avoid misunderstanding, anonymisation in this Guide refers to the transformation of existing data already available to an Organisation. It does not refer to the aspect of "anonymity" of individuals, where individuals attempt to hide their identity from being known.



For more information on the PDPC's interpretation of "anonymisation" and "anonymised data", please refer to the Advisory Guidelines.

2.4. The intent of this Guide is to provide information on techniques that could be applied in anonymising data. This Guide primarily addresses organisations which do not intend to release the anonymised data into the public domain, but who share data with other organisations or entities, where additional administrative and technical controls may be imposed to reduce the risk of unauthorised disclosure of personal data. Application of these techniques may not necessarily ensure that the data does not pose any serious risk of re-identification and therefore constitutes "anonymised data" to which the PDPA does not apply.

2.5. This Guide is not a substitute for professional training, literature and services. Unless Organisations are familiar with the risks and countermeasures, it is recommended for Organisations, when disclosing anonymised data – especially if the disclosure is intended for release into the public domain or the release involves multiple datasets or updates of anonymised data over time – to seek professional advice or services for data anonymisation.

2.6. This Guide describes anonymisation techniques for static, structured, well-defined, textual, and single-level datasets, whereby:

- "Static" refers to the fact that the data is fully available at the time of anonymisation; this is in contrast to streaming data, where relationships between data may not be fully established because streaming constantly

provides new data. Hence, streaming data may need other anonymisation techniques than those discussed in this Guide.

- "Structured" refers to the fact that the anonymisation technique is applied to data within a known format and a known location within the data pool. "Structured" is therefore not limited to data in a tabular format like in a spreadsheet or a relational database, but may be held or released in other defined formats, for example XML, CSV, JSON, etc. This Guide describes the techniques and provides examples in the more common tabular format, but this does not imply that the techniques only apply to tabular format.

- "Well-defined" refers to the fact that the original dataset conforms to pre-defined rules. E.g. data from relational databases tend to be more well-defined. Anonymising datasets which are not well-defined may create additional challenges to data anonymisation, and is outside the scope of this Guide.

- "Textual" refers to text, numbers, dates, etc., that is, alphanumeric data already in digital form. Anonymisation techniques for streaming data like audio, video, images, big data (in its raw form), geolocation, bio-metrics etc. create additional challenges and require entirely different anonymisation techniques, which are outside the scope of this Guide.

- "Single-level" refers to data pertaining to different individuals. Datasets which contain multiple entries for the same individuals (e.g. different transactions done by an individual) may still use some of the techniques explained in this Guide, but additional criteria may need to be applied; such criteria are outside the scope of this Guide.

2.7. This Guide is for persons who are responsible for data protection within an organisation, without prior knowledge or experience in data anonymisation. A basic mathematical background will be required to understand some of the terminology and concepts used, and a basic understanding of risk management is needed in the application of the techniques.

2.8. While this Guide seeks to assist organisations in anonymising personal data, the Commission recognises that there is no "one size fits all" solution for organisations. Each organisation should therefore utilise anonymisation approaches that are appropriate for their circumstances. Some factors that organisations can take into account when deciding on the anonymisation technique(s) to use include:

- the nature and type of personal data that the organisation intends to anonymise, as different anonymisation techniques are suitable for different types of data and circumstances;

- risk management by the organisation to impose controls to protect the anonymised data, in addition to the anonymisation techniques;
- the utility required from the anonymised data (refer to section 4 on anonymisation concepts).

3 Terminology

3.1. Due to the variance of terms and meanings used in literature on the subject of data anonymisation, this section explains the meaning of some key terms as they are used in *this* Guide.

Term	Meaning in this Guide
Adversary	A party which attempts to re-identify individual(s) from a dataset that is supposed to be anonymised.
Anonymisation	The conversion of personal data into "anonymised data" by applying a range of anonymisation techniques. (This guide focusses only on the technical aspects of this conversion)
Anonymised dataset	The resultant dataset after anonymisation technique(s) has/have been applied in combination with adequate risk assessment.
Attribute	Also referred to as data field, data column or variable. An information that can be found across the data records in a dataset. Name, gender and address are examples of attributes.
Dataset	A set of data records. Conceptually similar to a table in a typical relational database or spreadsheet, having records (rows) and attributes (columns).
Direct identifier	A data attribute that on its own identifies an individual (e.g. fingerprint) or has been assigned to an individual. (e.g. NRIC number)
Equivalence class	The records in a dataset that share the same values within certain attributes, typically indirect identifiers.
Identifiability vs Re-identifiability	The degree to which an individual can be identified from one or more datasets containing direct and indirect identifiers, versus the degree to which an individual can be identified from anonymised dataset(s).
Indirect identifier	Also referred to as quasi-identifiers. A data attribute that, by itself/on its own, does not identify an individual, but may identify an individual when combined with other information.
Non-identifier	Datasets may contain data attributes which are neither categorised as direct nor indirect identifiers. Such attributes need not undergo anonymisation (Note that

	the examples provided in this guide do not include such attributes, but this does not mean they cannot be part of the anonymised data)
Original dataset	The dataset before any anonymisation technique is applied.
Pseudonymisation[2]	The technique of replacing an identifier with an unrelated yet typically still unique value. E.g. Replacing "Joshua Quek" with "274927473"
Record	Also referred to as a row. A group of information typically relating to a subject (e.g. an individual) or transaction.
Re-identification	Identifying a person from an anonymised dataset. Spontaneous re-identification refers to unintended re-identification due to having special knowledge of individuals.

Additional notes on terminology

3.2. Chapter 5 of the "Advisory Guidelines On Key Concepts in the PDPA" clarifies what "identifiers" are. The Guidelines use the term "unique identifier", which is equivalent to the term "direct identifier" used in this Guide. The term "direct identifier" is used instead of "unique identifier" in this Guide as the former is more commonly used in the area of data anonymisation.

3.3. The Advisory Guidelines do not provide a specific term equivalent to "indirect identifier", but explain based on an example that "although each of these data points, on its own, would not be able to identify an individual", the organisation "should be mindful that the dataset" (e.g. data points in combination) "may be able to identify the respondent". It also clarifies that "so long as any combination of data contains a unique identifier of an individual, that combination of data will constitute personal data".

3.4. Note also that there is no common term in typical anonymisation literature to describe the third type of data, referred to in this Guide as 'non-identifiers'. Such non-identifiers would not be considered Personal Data, if they were isolated from any direct and indirect identifiers (e.g. not all data is necessarily Personal Data). But once they are linked to direct or indirect identifiers, they need to be protected and treated just like Personal Data. As long as the use or appearance of such data within an anonymised

[2] Some literature (e.g. "Opinion 05/2014 on Anonymisation Techniques" by the Article 29 Data Protection Working Party) emphasise the risk of using pseudonyms as an anonymisation technique. In this Guide pseudonymisation is not excluded from the anonymisation techniques, because it may still serve its purpose when applied diligently.

dataset does not violate any of the other PDPA obligations, they need not be further anonymised, as they would not be able to identify an individual.[3]

3.5. It is not the intent of this Guide to define which of the three types are Personal Data under the PDPA and which are not, but for the purpose of discussing anonymisation techniques, this additional distinction is important, and therefore this Guide follows the common terminology in anonymisation literature using "direct" and "indirect" identifiers, and where "data points" are termed as data fields or attributes.

3.6. Similarly, it should be noted that this Guide does not differentiate between "data" and "metadata"; the techniques can (and where needed, should) be applied to metadata and any other type of data as well. However, the anonymisation of a specific kind of meta-data within the dataset itself, namely column header names in spreadsheets or tags in XML files, is not discussed, as only a few techniques would apply to address this type of data.

PART 2: BACKGROUND

4 Data Anonymisation Concepts

4.1. Data anonymisation requires a good understanding of the following elements, which should be taken into consideration when determining suitable anonymisation techniques and an appropriate anonymisation level:

a. **Purpose of anonymisation and utility**: The purpose of the anonymisation should be clear, because anonymisation should be done specifically to the purpose on hand. The process of anonymisation, regardless of the techniques used, reduces the original information in the dataset by some extent. And hence, generally, as the extent of anonymisation increases, the utility (e.g. clarity and/or precision) of the dataset reduces. Hence the organisation needs to decide on the degree of the trade-off between acceptable (or expected) utility and trying to reduce the risk of re-

[3] For example: A car dealer, for the purpose of utilising Artificial Intelligence and Machine Learning, has very detailed customer records, e.g. down to the colour of the car purchased and the year of the tyre production. The car producer wants to determine which default car colour should be produced in more quantity. After anonymising (e.g. suppressing) the direct and indirect identifiers, the car dealer can share the resulting data (e.g. containing purchaser's gender, car colour, tyre production date, etc.) with the car producer without the need to apply further anonymisation techniques (e.g. no need to generalise the tyre production date or *k*-anonymise the data set). However, the car dealer can only proceed in this manner when, among others, it is established that: a) the remaining data *in general* does not constitute direct or indirect identifiers, b) the use and sharing of this raw data is not against any of the other obligations (e.g. consent), c) individual, specific records (e.g. custom produced unique car colour) are removed.

identification - where the data subject is identified from data that is supposed to be anonymised.

It should be noted that utility should not be measured on the level of the entire dataset, but is typically different for different attributes; one extreme is that a specific attribute is the main item of interest and no generalisation/anonymisation technique should be applied (e.g. due to data accuracy being crucial), whereas the other extreme could be that a certain attribute is of no use for the intended purpose and may be dropped entirely without affecting the utility of the data to the recipient.

Another important consideration in terms of utility is whether it poses an additional risk if the recipient knows which anonymisation technique and what degree of granularity have been applied; on the one side it might help the analyst to understand the results better or interpret them better, but on the other side it might contain hints which could lead to a higher risk of re-identification (however, some outcomes simply cannot hide their granularity, e.g. *k*-anonymity).

b. **Characteristics of anonymisation techniques**: The different characteristics of the various anonymisation techniques mean that certain techniques may be more suitable for a situation than others. For instance, certain techniques (e.g. character masking) are usually used on direct identifiers and others (e.g. aggregation) for indirect identifiers. Another example is to consider if the attribute value is a continuous value or discrete (e.g. “yes” or “no”) value, because techniques like data perturbation work much better for continuous values.

   The various anonymisation techniques also modify data in significantly different ways. Some modify only parts of an attribute (e.g. character masking); some replace the value of an attribute across multiple records (e.g. aggregation); some replace the entire attribute with unrelated, but consistent information (e.g. pseudonymisation); and some remove the attribute entirely (e.g. attribute suppression).

   Some anonymisation techniques can be used in combination. E.g. suppressing or removing (outlier) records after generalisation is done.

c. **Inferred information**: It may be possible for certain information to be inferred from anonymised data. E.g. masking may hide personal data, but it does not hide the length of the original data in terms of the number of characters.

   The problem of inference is not limited to a single attribute, but may also apply across attributes, even if all have had anonymisation techniques applied. The anonymisation process must therefore take note of every possibility, both before deciding on the actual techniques and after applying the techniques.

The approach may also want to consider in which order the anonymised data is presented: if the recipient knows that the data records were collected in serial order (e.g. visitors as they come), it might be prudent (as long as it does not affect the utility) to reshuffle the entire dataset to avoid inference based on the order of the data records.

d. **Expertise with the subject matter**: Anonymisation techniques basically reduce the "identifiability" of one or more individuals from the original dataset to a level acceptable by the organisation's risk portfolio.

   An "identifiability" assessment should be performed before and after anonymisation techniques are applied, and this requires a good understanding of the subject matter which the data pertains to. The assessment before the anonymisation process ensures that the structure and information within an attribute is clearly identified and understood, and the risk of explicit and implicit inference from such data is assessed; e.g. an attribute containing the year of birth implicitly provides age, to some extent similar to an NRIC number. The assessment after the anonymisation process will determine the residual risk of re-identification. Hence, if the dataset is healthcare data, it likely requires someone with sufficient healthcare knowledge to assess how unique (i.e. how identifiable) a record is.

   Another example is where a synthetic dataset is created or data attributes are swapped between records, it takes a subject matter expert to recognise if the anonymised records even make sense.

   The right choice of anonymisation techniques therefore depends on the awareness of the explicit and implicit information contained in the dataset and the amount or type of information intended to be anonymised.

e. **Competency in anonymisation process and techniques**: Anonymisation is complex. Besides having subject matter expertise (as explained above), Organisations wishing to share anonymised datasets should also ensure that the anonymisation process is undertaken by persons well-versed in anonymisation techniques and principles. If the necessary expertise is not found within the Organisation, external help should be engaged.

f. **The recipient**: Factors such as the recipients' expertise with the subject matter, controls implemented to limit the recipients and to prevent the data from being shared with unauthorised parties play an important role in the choice of the anonymisation techniques. In particular, the expected use of the anonymised data by the recipient may impose limitations on the applied techniques, because the utility of the data may be lost beyond acceptable limits. Extreme caution need to be taken

when making public releases of data, and will require a much stronger form of anonymisation compared to data shared under a contractual arrangement.

g. **Tools**: Due to the complexity and computation required, software tools can be very useful to aid in executing anonymisation techniques. There are some dedicated tools[4] available, but this Guide does not provide any assessment nor recommendation of anonymisation or re-identification assessment tools. Note that even the best tools will need adequate inputs (e.g. appropriate parameters to be used), or may contain limitations, hence human oversight and familiarity with the tools and data, are still required.

5 Disclosure risks

5.1. There are various types of disclosure risks. This section explains some fundamental ones to facilitate further discussion on data anonymisation.

- Identity disclosure (re-identification): determining, with a high level of confidence, the identity of an individual described by a specific record. This could arise from scenarios such as insufficient anonymisation, re-identification by linking, or pseudonym reversal. E.g. an anonymisation process which creates pseudonyms based on as easily guessable and reversible algorithm, such as replacing '1' with 'a', '2' with 'b', and so on.
- Attribute disclosure: determining, with a high level of confidence, that an attribute described in the dataset belongs to a specific individual, even if the individual's record cannot be distinguished. E.g. a dataset containing anonymised client records of a particular aesthetic surgeon reveals that all his clients below the age of 30 have undergone a particular procedure. If it is known that a particular individual is 28 years old and is a client of this surgeon, we then know that this individual has undergone the particular procedure, even if the individual's record cannot be distinguished from others in the anonymised dataset.
- Inference disclosure: making an inference, with a high level of confidence, about an individual even if he/she is not in the dataset, by statistical properties of the dataset. E.g. if a dataset released by a medical researcher reveals that 70% of individual aged above the age of 75 have a certain medical condition, this information could be inferred on an individual who is not even in the dataset.

5.2. In general, most traditional anonymisation techniques aim to protect against identify disclosure and not necessarily other types of disclosure risks.

[4] Anonymisation tools include ARGUS, sdcMicro, ARX, Privacy Analytics Eclipse, Arcad DOT-Anonymizer

PART 3: BASIC DATA ANONYMISATION TECHNIQUES

6 Attribute Suppression

6.1. **Description**: Attribute suppression refers to the removal of an entire part of data (also referred to as "column" in databases and spreadsheets) in a dataset.

6.2. **When to use it**: When an attribute is not required in the anonymised dataset, or when the attribute cannot otherwise be suitably anonymised with another technique. This technique should be applied at the start of the anonymisation process, as it is an easy way to decrease identifiability at this point.

6.3. **How to use it**: Delete (e.g. remove) the attribute(s), or if the structure of the dataset needs to be maintained, clear the data (and possibly the header). Note that the suppression should be actual removal (i.e. permanent) , and not just '"hiding the column"[5]. Similarly, 'redacting' may not be sufficient if the underlying data remains somewhat accessible.

Other tips:

6.4. This is the strongest type of anonymisation technique, because there is no way of recovering any information from such an attribute.

6.5. In certain scenarios, it may be possible to create a "derived attribute" that provides utility and yet is less sensitive than the original attribute(s) which can thus be suppressed. E.g. to create a "duration in premise" attribute based on the "date & time of entry" and "date and time of exit" attributes.

6.6. Example

In this example, the dataset consists of test scores. As the recipient only needs to analyse test scores obtained by students with respect to their various trainers but without analysis on the students themselves, the "student" attribute was removed.

Before anonymisation:

Student	Trainer	Test Score
John	Tina	87
Yong	Tina	56
Ming	Tina	92
Poh	Huang	83
Linnie	Huang	45
Jake	Huang	67

[5] Found in spreadsheet software

After suppressing the "student" attribute:

Trainer	Test Score
Tina	87
Tina	56
Tina	92
Huang	83
Huang	45
Huang	67

7 Record Suppression

7.1. **Description**: Record suppression refers to the removal of an entire record in a dataset. In contrast to most other techniques, this technique affects multiple attributes at the same time.

7.2. **When to use it**: To remove outlier records which are unique or do not meet other criteria such as *k*-anonymity, and not to keep in the anonymised dataset. Outliers can lead to easy re-identification. It can be applied before or after other techniques (e.g. generalisation) have been applied.

7.3. **How to use it**: Delete the entire record. Note that the suppression should be permanent, and not just a '"hide row"[6] function; similarly, 'redacting' may not be sufficient if the underlying data remains accessible.

Other tips:

7.4. Refer to the example in the section on generalisation for illustration of how record suppression is used.

7.5. Note that removal of a record can impact the dataset, e.g. in terms of statistics such as average and median.

8 Character Masking

8.1. **Description**: Character masking is the change of the characters of a data value, e.g. by using a constant symbol (e.g. "*" or "x"). Masking is typically partial, i.e. applied only to some characters in the attribute.

8.2. **When to use it**: When the data value is a string of characters and hiding part of it is sufficient to provide the extent of anonymity required.

[6] Found in spreadsheet software

8.3. **How to use it**: Depending on the nature of attribute, replace the appropriate characters with a chosen symbol. Depending on the attribute type, you may decide to 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:

8.4. Note that masking may need to take into account whether the length of the original data provides information about the original data. Subject matter knowledge is critical especially for partial masking to ensure the right characters are masked. Special consideration may also apply to checksums within the data; sometimes the checksum could be used to recover (other parts of) the masked data. As for complete masking, the attribute could alternatively be suppressed unless the length of the data is of some relevance.

8.5. The scenario of masking data in such a way that data subjects are meant to recognise their own data is a special one, and does not belong to the usual objectives of data anonymisation. An example of this is the publishing of lucky draw results, whereby typically the names and partially masked NRIC numbers of lucky draw winners are published for the individuals to recognise themselves as winners. Note that generally, anonymised data should *not* be recognisable even to the data subject themselves.

8.6. Example

This example shows an online grocery store conducting a study of its delivery demand from historical data, in order to improve operational efficiency. The company masked out the last 4 digits of the postal codes, leaving the first 2 digits, which correspond to the "sector code" within Singapore.

Before anonymisation:

Postal Code	Favourite Delivery Time Slot	Average No. of Orders Per Month
100111	8 pm to 9 pm	2
200222	11 am to 12 noon	8
300333	2 pm to 3pm	1

After partial masking of postal code:

Postal Code	Favourite Delivery Time Slot	Average No. of Orders Per Month
10xxxx	8 pm to 9 pm	2
20xxxx	11 am to 12 noon	8
30xxxx	2 pm to 3pm	1

9 Pseudonymisation

Description:

9.1. The replacement of identifying data with made up values. Pseudonymisation is also referred to as coding. Pseudonyms can be irreversible, where the original values are properly disposed and the pseudonymisation was done in a non-repeatable fashion, or reversible (by the owner of the original data), where the original values are securely kept but can be retrieved and linked back to the pseudonym, should the need arises[7].

9.2. Persistent pseudonyms allow linkage by using the same pseudonym values to represent the same individual across different datasets. On the other hand, different pseudonyms may be used to represent the same individual in different datasets to prevent linking of the different datasets.

9.3. Pseudonyms can also be randomly or deterministically generated.

9.4. **When to use it**: When data values need to be uniquely distinguished and where no character or any other implied information of the original attribute shall be kept.

9.5. **How to use it**: Replace the respective attribute values with 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 should have no relationship to the original values (such that one can derive the original values from the pseudonyms).

Other tips:

9.6. When allocating pseudonyms, ensure not to re-use pseudonyms that have already been utilised (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).

9.7. Persistent pseudonyms usually provide better utility by maintaining referential integrity across datasets.

9.8. For reversible pseudonyms, the identity database cannot be shared with the recipient; it should be securely kept and can only be used by the organisation to resolve any specific queries (however, the number of such queries must be controlled, otherwise they can be used to "decode" the entire pseudonymisation).

9.9. Similarly, if encryption is used, the encryption key cannot be shared, and in fact must be securely protected from unauthorised access, because a leak of such a key could

[7] For example, in the event that a research study yields results that would be able to provide useful warning to a data subject.

result in a data breach by enabling the reversal of the encryption. The same applies for pseudo-random number generators, which require a seed. Security of any key used must be ensured like with any other type of encryption or reversible process[8].

9.10. If encryption is used, review the method of encryption (e.g. algorithm and key length) periodically to ensure that it is recognised by the industry as relevant and secure.

9.11. 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, or simply to look more similar to the original attribute); 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 which have the same format as the original data.

9.12. Example

This example shows pseudonymisation being applied to the names of persons who obtained their driving licenses, and some information about them. In this example, the names were replaced with pseudonyms instead of the attribute being suppressed, because the organisation wanted to be able to reverse the pseudonymisation if necessary.

Before anonymisation:

Person	Pre Assessment Result	Hours of Lessons Taken Before Passing
Joe Phang	A	20
Zack Lim	B	26
Eu Cheng San	C	30
Linnie Mok	D	29
Jeslyn Tan	B	32
Chan Siew Lee	A	25

After pseudonymising the Person attribute:

Person	Pre Assessment Result	Hours of Lessons Taken Before Passing
416765	A	20
562396	B	26
964825	C	30
873892	D	29
239976	B	32
943145	A	25

[8] Note that relying on a proprietary or "secret" reversal process (with or without key) is likely more prone to decoding and the risk of being broken than relying on standard key based encryption.

For reversible pseudonymisation, the identity database is securely kept in case there is a future legitimate need to identify individuals. Security controls (including administrative and technical ones) should also be used to protect the identity database.

Identity database (single coding):

Pseudonym	Person
416765	Joe Phang
562396	Zack Lim
964825	Eu Cheng San
873892	Linnie Mok
239976	Jeslyn Tan
943145	Chan Siew Lee

9.13. Example

For added security regarding the identity database, double coding can be used. Continuing from the previous example, this example shows the additional linking database, which is placed with a trusted third party. With double coding, the identity of the individuals can only be known when both the trusted third party (having the linking database) and the organisation (having the identity database) put their databases together.

After anonymisation:

Person	Pre Assessment Result	Hours of Lessons Taken Before Passing
373666	A	20
594824	B	26
839933	C	30
280074	D	29
746791	B	32
785282	A	25

Linking database (securely kept by a trusted third party only; even the organisation will remove it eventually. The third party is not given any other information)

Pseudonym	Interim Pseudonym
373666	OQCPBL
594824	ALGKTY
839933	CGFFNF
280074	BZMHCP
746791	RTJYGR
785282	RCNVJD

Identity database (securely kept by the organisation)

Interim pseudonym	Person
OQCPBL	Joe Phang
ALGKTY	Zack Lim
CGFFNF	Eu Cheng San
BZMHCP	Linnie Mok
RTJYGR	Jeslyn Tan
RCNVJD	Chan Siew Lee

Note: in both the linking database and identity database, it is good practice to scramble the order of the records rather than leave it in the same order as the dataset. In this example the two are left in the original order for easier visualisation.

10 Generalisation

10.1. **Description**: a deliberate reduction in the precision of data. E.g. converting a person's age into an age range, or a precise location into a less precise location. This technique is also referred to as recoding.

10.2. **When to use it**: for values that can be generalised and still be useful for the intended purpose.

10.3. **How to use it**: Design appropriate data categories and rules for translating data. Consider suppressing any records that still stand out after the translation (i.e. the generalisation).

Other tips:

10.4. Design the data ranges with appropriate sizes. Data ranges that are too large may mean that the data may be "modified" very much, while data ranges that are 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 will affect the data ranges too. Note that the first and the last range may be a larger range to accommodate the typically lower number of records at these ends; this is often referred to as top/bottom coding.

10.5. Example

In this example, this dataset contains person name (which has already been pseudonymised), age in years, and residential address.

Before anonymisation:

S/n	Person	Age	Address
1	357703	24	700 Toa Payoh Lorong 5
2	233121	31	800 Ang Mo Kio Avenue 12
3	938637	44	900 Jurong East Street 70
4	591493	29	750 Toa Payoh Lorong 5
5	202626	23	5 Tampines Street 90
6	888948	75	1 Stonehenge Road
7	175878	28	10 Tampines Street 90
8	312304	50	50 Jurong East Street 70
9	214025	30	720 Toa Payoh Lorong 5
10	271714	37	830 Ang Mo Kio Avenue 12
11	341338	22	15 Tampines Street 90
12	529057	25	18 Tampines Street 90
13	390438	39	840 Ang Mo Kio Avenue 12

For the age, the approach taken is to generalise into the following age ranges.

< 20
21-30
31-40
41-50
51-60
> 60

For the address, one possible approach is to remove the block/house number and retain only the road name.

After generalisation of Age and Address:

S/n	Person	Age	Address
1	357703	21-30	Toa Payoh Lorong 5
2	233121	31-40	Ang Mo Kio Avenue 12
3	938637	41-50	Jurong East Street 70
4	591493	21-30	Toa Payoh Lorong 5
5	202626	21-30	Tampines Street 90
6	888948	>60	Stonehenge Road
7	175878	21-30	Tampines Street 90
8	312304	41-50	Jurong East Street 70
9	214025	21-30	Toa Payoh Lorong 5
10	271714	31-40	Ang Mo Kio Avenue 12
11	341338	21-30	Tampines Street 90
12	529057	21-30	Tampines Street 90
13	390438	31-40	Ang Mo Kio Avenue 12

Supposing there is, in fact, only 1 residential unit on Stonehenge Road, as an example. The exact address can hence be derived, even though the data has gone through the generalisation. This could be considered as being still "too unique".

Hence, as a next step of generalisation, record 6 could be removed (i.e. using the suppression technique) as the address is still too unique after removing the unit number. Alternatively, all the addresses could be generalised to a greater extent (e.g. town or district) such that suppression is not needed, but this might affect the utility of the data much more than suppression a few records from the dataset.

11 Swapping

11.1. **Description**: The purpose of swapping is to rearrange data in the dataset such that the individual attribute values are still represented in the dataset, but generally, do not correspond to the original records. This technique is also referred to as shuffling and permutation.

11.2. **When to use it**: when subsequent analysis only needs to look at aggregated data, or analysis is at the intra-attribute level; in other words, there is no need for analysis of relationships between attributes at the record level.

11.3. **How to use it**: First, identify which attributes to swap. Then, for each, swap or reassign the attribute values to any record in the dataset.

11.4. **Other tips**: Assess and decide which attributes (columns) need to be swapped. Depending on the situation, organisations may decide that, for instance, only attributes (columns) containing values that are relatively identifying, need to be swapped.

11.5. Example

In this example, the dataset contains information about customer records for a business organisation.

Before anonymisation:

Person	Job Title	Date of Birth	Membership Type	Average Visits per Month
A	University dean	3 Jan 1970	Silver	0
B	Salesman	5 Feb 1972	Platinum	5
C	Lawyer	7 Mar 1985	Gold	2
D	IT professional	10 Apr 1990	Silver	1
E	Nurse	13 May 1995	Silver	2

After anonymisation:

In this example, all values for all attributes have been swapped.

Person	Job Title	Date of Birth	Membership Type	Average Visits per Month
A	Lawyer	10 Apr 1990	Silver	1
B	Nurse	7 Mar 1985	Silver	2
C	Salesman	13 May 1995	Platinum	5
D	IT professional	3 Jan 1970	Silver	2
E	University dean	5 Feb 1972	Gold	0

Note: On the other hand, if the purpose of the anonymised dataset is to study the relationships between job profile and consumption patterns, other methods of anonymisation may be more suitable, e.g. via generalisation of job titles, which could result in "university dean" being modified to become "educator".

12 Data Perturbation

12.1. **Description**: the values from the original dataset are modified to be slightly different.

12.2. **When to use it**: for quasi-identifiers (typically numbers and dates) which may potentially be identifying when combined with other data sources, and slight changes in value are acceptable. This technique should not be used where data accuracy is crucial.

12.3. **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 base-*x* rounding.

Other tips:

12.4. 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.

12.5. Note that where computation is performed on attribute values which have been perturbed, the resulting value may experience perturbation to an even larger extent.

12.6. Example

In this example, the dataset contains information to be used for research on possible linkage between a person's height, weight, age, whether the person smokes, and whether the person has "disease A" and/or "disease B". The person's name has already been pseudonymised.

The following rounding is then applied:

Attribute	Anonymisation technique
Height (in cm)	Base-5 rounding (5 is chosen to be somewhat proportionate to the typical height value of, e.g. 120 to 190 cm)
Weight (in kg)	Base-3 rounding (3 is chosen to be somewhat proportionate to the typical weight value of, e.g. 40 to 100 kg)
Age (in years)	Base-3 rounding (3 is chosen to be somewhat proportionate to the typical age value of, e.g. 10 to 100 years)
(the remaining attributes)	Nil, due to being non-numerical and difficult to modify without substantial change in value

Dataset before anonymisation:

Person	Height (cm)	Weight (kg)	Age (years)	Smokes?	Disease A?	Disease B?
198740	160	50	30	No	No	No
287402	177	70	36	No	No	Yes
398747	158	46	20	Yes	Yes	No
498732	173	75	22	No	No	No
598772	169	82	44	Yes	Yes	Yes

Dataset after anonymisation (shaded columns represent the affected attributes):

Person	Height (cm)	Weight (kg)	Age (years)	Smokes?	Disease A?	Disease B?
198740	160	51	30	No	No	No
287402	175	69	36	No	No	Yes
398747	160	45	18	Yes	Yes	No
498732	175	75	21	No	No	No
598772	170	81	42	Yes	Yes	Yes

Note: for base-*x* rounding, the attribute values to be rounded are rounded to the nearest multiple of *x*.

13 Synthetic Data

13.1. **Description**: this technique is slightly different as compared to the other techniques described in this Guide, as it is mainly used to generate synthetic datasets directly and separately from the original data, instead of modifying the original dataset.

13.2. **When to use it**: typically, when a large amount of data is required for system testing, but the actual data cannot be used and yet the data should be "realistic" in certain aspects, like format, relationship among attributes, etc.

13.3. **How to use it:** study the patterns from the original dataset (i.e. the actual data) and apply the patterns when creating the "anonymised" dataset (i.e. the synthetic data). The degree to which the patterns from the original dataset need to be replicated depends on the how the anonymised dataset is to be used.

Other tips:

13.4. Depending on the test scope and the administrative controls, fully or partially synthetic data can be generated; e.g. where tests are conducted, which need to reference to other datasets, then those few items being tested need to remain in their original form, but other information could be synthetic.

13.5. While in the other techniques, typically the anonymised data is the same or about the same (e.g. when suppression or aggregation are applied) volume as the original data, synthetic data can be generated in any volume, as needed.

13.6. When applying this technique, outliers may need additional attention. For testing purposes outliers are often very valuable, but outliers in the synthetic data may also indicate certain outliers within the original dataset. It is therefore recommended to create outliers in synthetic data intentionally and independent of the original data.

13.7. This technique is of rather little utility for data analysis, because the data is not "real" and the data was created based on a pre-conceived model.

13.8. Example

In this example, an office facility which specialises in providing "hot-desking" facilities keep records of the time that users start and end using their facilities. They would like to create a set of synthetic data to perform simulation testing on a new facility allocation algorithm.

A detailed discussion of statistical measures is beyond the scope of this Guide, however in this example, some possible measures could be the average or median number of users during each hour of the day.

Original dataset:

User	Date	Time in	Time out
User A	1-Mar-17	8:27	18:04
User A	2-Mar-17	8:20	18:10
User B	1-Mar-17	8:45	17:17
User B	2-Mar-17	8:55	17:54
User C	1-Mar-17	13:18	15:48
User C	2-Mar-17	13:02	16:02
User D	1-Mar-17	17:55	7:31
User D	2-Mar-17	18:04	7:39
(etc.)	(etc.)	(etc.)	(etc.)

Statistics obtained from original dataset

Start Time	End Time	Average No. of Users
0:00	1:00	130
1:00	2:00	98
2:00	3:00	102
3:00	4:00	95
4:00	5:00	84
5:00	6:00	72
6:00	7:00	62
7:00	8:00	144
8:00	9:00	450
9:00	10:00	506
(etc.)	(etc.)	(etc.)
22:00	23:00	138
23:00	0:00	132

Synthetic dataset (for 1 day):

User	Date	Time in	Time out
100001	3-Apr-17	8:25	17:53
100002	3-Apr-17	8:00	18:04
100003	3-Apr-17	8:12	18:48
100004	3-Apr-17	8:49	18:02
100005	3-Apr-17	8:33	18:11
100006	3-Apr-17	8:37	18:05
100007	3-Apr-17	8:55	20:05
100008	3-Apr-17	8:23	18:34
100009	3-Apr-17	13:16	15:48
100010	3-Apr-17	13:03	15:11
100011	3-Apr-17	13:28	15:25
100012	3-Apr-17	13:18	15:32
100013	3-Apr-17	17:55	7:38
100014	3-Apr-17	18:04	7:32
100015	3-Apr-17	17:57	7:02
(etc.)	(etc.)	(etc.)	(etc.)

Note: basically, the synthetic dataset is created based on the statistics derived from the original dataset, e.g. the average number of users in office at different time periods of the day.

14 Data Aggregation

14.1. **Description**: converting a dataset from a list of records to summarised values.

14.2. **When to use it**: when individual records are not required and aggregated data is sufficient for the purpose.

14.3. **How to use it**: a detailed discussion of statistical measures is beyond the scope of this Guide, however typical ways include using totals or averages, etc. It might also be also useful to discuss with the data recipient about the expected utility and find a suitable compromise.

Other tips:

14.4. Where applicable, watch out for groups having too few records after performing aggregation. E.g. in the below example if the aggregated data includes a single record in any of the categories, it could be easy for someone with some additional knowledge to identify a donor.

14.5. Hence, aggregation may need to be applied in combination with suppression. Some attribute may need to be removed, as they contain details which cannot be aggregated, and new attributes might need be added, e.g. to contain the newly computed aggregate values.

14.6. Example

In this example, a charity organisation has records of the donations made, as well as some information about the donors.

The charity organisation assessed that aggregated data is sufficient for an external consultant to perform data analysis, hence performs data aggregation on the original dataset.

Original dataset:

Donor	Monthly Income ($)	Amount donated in 2016 ($)
Donor A	4000	210
Donor B	4900	420
Donor C	2200	150
Donor D	4200	110
Donor E	5500	260
Donor F	2600	40
Donor G	3300	130
Donor H	5500	210
Donor I	1600	380
Donor J	3200	80
Donor K	2000	440

Donor L	5800	400
Donor M	4600	390
Donor N	1900	480
Donor O	1700	320
Donor P	2400	330
Donor Q	4300	390
Donor R	2300	260
Donor S	3500	80
Donor T	1700	290

Anonymised dataset:

Monthly Income ($)	**No. of Donations Received (2016)**	**Sum of Amount donated in 2016 ($)**
1000-1999	4	1470
2000-2999	5	1220
3000-3999	3	290
4000-4999	5	1520
5000-6000	3	870
Grand Total	**20**	**5370**

PART 4: PUTTING IT TOGETHER

15 Anonymisation Methodology

15.1. While Part 3 of this Guide focussed on various basic anonymisation techniques, anonymisation requires more than just applying the appropriate technique(s). Part 4 looks as the bigger picture and discusses what else needs to be considered. Please note that this description is mainly focussing on non-public release; public release models might need additional and more detailed considerations.

15.2. The following is a suggested methodology for performing anonymisation:

1) Determine the release model.

   This refers to how the anonymised dataset will be released. *Public* refers to making it available to basically anyone. *Non-public* refers to a controlled release to limited (and often, a fixed number of) known recipients. The public release model poses inherently more challenges on the anonymisation techniques.

2) Determine the acceptable re-identification risk threshold as well as the expected utility and risk threshold intended or required.
   Refer to section 17 for more details. Note that the risk threshold set at this stage must be clearly distinguished if the additional controls are taken into consideration or only reflect the risk of the data.

3) Classify data attributes.

   This is about classifying the attributes in the dataset as either direct identifiers, indirect identifiers, or non-identifiers, which affects how the attributes will subsequently be processed.

4) Remove unused data attributes.

   In the process of anonymisation, usually most attributes, whether direct or indirect identifiers, require processing or at least consideration, so as to become less identifying. Hence, any attribute that is clearly not required in the anonymised dataset should be suppressed.

5) Anonymise direct and indirect identifiers.

   This is done by applying techniques such as those described in this Guide. Different techniques are applicable for types of identifiers. Some techniques can (and often, should) be used in combination. Outlier records should be considered for record suppression.

6) Determine actual risk and compare against threshold.

   Refer to Section 17 for more details.

7) Perform more anonymisation, if necessary.

   If the actual risk is higher than the threshold, “stronger” anonymisation is required and steps 5 to 7 should be performed again with the necessary adjustments, until the actual risk is lower than the threshold.

8) Evaluate the solution.

   This includes examining the anonymised dataset to assess if the utility meets the target. If the utility is insufficient, the anonymisation process may need to be re-designed, or it may be considered whether anonymisation is feasible for this dataset.

9) Determine controls required.

   Controls include both technical and non-technical (e.g. legal and organisational measure). Technical controls are further described in section 18.

10) Document the anonymisation process.

    The details of the anonymisation process, parameters used and controls should be clearly recorded for future reference. Such documentation facilitates review, maintenance, fine-tuning and audits. Note that such documentation should be kept securely as the release of the parameters may facilitate re-identification.

16 *K*-anonymity – a measure of risk

16.1. *K*-anonymity (and similar extensions to it like *L*-diversity and *T*-closeness) is sometimes thought of as an anonymisation technique, but it is more of a measure used to ensure that risk threshold has not been surpassed, as part of the anonymisation methodology (see in particular step 6).

16.2. *K*-anonymity is not the only measure available nor is it without its limitations but it is relatively well understood and easy to apply. Alternative methods such as differential privacy[9] have emerged over the past few years.

16.3. **Description**: The *k*-anonymity model is used as a guideline before and for verification, after anonymisation techniques (e.g. generalisation) have been applied, to ensure any record's direct and/or indirect identifiers are shared by at least *k*-1 other records.

This is the key protection provided by *k*-anonymity against linking attacks, because *k* records (or at least different direct and indirect identifiers) are identical in the identifying attributes (and thereby create an "equivalence class" with *k* members), and therefore it is not possible to link or single out an individual's record; there are always *k* identical attributes.

An anonymised dataset may have different *k*-anonymity levels for different sets of indirect identifiers, but to assess the protection against linking, the lowest *k* is used for comparison against the threshold.

16.4. **When to use it**: to confirm that the anonymisation measures put in place achieve the desired threshold against linking attacks.

16.5. **How to use it**: First, decide on a value for *k* (which is basically equal to or higher than the inverse of the equivalence class size), which provides the lowest k to be achieved among all equivalence classes. Generally, the higher the value of *k*, the harder it is for data subjects to be identified; however, utility may become lower as *k* increases and more records may need to be suppressed. After other anonymisation techniques are applied, check that each record has at least *k*-1 other records with the same attributes addressed by the *k*-anonymisation. Records in equivalence classes with less than *k* records, should be considered for suppression; alternatively, more anonymisation can be done.

[9] Differential privacy involves several concepts, including answering queries instead of providing the anonymised dataset, adding randomised noise to the protect individual records, providing mathematical guarantees that the pre-defined "privacy budget" is not exceeded, etc.

Other tips:

16.6. Besides generalisation and suppression, synthetic data can also be created (e.g. near to the outliers) to achieve *k*-anonymity. These techniques (and others) can sometimes be used in combination, but note that the exact way chosen can affect data utility. Consider the trade-offs between dropping the outliers or inserting synthetic data.

16.7. *K*-anonymity assumes that each record pertains to a different individual. If the same individual has multiple records (e.g. visiting the hospital on several occasions), then *k*-anonymity will need to be higher than the repeat records, else the records may not only be linkable, but might despite seemingly fulfilling "*k* equivalence classes" be re-identifiable from the records.

16.8. Example

In this example, the dataset contains information about people taking taxis.

K = 2 is used, i.e. each record should eventually share the same attributes with 1 other record, after anonymisation. Note: k=2 is used for simplifying the example but it is probably too low a value for actual data, because this means the risk of identification would be 50%.

The following anonymisation techniques are used in combination and the level of granularity is one example which allows to achieve the required *k* level.

Attribute	Anonymisation technique
Age	Generalisation (10-year intervals)
Occupation	Generalisation – e.g. both "Database administrator" and "programmer" are generalised to "IT"
Record suppression	Records that do not meet the 2-anonymity criteria after anonymisation techniques have been applied (in this case, generalisation), are removed. E.g. for the case of the banker who is the only such data subject.

Dataset before anonymisation:

Age	Gender	Occupation	Average No. of Trips per Week
21	Female	Legal Counsel	15
38	Male	Data Privacy Officer	2
25	Female	Banker	8
44	Female	Database Administrator	3
25	Female	Administrative Assistant	1
31	Male	Data Privacy Officer	5
42	Female	Programmer	3

22	Female	Administrative Assistant	4
30	Female	Legal Counsel	2

Dataset after anonymisation of age and occupation and suppression of outlier (the respective equivalence classes are highlighted in different colours):

Age	Gender	Occupation	Average number of Trips per Week
21 to 30	Female	Legal Counsel	15
31 to 40	Male	Data Privacy Officer	2
~~21 to 30~~	~~Female~~	~~Banker~~	~~8~~
41 to 50	Female	IT	3
21 to 30	Female	Administrative Assistant	1
31 to 40	Male	Data Privacy Officer	5
41 to 50	Female	IT	3
21 to 30	Female	Administrative Assistant	4
21 to 30	Female	Legal Counsel	2

Note: The average number of trips per week is taken here as an example for a non-identifier, without a need to further anonymise this attribute. Also note that a dataset following k-anonymity without other non-identifiers or other attributes can be simplified by removing all the duplicates and just indicating the value of k.

17 Assessing the Risk of Re-Identification

17.1. There are various ways to assess the risk of re-identification, and these may involve rather complex calculations involving computation of probabilities. Refer to the reference publications in Annex B for detailed information.

17.2. This section describes a simplified model, using *k*-anonymity[10], and making certain assumptions. One of the assumptions is that the release model is non-public. The second assumption is that attack attempts to link an individual to the anonymised dataset. The third assumption is that the content of the anonymised data is not taken into consideration and that the risk is calculated independent of what kind of information the attacker actually has available.

17.3. First, the risk threshold should be established. This value, reflecting a probability, ranges between 0 to 1. It reflects the risk level that the organisation is willing to accept. The main factors affecting this should include the harm that could be caused to the data subject, as well as the harm to the organisation, should re-identification take place; but

[10] The calculations would be different if done using, e.g. differential privacy or traditional statistical disclosure controls.

it also takes into consideration what other controls have been put in place to mitigate risk in other forms than anonymisation. The higher the potential harm, the higher the risk threshold should be. There are no hard and fast rules as to what risk threshold values should be used; the following are just examples:

Potential Harm	Risk Threshold
Low	0.2
Medium	0.1
High	0.01

17.4. In computing the actual risk, this Guide explains looking into the "Prosecutor Risk", which assumes the adversary knows a specific person in the dataset and is trying to establish which record in the dataset refers to that person.

17.5. The simple rule for calculating the probability of re-identification for a single record in a dataset, is to take the inverse of the record's equivalence class size, i.e.

> P (link individual to a single record) = 1 / record's equivalence class size

17.6. Now, to compute the probability of re-identification of any record in the entire dataset, again, *given that there is a re-identification attempt*, a conservative approach would be to equate it to the maximum probability of re-identification among all records in the dataset.

> P (re-ID any record in dataset) = 1 / Min. equivalence class size in dataset
>
> Note: if the dataset has been *k*-anonymised,
> P (re-ID any record in dataset) <= 1 / k

17.7. We can consider 3 re-identification attack scenarios: (1) the deliberate insider attack; (2) inadvertent recognition by an acquaintance, and (3) data breach.

> P (re-ID) = P (re-ID | re-ID attempt) x P (re-ID attempt)
>
> Where P (re-ID | re-ID attempt) refers to the probability of successful re-identification, given there is a re-identification attempt. As discussed earlier, we can take P (re-ID | re-ID attempt) to be (1 / Min. equivalence class size in dataset)
>
> Therefore, P (re-ID) = (1 / Min. equivalence class size in dataset) x P (re-ID attempt)

17.8. For scenario #1 – the deliberate insider attack, we assume a party receiving the dataset attempts re-identification. To estimate P (re-ID attempt), i.e. the probability of a re-identification attempt, factors that can be considered include the extent of mitigating controls put in place as well as the motives and capacity of the adversary. The following table presents example values; again it is for the party anonymising the dataset to decide on suitable values to use.

P (re-ID attempt) for scenario #1 – the deliberate insider attack		Motivation and Resources of Adversary		
		Low	Medium	High
Extent of Mitigating Controls	High	0.03	0.05	0.1
	Medium	0.2	0.25	0.3
	Low	0.4	0.5	0.6
	None	1.0	1.0	1.0

Factors affecting the motivation and resources of the adversary may include:

- Willingness to violate contract (assuming contract preventing re-identification) is in place
- Financial and time constrains
- Inclusion of high profile personalities (e.g. celebrities) in the dataset

Factors affecting the extent of mitigating controls include:

- Organisational structures
- Administrative controls (e.g. contracts)
- Technical and physical measures (refer to section 18)

17.9. For scenario #2 - inadvertent recognition by an acquaintance, we assume a party receiving the dataset inadvertently re-identifies a data subject while examining the dataset. This is possible because the party has some additional knowledge about the data subject due to their relationship (e.g. friend, neighbour, relative, colleague, etc.). To estimate P (re-ID attempt), i.e. the probability of a re-identification attempt, the main factor to be considered is the likelihood that the data recipient knows someone in the dataset.

17.10. For scenario #3 – a data breach occurring at the data recipient's ICT system, the probability can be estimated based on available statistics on the prevalence of data breaches in the data recipient's industry. This is based on the assumption that the attackers who obtained the dataset will attempt re-identification.

Scenario #3 – a data breach

P (re-ID attempt) = P (data breach in data recipient's industry)

17.11. The highest probability among the 3 scenarios should be used as P (re-ID attempt).

P (re-ID attempt) = Max (P(deliberate insider attack), P(inadvertent recognition by an acquaintance), P(data breach))

17.12. To put everything together,

P (re-ID) = (1 / Min. equivalence class size in dataset) x P (re-ID attempt)
= (1 / k) x P (re-ID attempt) *for k-anonymised dataset*

Where P (re-ID attempt) = Max (P (deliberate insider attack),
P (inadvertent recognition by an acquaintance), P(data breach))

18 Technical Controls

18.1. This section discusses *technical* controls that can be implemented to further reduce the risk of re-identification after anonymisation. The controls may or may not be suitable for the situation, depending on the policy of access to the anonymised dataset. Where relevant, these controls can generally be implemented in combination with one another. Note that some of these are only effective provided the anonymised dataset with high residual risk is not passed over to the recipient, as once this has been done, technical controls are typically not possible anymore. Also note that a risk-based approach should be taken; hence the controls discussed in this section are for consideration and not mandatory to adopt.

18.2. Revocable access – with records of access granted, it may be possible, depending on the type of technical control used, to revoke access where deemed necessary. Typically, this is easier to implement where only online access to the dataset is given.

18.3. Query only - Allowing queries to be made instead of providing direct access to the dataset. An even more secure mode is for each query to be vetted by a curator who assesses whether the specific query should even be granted.

18.4. Limit recipients – This is often done by implementing user authentication and authorisation where online access to the dataset or to query is given, or password protection or encryption where offline access to the dataset is given.

18.5. Digital Rights Management (DRM) Controls – This is usually done by providing online access but implementing additional controls such as not allowing the user to save or print the data. Note that there are limitations such as not being able to prevent photographs to be taken of the onscreen data.

18.6. On-site access – Requiring the user to be physically present at the site where access to the dataset is provided, or where access to perform queries is given. The additional security comes being able to control what the user does with the data, e.g. to disallow even photographs to be taken of the onscreen data. Additional security measures taken at the site could include no network/internet connection provided, no phones or external computers being allowed, CCTV surveillance, etc.

18.7. Providing only a subset of the anonymised dataset – The subset can also be a randomly selected and/or perturbed.

18.8. Physical measures – The above measures are mostly relating to the control of access to the anonymised data in digital form. Physical measures apply too; examples of these include restricting physical access to devices or storage devices containing or able to access anonymised data, as well as restricting access to printouts containing anonymised data.

19 Governance

19.1. The methodology given in section 15 describes the steps required to methodically anonymise a dataset. However, responsible anonymisation does not end there. Note that a risk-based approach should be taken; hence the suggestions discussed in this section are for consideration and not mandatory to adopt.

19.2. After the release of an anonymised dataset, proper governance relating to the anonymised dataset is required even for non-public release. This may include the following:

- Keeping track of anonymised datasets released by the organisation. Details include the recipients and the method of access (e.g. providing a copy of the anonymised dataset, or online access, or physical access, or query only, etc.) This includes different variants/subsets of datasets, as well as datasets released by different parts of the organisation; in both of these scenarios, combination of different datasets can lead to re-identification.

- Management of keys and mapping tables – some anonymisation techniques, including pseudonymisation, require the use of encryption keys, mapping tables, etc. It is crucial for these to be properly managed and securely kept, as any party that gets hold of these can immediately undo the anonymisation.
- Regularly reviewing re-identification risk and controls put in place
- Conducting audits on data recipients, to ensure they are complying with contractual requirements
- Notifying the relevant parties if a data breach occurs[11].
- Keeping track of compliance requirements and best practices regarding data anonymisation.

20 Acknowledgements

20.1. In developing this Guide, best practices from personal data protection agencies and other authorities in other countries were considered. Refer to Annex B for the guides that were referenced.

20.2. We would like to express our appreciation to the following organisations for their valuable input in the development of this Guide:

- Agency of Integrated Care (AIC)
- Changi General Hospital (CGH)
- Cyber Security Agency of Singapore (CSA)
- Government Technology Agency (GovTech)
- Institute of Mental Health (IMH)
- Integrated Health Information Systems Pte Ltd (IHIS)
- JurongHealth
- Khoo Teck Puat Hospital
- KK Women's & Children's Hospital (KKH)
- Ministry of Health (MOH)
- MOH Holdings Pte Ltd (MOHH)
- Nanyang Technological University (NTU)
- National Cancer Centre Singapore
- National Dental Centre Singapore
- National Healthcare Group (NHG)
- National Healthcare Group Polyclinics (NHGP)
- National Heart Centre Singapore
- National Neuroscience Institute

[11] Refer to PDPC's *Guide to Managing Data Breaches*

- National Skin Centre
- National University Health System (NUHS)
- National University Hospital (NUH)
- National University of Singapore (NUS)
- National University Polyclinics (NUP)
- Privitar Ltd
- Saw Swee Hock School of Public Health
- Sengkang Health
- Singapore Department of Statistics (SingStat)
- Singapore General Hospital (SGH)
- Singapore Health Services (Singhealth HQ)
- Singapore Management University (SMU)
- Singapore National Eye Centre
- Singhealth Polyclinics
- Tan Tock Seng Hospital (TTSH)

END OF DOCUMENT

Annex A: Summary of Anonymisation Techniques

Technique Name	When to Use	Attribute type
Attribute suppression	Attribute is not required in the anonymised dataset	All
Record suppression	Presence of outlier records	N.A. (applies across entire record, hence all attributes affected)
Character masking	Masking some characters in an attribute provides sufficient anonymity	Direct identifier
Pseudonymisation	Records still need to be distinguished from each other in the anonymised dataset but no part of the original attribute value can be retained	Direct identifier
Generalisation	Attributes can be modified to be less precise but still be useful	All
Swapping	No need for analysis of relationships between attributes at the record level	All
Data perturbation	Slight modification to the attributes are acceptable	Indirect identifier
Synthetic data	Large amounts of made up data similar in nature to the original data are required, e.g. for system testing	All
Data aggregation	Individual records are not required and aggregated data is sufficient	Indirect identifiers

Annex B: Main References

- "Advisory Guidelines on Key Concepts In the PDPA" (Chapter 5 – Personal Data). https://www.pdpc.gov.sg/AG. Personal Data Protection Commission (Singapore), revised 27 July 2017
- "Advisory Guidelines on the PDPA for Selected Topics" (Chapter 3 – Anonymisation). https://www.pdpc.gov.sg/AG. Personal Data Protection Commission (Singapore), revised 28 March 2017
- "De-identification Guidelines for Structured Data". https://www.ipc.on.ca/wp-content/uploads/2016/08/Deidentification-Guidelines-for-Structured-Data.pdf. Information and Privacy Commissioner of Ontario, June 2016.
- "Guide to Managing Data Breaches". https://www.pdpc.gov.sg/OG. Personal Data Protection Commission (Singapore), 8 May 2015
- El Emam K. *Guide to the De-Identification of Personal Health Information.* CRC Press, 2013.
- "Opinion 05/2014 on Anonymisation Techniques". http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf. Article 29 Data Protection Working Party (European Commission),10 April 2014.
- "Personal Data Protection Act 2012". *Government Gazette*. https://sso.agc.gov.sg/Act/PDPA2012. Republic of Singapore, 7 December 2012
- S L Garfinkel. "NISTIR 8053: De-Identification of Personal Information". http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf. National Institute of Standards and Technology (NIST), October 2015.

BROUGHT TO YOU BY