3 Anonymisation

What is anonymisation?

3.1 In general, anonymisation refers to the process of removing identifying information such that the remaining data cannot be used to identify any particular individual. The terminology that has been used to refer to this process varies across different jurisdictions and sectors. For instance, some jurisdictions use 'anonymisation' and 'de-identification' interchangeably to refer to the process of converting personal data into data that can no longer be used to identify an individual, whether alone or in combination with other available information. Others use 'anonymisation' to refer to de-identification that is irreversible.[2] For the purposes of these Guidelines, the term 'anonymisation' refers to the process of converting personal data into data that cannot be used to identify any particular individual, and can be reversible or irreversible. The reversibility of the specific process used would be a relevant consideration for organisations when managing the risk of re-identification.

3.2 Anonymisation is a set of risk management controls and has to be understood in the context of the following definition of personal data under section 2(1) of the PDPA: "data, whether true or not, about an individual who can be identified — (a) from that data; or (b) from that data and other information to which the organisation has or is likely to have access". Please see the section on "Personal Data" in the Key Concepts Guidelines for more details.

3.3 Data would not be considered anonymised if there is a serious possibility that an individual could be re-identified, taking into consideration both:

a) the data itself, or the data combined with other information to which the organisation has or is likely to have access; and

b) the measures and safeguards (or lack thereof) implemented by the organisation to mitigate the risk of identification.

3.4 Data that has been anonymised is not personal data, and the Data Protection Provisions in Parts III to VI of the PDPA do not apply to the collection, use or disclosure of such data.

Why anonymise personal data?

3.5 Anonymisation of personal data is carried out to render the resultant data suitable for more uses than its original state would permit under data protection regimes. For

[2] National Institute of Standards and Technology, *De-identification of Personal Information (NISTIR 8053)* http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf

example, anonymised data may be used for research and data mining where personal identifiers in the data are unnecessary or undesired. Anonymised datasets could also be a protection measure against inadvertent disclosures and security breaches.

3.6 Where individuals need not be identifiable for the purposes in question, it is usually good practice to collect, use and disclose data in an anonymised form.

3.7 In circumstances where an organisation must cease to retain personal data under section 25 of the PDPA, the organisation may anonymise the data. By doing so, it will have ceased to retain personal data, since anonymised data is not personal data.

Anonymisation techniques

3.8 The following is a non-exhaustive list of commonly used anonymisation techniques, and examples of how each technique can be used.

3.9 **Examples of Anonymisation Techniques:**

a) *Pseudonymisation*: replacing personal identifiers with other references. For example, replacing an individual's name with a tag or reference number, which is randomly generated.

b) *Aggregation*: displaying values as totals, so that none of the individual values which could identify an individual is shown. For example, given a dataset with the ages of eight individuals (i.e. 33, 35, 34, 37, 42, 45, 37, 40), displaying the sum of the individual ages of the total number of individuals in a group (i.e. 303), rather than the age of each individual represented discretely.

c) *Replacement*: replacing values or a subset of the values with a computed average or a number derived from the values. For example, replacing the individuals with ages of 15, 18 and 20 with an age value of 17 to blur the distinction, if exact age is not required for the desired purposes.

d) *Data suppression*: removing values that are not required for the purpose. For example, removing 'ethnicity' from a dataset of individuals' attributes.

e) *Data recoding or generalisation*: banding or grouping of categories into broader categories. For example, the grouping of exact education level (e.g. K1, Primary 3, Secondary 2) into broader categories (e.g. pre-primary, primary, secondary etc.), or hiding the value within a given range (e.g. replacing age '43' with the range '40-50').

	f)	*Data shuffling*: mixing up or replacing values with those of the same type so that the information looks similar but is unrelated to the actual details. For example, the surnames in a customer database could be sanitised by replacing them with those drawn from another database.
	g)	*Masking*: removing certain details while preserving the look and feel of the data. For example, representing a full string of NRIC numbers as '#####567A' instead of 'S1234567A'.

3.10 The Commission does not recommend or endorse the use of any particular technique. Organisations should make their own assessment of their situation(s) and operational context, and adopt the most appropriate anonymisation technique.

Considerations for anonymising data

3.11 When deciding whether to anonymise data for use or disclosure, organisations should keep in mind that not all datasets can be effectively or meaningfully anonymised. The following section elaborates on some relevant considerations.

Nature and type of data

3.12 The nature of the dataset affects how much identifying information needs to be removed so that it no longer can be used to identify individuals. Some data types are inherently 'rich' and full of information (e.g. portrait photographs taken for facial recognition purposes), such that any alteration to anonymise the data might render it useless for its intended purposes.

3.13 The uniqueness of a record or data point within a sample dataset or population also contributes to the challenge of anonymisation. There are instances where the use of particular methods may not anonymise the dataset, because data points for certain individuals remain unique. For example, a dataset containing the ages of individuals has one outlier of age 89 while the other ages are below 50. No matter how the ages are generalised or recoded into ranges, the data point for the 89-year-old stands out. Where there are records or data points that are unique to the population or sample (i.e. population unique or sample unique)[3] the risk of re-identification increases. Organisations must be careful to manage these risks, and ensure that the dataset is effectively anonymised.

[3] "Population unique" refers to a record within a dataset which is unique within the population on a given key. "Sample unique" refers to a record within a dataset which is unique within that dataset on a given key. Source: OECD Glossary of Statistical Terms

Potential impact on individuals

3.14 When deciding whether to anonymise personal data for use or disclosure, organisations should also consider any potential negative impact on the individuals if they were to be re-identified. This is particularly important if the personal data involved is of a highly sensitive nature, (e.g. records of individuals with HIV, records of adopted children and their biological parents). In such circumstances, even if the organisation assesses that there is a less than serious possibility of an individual being identified from the data, the organisation should carefully consider whether using or disclosing such data would be appropriate.

Assessing the risks of re-identification

3.15 Re-identification can occur as a result of combining separate datasets. When determining whether a dataset is anonymised, the organisation should consider whether there is a serious possibility that an individual can be identified from the dataset when it is combined with other information, by carrying out an assessment of the risk of re-identification. In general, factors affecting the risk of re-identification include the extent of disclosure, the amount of alteration the data has been subjected to in the course of anonymisation, the availability of other relevant information, and the recipients' ability and motivation to re-identify the data. Some of the considerations are discussed in the following paragraphs.

Nature of use and extent of disclosure

3.16 Organisations should consider how the nature of use and extent of disclosure of the anonymised data can affect the risk of re-identification. In general, where an organisation intends to use the anonymised data within the organisation itself, or where the organisation intends to disclose anonymised data only to a restricted group (or groups) of users (e.g. a closed community of researchers), re-identification risks can be managed better, compared to the situation where the organisation discloses the anonymised data to any users by publishing it.

Public knowledge and personal knowledge

3.17 Organisations should consider the types of information that could enable re-identification if combined with the anonymised data, as well as the ease with which such information can be accessed. It is necessary to understand the intended use and recipient of the anonymised data to perform this assessment and tailor an appropriate set of risk management controls for the intended use and recipient. For example, the risk of re-identification in a disclosure of data to a single entity for research and development of new products and services under a non-disclosure

agreement, would likely differ from the risk of re-identification in publishing the data to the world at large.

3.18 Even after the risks of re-identification have been assessed and appropriate controls implemented, the risk of spontaneous re-identification of data by persons with special knowledge of a particular individual can occur. For example, a patient's attending physician, who is also conducting medical research using datasets containing the patient's data, may be able to recognise the patient's medical profile from the data (even if such data may be assessed to be anonymised if the recipient did not have such special knowledge). If it is known or foreseeable to the organisation that the data might be accessed by any persons with special knowledge that could be used to re-identify any individual from the data, such risk must be accounted for in the risk assessment exercise.

Disclosing multiple datasets

3.19 If an organisation intends to disclose multiple anonymised datasets (e.g. by publishing seemingly unrelated datasets as part of an open data initiative), organisations should carry out a careful assessment of the risk of re-identification from such a disclosure, particularly if those datasets are extracted from the same database. Organisations would have to take particular care to ensure that risk management controls are adequate to prevent re-identification of individuals by the recipient organisation, particularly when the datasets are combined with each other or with additional datasets previously released by the organisation. Organisations could consider maintaining a centralised record to track datasets that have been disclosed or published.

3.20	**Example: Publishing multiple datasets** As part of an open data initiative, two divisions in Company WXY have independently decided to release related datasets, unaware of the other division's intentions. Both divisions then carried out assessments with incomplete and inaccurate information about the datasets being released by other divisions in the company. If each division then proceeds to publish their own datasets, this may increase the risk of re-identification where the two datasets are related.

Data recipient's ability and motivation to re-identify

3.21 The risk of re-identification is likely to differ depending on the ability and the motivation of the data recipient to re-identify individuals from the dataset. A data

recipient in possession of complementary information, specialised skills or technologies would more likely be capable of re-identifying individuals from the data than one that does not, assuming both have similar motivations.

3.22	**Example:** Anonymised dataset Y contains complex genome sequencing, a specialised type of data not easily understood by the layman. If the data recipient is an expert in genome sequencing, and understands how to identify individuals from such a dataset, the risk of re-identification may be higher. In contrast, if the data recipient has no knowledge of that data type, the risk re-identification would be comparatively lower.

3.23 Even if the data recipient has the requisite skills and information for re-identification, it does not necessarily mean the risk of re-identification is high. The motivation to re-identify must also be considered. The motivation to re-identify an individual may be low if there are barriers to re-identification, such as legal (e.g. via contractual obligations) or regulatory consequences (e.g. pursuant to government regulations or legally binding industry codes issued by regulators), for the data recipient, or if there is simply no incentive or benefit for a data recipient to re-identify individuals.

Changing environment

3.24 While a dataset may be anonymised at a particular point in time, it is not guaranteed that the dataset will stay anonymised permanently. The likelihood of re-identification for any given anonymised dataset is likely to increase over time, due to greater ease of access to and volume of other relevant information, increase in computing power, and improvements in data-linking techniques. A dataset that is sufficiently anonymised based on current technology might be more easily re-identified with technological advancements.

3.25 The adequacy of anonymisation techniques and risk management controls would have to be assessed in relation to the current state of technology. However, it is also important to build in robust organisational, legal and non-technical measures to manage the risks of re-identification, taking into account the possibility of technological developments over the period for which the data may be retained. Therefore, organisations should consider whether periodic re-assessment of re-identification risks should be included as a safeguard, particularly where there are developments that would significantly affect the risk of re-identification, or where anonymised data is disclosed and available over an extended period of time, for example in an ongoing relationship. For example, if the encryption used for

anonymisation of a particular dataset has been compromised, organisations should re-assess the risk of that anonymised dataset being re-identified and put in place additional safeguards to mitigate such risk.

General test for assessing risks of re-identification

3.26 As a general test for assessing the risks of re-identification and the robustness of the anonymisation, a useful starting point is the 'motivated intruder' test highlighted in the ICO's Code of Practice *Anonymisation: Managing Data Protection Risk Code of Practice*, which we adapt to our vernacular.

3.27 The 'motivated intruder' test considers whether individuals can be re-identified from anonymised data by someone who is motivated, reasonably competent, has access to standard resources (e.g. the Internet and published information such as public directories), and employs standard investigative techniques (e.g. making enquiries of people who may have additional knowledge of the identity of the data subject).

3.28 As anonymisation is a package of risk control measures tailored to the purpose of disclosure and the recipient, the 'motivated intruder' test has to accommodate the features of the intended recipient organisation.[4]

a) The assessment should include the totality of the risk management controls that are applicable to the recipient organisation. This refers to both technical measures as well as legal, regulatory or organisational measures. For example, the risk assessment could take into account the kind of safeguards accorded to the data, or how long the data is to be retained, among others.

b) The 'motivated intruder' test assumes that no particular individual has been targeted for identification and that the intruder does not resort to criminality or any specialist equipment or skills.

c) Where disclosure of a particular dataset is to a specific recipient whose motivations, re-identification capabilities, and other information available to that recipient are known or can be reasonably inferred, these known characteristics should also be accounted for.

d) In addition, the risk assessment for re-identification should also consider the other factors (not specific to anonymised data) that subject the anonymised data to re-identification risks. This includes all other 'residual'

[4] This can include disclosures to the public in cases where the dataset is published as part of open data efforts or in a publication that is freely available.

risks that are not directly related to a recipient's motivation and capability to re-identify or the risk management controls for the disclosed anonymised data; for example, risk of the data being compromised or mistakenly disclosed to unintended recipients.

Managing the risks of re-identification when using or disclosing anonymised data

3.29 Before using or disclosing anonymised data, the organisation should apply the appropriate anonymisation techniques to ensure robust anonymisation of the data. In general, the Commission would consider an organisation to have anonymised data if there is no serious possibility that a data user or recipient would be able to identify any individuals from the data.

3.30 When determining which anonymisation technique to adopt, some relevant considerations could include:

a) the nature or type of data to be anonymised; and/or

b) international best practices for anonymisation of the given data type.

3.31 Organisations may consider hiring anonymisation experts, statisticians, or independent risk assessors, to aid in their assessment of the appropriate anonymisation techniques to apply, particularly where the anonymisation issues are complex; for example, large datasets containing a wide range of personal data.

3.32 To further manage its risks, organisations may also consider putting in place other appropriate controls to lower the risk of re-identification, like:

a) limiting the number of data recipients to whom the information is disclosed and the number of persons that can access the information;

b) imposing restrictions on the data recipient on the use and subsequent disclosure of the data;

c) requiring the data recipient to implement processes to govern the proper use of the anonymised data in line with the restrictions; and/or

d) requiring the data recipient to implement processes and measures for the destruction of data as soon as the data no longer serves any business or legal purpose.

3.33 In addition, organisations may put in place controls to limit the data users' or recipients' access to "other information" that could re-identify the anonymised data. Depending on the circumstances, controls could be implemented through:

a) organisational structures;

b) legally binding agreements, administrative rules, or policies;

c) technical measures (e.g. using encryption to restrict access to the original dataset, limiting access to only authorised users, and controlling access through passwords); and/or

d) physical measures (e.g. restricted access areas).

3.34 It is not necessary that the most technically sophisticated anonymisation technology be used all the time. Rather, what is required is one that is sufficiently robust to manage the risk of re-identification, given the circumstances (e.g. the extent of disclosure, the intended recipient(s) and existing controls).

3.35 In the event that an organisation intentionally re-identifies an individual, such deliberate actions will constitute collection of personal data, for which consent is required from the relevant individual. There may be situations where the re-identification is unintentional. Generally, unintentional re-identification is not considered collection of personal data. However, the organisation should immediately delete the personal data or re-identifying information and should evaluate whether the risk management controls in place are adequate. If the organisation uses or discloses such unintentionally re-identified personal data, its actions will be considered to be use or disclosure of personal data. Generally, where such collection, use or disclosure is carried out for a purpose to which the relevant individuals did not consent[5], the organisation will have breached its PDPA obligations.

3.36 Given that management and assessment of re-identification risks is dependent on the circumstances of the use or disclosure, the following scenarios illustrate how re-identification risks can be assessed and managed in certain circumstances.

Use of anonymised data within the organisation

3.37 In general, if the departments within an organisation have, or are likely to have, access to other information (e.g. the decryption key or algorithm to reverse the anonymisation process) that can be combined with the data to re-identify individuals, the data will be considered personal data. and the Data Protection Provisions in Parts III to VI of the PDPA will apply.

3.38 Conversely, if anonymisation techniques and risk mitigation measures have been applied such that there is no serious possibility that the data can be used to identify

[5] And no relevant exception under the PDPA applies.

any individual, the Commission would consider such data anonymised. The Data Protection Provisions of the PDPA would not apply. For example, deploying one-way encryption in order to anonymise customer data for long term archival purposes.

3.39 There may be circumstances where an organisation wishes to convert personal data into anonymised datasets to be used for a particular purpose. However, the organisation may need to retain the original dataset or other information that can re-identify the individuals from the anonymised datasets for other purposes. In such situations, anonymisation can be relevant to the safe use of data for a particular purpose within an organisation. Organisational structures should establish effective barriers to access, by a group (or groups) of users within the organisation, to other information (e.g. the decryption key or algorithm that could reverse the anonymisation) held by the organisation that could be used to re-identify an individual.

3.40	**Example:** Department A and Department B are two departments within Organisation JKL. Department A manages personal data collected by the organisation for the purpose of customer relations. Department B wishes to use the data for business analytics purpose but does not require individually identifiable data. Department A proceeds to anonymise the data such that the anonymised dataset can no longer identify any individual. Organisation JKL puts in place controls on Department B to prevent re-identification of the anonymised dataset. This includes access restrictions to prevent data users in Department B from gaining access to other information held by Department A that can lead to re-identification, as well as administrative restrictions to prevent Department B from attempting re-identification. Organisation JKL also makes the unauthorised attempt to re-identify individuals from the anonymised dataset a breach of the terms of employment. Department B could be considered to be using anonymised data given that effective controls are imposed on Department B to prevent re-identification.

3.41 Once data is anonymised, in general, organisations must not be able to, in effect, use the same data as personal data (e.g. the data cannot be used to make a decision about, or in a manner that has a targeted impact on a specific individual). If organisations are, in effect, using the data as personal data, then organisations must ensure that consent has been obtained for such use, unless any exceptions apply.

3.42 Organisations must also be mindful of subsequent actions in respect of the anonymised data that could increase the risk of re-identification. For example, the Commission would consider there to be a serious possibility that an individual could be identified, if there are any subsequent disclosures of the anonymised data or information relating to the anonymised data to persons outside of Department B (e.g. where no effective controls are imposed on the recipients), and no further risk-mitigation measures are taken. Given that Department A holds the decryption key or other information (e.g. master database) that could be used to re-identify individuals from the anonymised data, the organisation must ensure the robust anonymisation of the dataset before any further disclosure (e.g. this may include further scrambling of the data).

Disclosure of anonymised data to a specific group (or groups) or data recipients

3.43 There may be instances where an organisation converts personal data into anonymised data in order to disclose it to a specific group (or groups) of recipients outside the organisation for other purposes, while the organisation continues to have access to other information that can re-identify the individuals (e.g. the decryption key or algorithm to reverse the anonymisation process). The risks of re-identification can be better managed where the recipients are limited to a specific group known to the disclosing organisation, compared to a disclosure to the world at large.

3.44 The Commission will consider it use or disclosure of anonymised data where the data is used by or disclosed to a specific group (or groups) of data users or recipients such that there is no serious possibility that any data user or recipient would be able to identify any individuals from the data. The Data Protection Provisions in Parts III to VI of the PDPA would not apply to such use and disclosure. In such instances, the recipients of the data cannot use it to make a decision about, or otherwise use in a manner that has an impact on, any specific individual.

3.45 Where disclosure of anonymised data is restricted to specific data recipients (e.g. a closed group of researchers) for their own use, without any further disclosure, the disclosing organisation may consider adopting legal measures to discourage any attempts by the data recipients to re-identify individuals from the anonymised data. For example, the disclosing organisation could put in place contractual safeguards, or require an undertaking from the data recipients not to attempt to re-identify the anonymised data. The disclosing organisation could also require that the data recipients put in place additional measures, such as governance frameworks, processes, and controls, to ensure the proper handling of the dataset and further reduce the risk of re-identification.

3.46 Further considerations when disclosing anonymised data are discussed in the previous section on *"Managing the risks of re-identification when using or disclosing anonymised data"*.

3.47	**Example:** Organisation JKL (disclosing organisation) intends to disclose an anonymised dataset to Organisation XYZ (data recipient). As part of its contractual agreement, Organisation XYZ has committed not to attempt re-identification of the dataset or to further disclose the anonymised data to another organisation. Organisation JKL also contractually requires that Organisation XYZ puts in place governance frameworks and controls to ensure the proper handling of the dataset, including: a) limiting the number of Organisation XYZ's employees who can access the anonymised dataset; b) requiring Organisation XYZ's employees authorised to access the dataset to be trained on data protection practices; and c) applying appropriate technical solutions to ensure better access controls (e.g. technical measures that limit employees' access to the dataset and the copying of or number of copies of the dataset). Organisation JKL follows through with periodic reviews of the governance frameworks and controls that Organisation XYZ has put in place.

3.48 In assessing the anonymisation and risk of re-identification of any dataset, the Commission will take a holistic view, not restricted to the measures discussed above, and include any other relevant facts of the case. For example, even if a disclosing organisation has employed robust anonymisation techniques and put in place legal safeguards to prevent attempts to re-identify and further disclose the anonymised data, if there are serious doubts about the reliability or reputation of the receiving organisation (e.g. in ensuring the proper use and handling of anonymised datasets), the Commission may consider the risk of re-identification significant, and the disclosure to be of personal data. On the other hand, the disclosure of anonymised data to a group of recipients with demonstrated reliability in ensuring the proper use and handling of anonymised datasets may be assessed to be of lower risk.

3.49 If there is no serious possibility that individuals could be identified from the disclosed dataset, the Commission would consider it a disclosure of anonymised data, and the

Data Protection Provisions in Parts III to VI of the PDPA would not apply to that disclosure.

3.50	**Example:** Organisation JKL (disclosing organisation) discloses an anonymised dataset to Organisation XYZ (data recipient), referred to as the Initial Disclosure. Organisation JKL lowers the risk of re-identification at the point of disclosure by: a) Ensuring that the anonymisation techniques applied to the data set are robust; b) Contractually requiring Organisation XYZ not to attempt re-identification or further disclosure of the anonymised data to another organisation; c) Contractually requiring Organisation XYZ to put in place governance frameworks and controls to ensure proper handling of the dataset; d) Ensuring that Organisation XYZ is a reliable organisation and has a good track record of adequate data protection. After receiving the dataset, Organisation XYZ decides that there is good reason for it to publish the dataset, and decides to do so (referred to as the Subsequent Disclosure), with the knowledge that this would constitute a breach of contract with Organisation JKL. An individual is subsequently re-identified from the published dataset. Generally, when determining whether Organisation JKL disclosed personal data or anonymised data during the Initial Disclosure, the Commission would consider whether Organisation JKL had implemented appropriate measures and safeguards, such that there was no serious possibility that individuals would have been identifiable, at the time of the Initial Disclosure. When investigating whether Organisation XYZ disclosed personal or anonymised data during the Subsequent Disclosure, the Commission will assess the measures and safeguards, taking into account that the Subsequent Disclosure was not limited to a specific data recipient but published to a wider audience.