Guia para Técnicas Básicas de Anonimização de Dados

(Traduçao não oficial)

Declaração

Esta tradução não oficial em Português do documento, disponibilizada pelo Gabinete para a Protecção de Dados Pessoais (GPDP) do Governo da RAEM, servindo apenas para a consulta dos interessados.

Este documento dedica-se ao objectivo referencial, não é lei ou regulamento vigente na RAEM, e não produz qualquer efeito legal. O GPDP ou qualquer outra entidade pública na RAEM não se responsabiliza por qualquer prejuízo ou dano provocado por este documento ou pela reprodução ou divulgação do mesmo.

Este documento pode ser publicado ou reproduzido para uso sem fins lucrativos. No entanto, o utilizador deve declarar que o documento é disponibilizado pelo GPDP e indicar a origem do documento em Inglês. Salvo autorização prévia por escrito do GPDP, ninguém pode reproduzir, reeditar, distribuir, divulgar ou proporcionar este documento para uso com fins lucrativos. O GPDP reserva o direito de responsabilizar o infractor nos termos da lei.

Governo da Região Administrativa Especial de Macau
Gabinete para a Protecção de Dados Pessoais

Abril de 2019 (1.ª versão)

Nota

O documento "Guia para Técnicas Básicas de Anonimização de Dados" é uma tradução para Português do documento "**GUIDE TO BASIC DATA ANONYMISATION TECHNIQUES**", publicado pela Comissão da Protecção de Dados Pessoais de Singapura (*Personal Data Protection Commission of Singapore*) (https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1(250118).pdf).

O texto original em Inglês foi publicado no dia 25 de Janeiro de 2018 pela Comissão da Protecção de Dados Pessoais de Singapura.

Este texto apresenta as diferentes ténicas de anonimização de dados, e elabora a interpretação da Comissão da Protecção de Dados Pessoais de Singapura em relação ao conceito "anonimização" no seu enquadramento legal, indicando os riscos de anonimização e a sua abordagem.

Recorda-se o leitor que esta Guia foi reparadas tendo como pano de fundo enquadramento legal de Singapura.

Assim alguns temas são discutidos sob uma perspectiva jurídica diferente da da RAEM, e deverá ter-se o cuidado de não retirar paralelos próximos com as soluções jurídicas de Macau para os mesmos problemas.

O Gabinete para a Protecção de Dados Pessoais formula votos de que os responsáveis pelo tratamento, os subcontratantes e o público em geral possam beneficiar dos úteis ensinamentos contidos no presente documento.

Governo da Região Administrativa Especial de Macau
Gabinete para a Protecção de Dados Pessoais

ÍNDICE

PARTE 1: PANORAMA

1. Introdução

1.1 A recolha, utilização e divulgação de dados pessoais de indivíduos por organizações em Singapura estão sujeitos à Lei da Protecção de Dados Pessoais de 2012 (*Personal Data Protection Act*, **PDPA**). A Comissão da Protecção de Dados Pessoais (*Personal Data Protection Commission*, **PDPC**) foi estabelecida para fiscalizar a PDPA e promover a sensibilização da protecção de dados pessoais em Singapura.

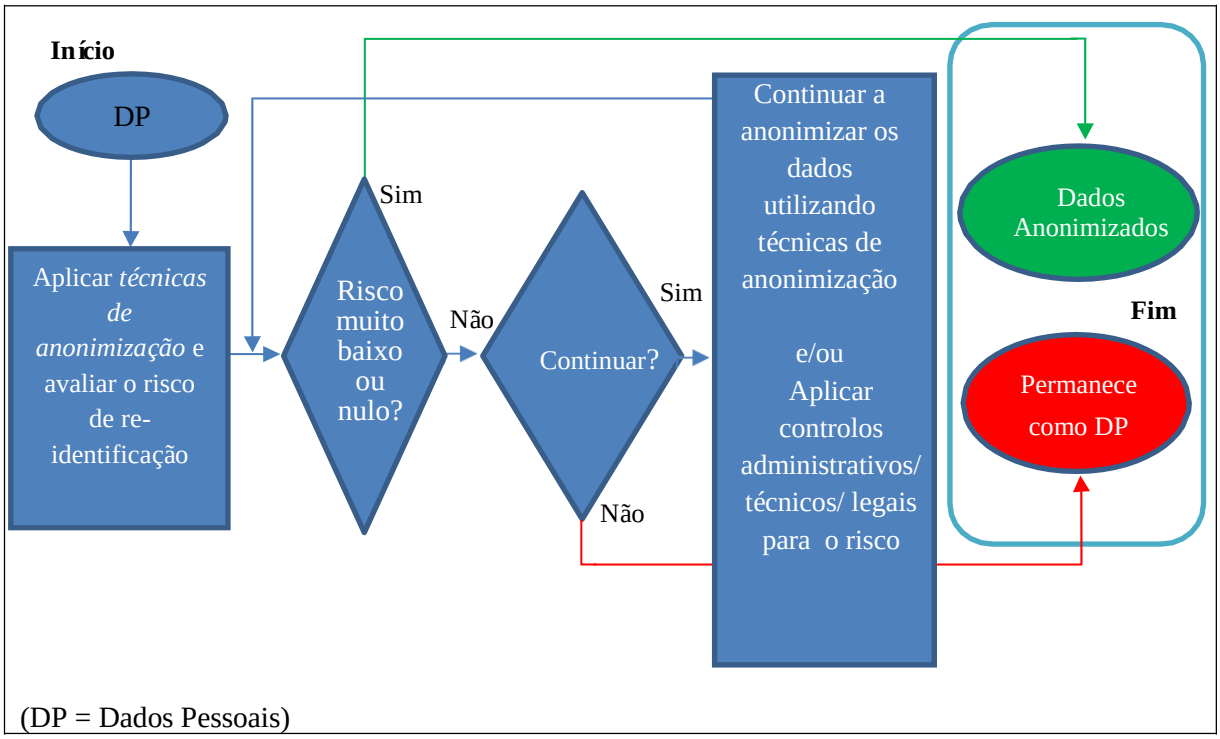
2. Objectivo e Âmbito da Guia

2.1 Esta Guia pretende fornecer uma introdução geral aos aspectos técnicos em relação a anonimização[1]. Deve ser consultado em conjunto com o Capítulo 3 (Anonimização) das Orientações Aconselhadas sobre os Tópicos Seleccionados na PDPA (adiante designadas por Orientações Aconselhadas) (*PDPC's Advisory Guidelines on the PDPA for Selected Topics*), emitidas pela PDPC, o qual propõe a interpretação e as considerações da PDPC para determinar o que constitui "anonimização" nos termos da PDPA.

2.2 Os básicos conceitos e técnicas discutidos nesta Guia fazem referência aos termos "anonimização de dados" e "dados anonimizados". A "Anonimização de dados" refere-se à conversão de dados pessoais em "dados anonimizados" com aplicação de um conjunto de "técnicas de anonimização". "Dados anonimizados", para os objectivos desta Guia, refere-se a dados que tenham sofrido transformação por técnicas de anonimização em combinação com a avaliação do risco de re-identificação. De modo geral, o processo de anonimização de dados seria "irreversível", e o recipiente do conjunto de dados anonimizado não conseguiria restabelecer os dados originais. Porém, há casos nos quais a organização que aplica a anonimização retenha a capacidade de recriar os dados originais dos dados anonimizados; nestes casos, o processo é "reversível".

2.3 Nesta Guia, pretende-se que os termos "anonimização de dados" e "dados anonimizados" sejam entendidos genericamente e alinhados com a literatura técnica do assunto. Não se pretende que sejam entendidos da mesma forma que os termos utilizados nas Orientações Aconselhadas, nem fornecer qualquer efeitos legais determinativos aos dados que sofreram transformação por técnicas de anonimização. O diagrama seguinte apresenta um resumo gráfico do conceito de anonimização de dados nas Orientações Aconselhadas:

[1] Para evitar mal-entendidos, anonimização nesta guia refere-se à transformação de dados existentes já disponíveis a uma organização. Não se refere ao aspecto de "anonimidade" de indivíduos, no qual, estes tentam impedir que a sua identidade seja identificada.



Para mais informações sobre a interpretação da PDPC em relação à "anonimização" e aos "dados anonimizados", é favor consultar as Orientações Aconselhadas.

2.4 Esta Guia tem como objectivo fornecer informações sobre as técnicas que poderão ser aplicadas a anonimização de dados. A Guia aborda principalmente as organizações que não pretendem libertar os dados anonimizados para o domínio público, mas que partilham dados com outras organizações ou entidades, nestes casos, os controlos técnicos e administrativos adicionais podem ser impostos para minimizar os riscos de revelação de dados pessoais sem a devida autorização . A utilização destas técnicas não garante necessariamente que os dados não coloquem qualquer risco sério de re-identificação e constituem por isso "dados anonimizados", aos quais a PDPA não se aplica.

2.5 Esta Guia não é um substituto para uma profissional acção de formação, literatura e serviços. A não ser que as organizações estejam familiarizadas com os riscos e contramedidas, sugere-se que as mesmas, ao divulgar dados anonimizados – especialmente se a intenção for libertar para o domínio público ou se a mesma envolver vários conjuntos de dados ou actualizações de dados anonimizados ao longo dos tempos –procurem aconselhamento profissional ou serviços para a anonimização dos dados.

2.6 Esta Guia descreve técnicas de anonimização para conjuntos de dados

estáticos, estruturados, bem-definidos, em formato textual e de nível único, onde:

- "Estáticos" se referem ao facto de os dados estarem totalmente disponíveis no momento da anonimização; isto contrasta com dados em fluxo, nos quais, as relações entre os dados podem não estar totalmente estabelecidas porque o fluxo fornece constantemente novos dados. Logo, é possível que os dados em fluxo precisem de outras técnicas de anonimização além daquelas que são descritas nesta Guia.
- "Estruturados" se referem ao facto de que a técnica de anonimização é aplicada a dados dentro de um formato conhecido e a uma localização conhecida dentro da base de dados. "Estruturados" não se limitam, portanto, a dados em formato tabular tais como uma folha de cálculo ou uma base de dados relacional, mas podem ser guardados ou divulgados em outros formatos definidos, como por exemplo XML, CSV, JSON, etc. Nesta Guia, se descrevem as técnicas e fornecem exemplos no formato tabular mais comum, mas isto não implica que as técnicas apenas se aplicam a esse formato, ou seja, formato tabular.
- "Bem-definidos" se referem ao facto de o conjunto de dados original estar em conformidade com regras pré-definidas. Por exemplo, os dados derivados de bases de dados relacionais tendem a ser mais bem-definidos. Anonimizar conjuntos de dados que não sejam bem-definidos pode provocar desafios adicionais à anonimização, e está fora do âmbito desta Guia.
- "Em formato textual" se referem a texto, números, datas, etc., ou seja, dados alfanuméricos já em formato digital. As técnicas de anonimização para dados em fluxo como áudio, vídeo, imagem, megadados (no seu formato em bruto), geolocalização, dados biométricos, etc., criam desafios adicionais e requerem técnicas de anonimização completamente diferentes, as quais estão fora do âmbito desta Guia.
- "De nível único" se referem a dados que relativos a indivíduos diferentes. Conjuntos de dados que contenham entradas múltiplas para os mesmos indivíduos (por exemplo, diferentes transacções efectuadas por um indivíduo) podem ainda utilizar algumas das técnicas explicadas nesta Guia, mas é necessário aplicar critérios adicionais; os quias estão fora do âmbito desta Guia.

2.7 Esta Guia destina-se a pessoas que estejam responsáveis pela protecção de dados dentro de uma organização, sem conhecimento ou experiência anterior em anonimização de dados. Os conhecimentos básicos de matemática serão necessários para entender alguma parte da terminologia e conceitos utlizados, e o entendimento básico de gestão de risco é necessário

na aplicação das técnicas.

2.8 Embora esta Guia pretenda ajudar as organizações à anonimização de dados pessoais, a Comissão reconhece que não existe uma solução única para todas as organizações. Cada organização deve, portanto, utilizar a abordagem de anonimização que seja apropriada para as suas circunstâncias. Alguns factores que as organizações podem ter em conta quando decidem sobre técnicas de anonimização a serem aplicadas incluem:

- A natureza e o tipo de dados pessoais que a organização pretende anonimizar, pelo que diferentes técnicas são adequadas a diferentes tipos de dados e circunstâncias;
- No contexto da gestão de risco, controlos impostos pela organização para proteger os dados anonimizados, além das técnicas de anonimização;
- A utilidade requerida dos dados anonimizados (*vide* secção 4 sobre os conceitos de anonimização).

3. Terminologia

3.1 Devido à variação dos termos e significados utilizados na literatura do tópico de anonimização de dados, esta secção explica o significado de alguns termos-chave usados nesta Guia.

Termo	**Significado na Guia**
Adversário	Um elemento que tente re-identificar individuo(s) através de conjunto de dados que é suposto estar anonimizado.
Anonimização	A conversão de dados pessoais em "dados anonimizados" através da aplicação de um conjunto de técnicas de anonimização. (Esta Guia foca-se apenas nos aspectos técnicos desta conversão)
Conjunto de dados anonimizado	O conjunto de dados resultante após as técnicas de anonimização terem sido aplicadas em combinação com a avaliação de risco adequada.
Atributo	Também chamado de campo de dados, coluna de dados, ou variável. Uma informação que pode ser encontrada nosregistos do conjunto de dados. Nome, género e endereço são exemplos de atributos.

Conjunto de dados	Uma colecção de registo de dados. Conceptualmente semelhante a uma tabela numa base de dados relacional ou folha de cálculo típica, tendo registos (linhas) e atributos (colunas).
Identificador directo	Um atributo de dados que por si só identifica um indivíduo (ex. impressão digital) ou foi atribuído a um indivíduo (ex. número de identificação nacional)
Classe de equivalência	Os registos num conjunto de dados que partilham os mesmos valores com certos atributos, tipicamente identificadores indirectos.
Identificabilidade vs. Re-identificabilidade	O grau ao qual um indivíduo pode ser identificado de um ou mais conjuntos de dados que contêm identificadores directos e indirectos, vs. o grau ao qual um indivíduo pode ser identificado a partir de conjuntos de dados anonimizados.
Identificador indirecto	Também chamado de quási-identificador. Um atributo de dados que, por si só, não identifica um individuo, mas pode fazê-lo em combinação com outra informação.
Não-identificador	Conjuntos de dados que podem conter atributos de dados que não são categorizados como identificadores directos nem indirectos. Tais atributos não precisam de ser sujeitos a anonimização (Nota-se que os exemplos fornecidos nesta guia não incluem este tipo de atributos, mas isto não significa que não podem fazer parte dos dados anonimizados).
Conjunto de dados original	O conjunto de dados antes de qualquer técnica de anonimização ser aplicada.
Pseudonimização[2]	A técnica de substituir um identificador com um valor não relacionado, mas ainda assim tipicamente único Ex. substituir "Joshua Quek"

[2] Alguma literatura (ex. "Parecer 05/2014 sobre técnicas de anonimização" pelo Grupo de Trabalho de Protecção de Dados do Artigo 29.º) enfatiza o risco da utilização de pseudónimos como técnica de anonimização. Nesta Guita, a pseudonimização não está excluída das técnicas de anonimização, porque pode ainda assim servir o seu propósito quando aplicada de forma cuidadosa.

	por "274927473"
Registo	Também denominado de linha. Um grupo de informação, tipicamente relacionada com o sujeito (ex. um individuo) ou transacção.
Re-identificação	Identificar uma pessoa de um conjunto de dados anonimizado. A re-identificação espontânea refere-se à re-identificação não intencional devido ao domínio de conhecimento particular sobre indivíduos.

Notas adicionais sobre a terminologia

3.2 O capítulo 5 das Orientações Aconselhadas sobre os Conceitos-Chave da PDPA clarifica o que são "identificadores". Essas Orientações usam o termo "identificador único", o que equivale ao termo "identificador directo" utilizado nesta Guia. O termo "identificador directo" é utilizado em vez de "identificador único" nesta Guia, pelo que o primeiro é habitualmente usado mais na área de anonimização de dados.

3.3 As Orientações Aconselhadas não fornecem um termo específico equivalente a "identificador indirecto", mas explicam com base no exemplo que "ainda que cada um destes pontos de dados, por si só, não seria suficiente para identificar um indivíduo", a organização "deve manter em mente que o conjunto de dados" (ex. os pontos de dados em combinação) "pode permitir identificar um individuo". Também esclarece que "desde que qualquer combinação de dados contenha um identificador único de um individuo, essa combinação de dados vai constituir dados pessoais".

3.4 Nota-se também que não existe nenhum termo comum na literatura de anonimização típica para descrever o terceiro tipo de dados, chamado de "não-identificadores" nestea Guia. Estes não-identificadores não seriam considerados Dados Pessoais se estivessem isolados de qualquer identificador directo e indirecto (ex. nem todos os dados são necessariamente Dados Pessoais). Mas assim que sejam ligados a identificadores directos e indirectos, precisam de ser protegidos e tratados tal como dados pessoais. Desde que o uso ou a aparência de tais dados dentro de um conjunto de dados anonimizado não viole qualquer outra das obrigações da PDPA, não precisam de ser ainda mais anonimizados, porque não permitiriam identificar um indivíduo.[3]

[3] Por exemplo: Um vendedor de carros, para o objectivo de utilizar IA e *Machine Learning*, tem

3.5 Não é o objectivo desta Guia definir quais dos três tipos são Dados Pessoais nos termos da PDPA e quais não são, mas para discutir as técnicas de anonimização, esta distinção adicional é importante, e por isso esta Guia segue a terminologia comum na literatura de anonimização utilizando identificadores "directos" e "indirectos", e onde "pontos de dados" são denominados campos de dados ou atributos.

3.6 De modo semelhante, deve ser notado que esta Guia não diferencia entre "dados" e "metadados"; as técnicas podem (e onde necessário, devem) ser aplicadas aos metadados e a outros tipos de dados também. Porém, a anonimização de um tipo específico de metadados dentro do conjunto de dados em si, nomeadamente o nome do cabeçalho em folhas de cálculo ou marcadores em ficheiros XML, não é discutida, dado que apenas poucas técnicas se aplicam a esse tipo de dados.

PARTE 2: CONTEXTO

4. Conceitos de Anonimização de Dados

4.1 A anonimização de dados requer um bom entendimento dos seguintes elementos, os quais devem ser tidos em consideração ao determinar as técnicas de anonimização adequadas e um nível apropriado de anonimização:

a. **Objectivo da anonimização e utilidade**: O objectivo da anonimização deve ser claro, porque a anonimização deve ser feita especificamente para o objectivo em causa. O processo de anonimização, independentemente das técnicas aplicadas, reduz a informação original no conjunto de dados em certa medida. Por isso, geralmente, à medida que a extensão da anonimização aumenta, a utilidade (ex. clareza e/ou precisão) do conjunto de dados é reduzida. Assim, a organização precisa de decidir sobre o grau de *trade-off* entre a utilidade aceitável (ou esperada) e a tentativa de reduzir o risco de re-identificação – onde o objecto dos dados é identificado a partir de dados que supostamente estariam anonimizados.

Nota-se que a utilidade não deve ser medida no nível do conjunto inteiro

---

registos dos clientes muito detalhados, ex. incluindo a cor do carro adquirido e o ano de produção dos pneus. O produtor de carros quer determinar qual é a cor do carro que deve produzir em maior quantidade. Após anonimizar (ex. suprimir) os identificadores directos e indirectos, o vendedor pode partilhar os dados resultantes (ex. contendo o género do comprador, a cor do carro e a data de produção dos pneus, etc.) com o fabricante de carros sem a necessidade de aplicar mais técnicas de anonimização (ex. não é necessário generalizar a data de produção dos pneus ou *k*-anonimizar o conjunto de dados). Porém, o vendedor de carros apenas pode proceder deste modo quando, entre outras coisas, é estabelecido que: a) os dados remanescentes em geral não constituem identificadores directos ou indirectos, b) o uso e a partilha destes dados em bruto não contradiz alguma das outras obrigações (ex. consentimento), c) registos individuais, específicos (ex. cores de carro de produção personalizada) são removidos.

de dados, mas é tipicamente diferente para diferentes atributos; um extremo é que um atributo específico é um item de interesse principal e nenhuma técnica de generalização/anonimização deve ser aplicada (ex. devido à precisão dos dados ser crucial), onde o outro extremo pode ser que um certo atributo não tem qualquer utilidade para o objectivo pretendido e pode ser removido totalmente sem que isso afecte a utilidade dos dados ao receptor.

Outra consideração importante em termos de utilidade é se coloca risco adicional se o recipiente souber qual técnica de anonimização e qual o nível de resolução que foi aplicado; por um lado pode ajudar o analista a entender os resultados melhor ou interpretá-los melhor, mas por outro lado pode contar pistas que podem levar a um risco elevado de re-identificação (porém, alguns desfechos simplesmente não podem mascarar a sua baixa resolução, ex. *k*-anonimato).

b. **Características das técnicas de anonimização**: As diferentes características das várias técnicas de anonimização significam que certas técnicas podem ser mais adequadas para uma situação do que outras. Por exemplo, certas técnicas (ex. encobrimento de caracteres) são habitualmente utilizadas em identificadores directos e outros (ex. agregação) para identificadores indirectos. Outro exemplo é considerar se o valor dos atributos é um contínuo ou discreto (ex. "sim" ou "não"), porque as técnicas como perturbação de dados funcionam melhor com dados contínuos.

As várias técnicas de anonimização também modificam os dados de modos significativamente diferentes. Alguns modificam apenas partes de um atributo (ex. encobrimento de caracteres); outros substituem o valor de um atributo em vários registos (ex. agregação); alguns substituem o atributo com informação não relacionada, mas consistente (ex. pseudonimização); e alguns removem o atributo totalmente (ex. supressão de atributos).

As várias técnicas de anonimização podem ser combinadas. Ex. suprimir ou remover registos (nas extremidades) após a generalização estar concluída.

c. **Informação inferida**: Pode ser possível inferir certa informação a partir de dados anonimizados. Ex. encobrir caracteres pode ocultar dados pessoais, mas não oculta o comprimento dos dados originais em termos do número de caracteres.

O problema da interferência não se limita a um atributo único, mas também entre atributos, mesmo que a todos tenham sido aplicadas técnicas de anonimização. O processo de anonimização tem, por isso, de registar todas as possibilidades, tanto antes de decidir a técnica em si como depois de aplicar as técnicas.

A abordagem também pode ter interesse em considerar em que ordem são apresentados os dados anonimizados: se o receptor saiba que os registos de dados foram recolhidos por ordem (ex. visitantes por ordem de chegada), pode ser prudente (desde que não afecte a utilidade) reorganizar o conjunto de dados com base na ordem do registo de dados.

d. **Conhecimento sobre o assunto em causa**: As técnicas de anonimização essencialmente reduzem a "identificabilidade" de um ou mais indivíduos do conjunto de dados original para um nível aceitável pelo portfolio de risco da organização.

Uma avaliação de "identificabilidade" deve ser realizada antes e depois de técnicas de anonimização serem aplicadas, e isto requer um bom entendimento do assunto ao qual os dados se referem. A avaliação antes do processo de anonimização assegura que a estrutura e a informação dentro de um atributo são claramente identificadas e compreendidas, e o risco de inferência implícita e explicita de tais dados é avaliada; ex. um atributo que contenha o ano de nascimento implicitamente fornece a idade, à semelhança do número de identificação nacional. A avaliação após a anonimização irá determinar o risco residual de re-identificação. Assim, se o conjunto de dados for dados de saúde, é provável que necessite de alguém com conhecimentos de saúde suficientes para avaliar o quão único (ou seja, o quão identificável) um registo é.

Outro exemplo é que um conjunto de dados sintético é criado ou os atributos dos dados são trocados entre registos. Isto requer alguém com conhecimento da área para reconhecer se os dados anonimizados sequer fazem sentido.

A opção correcta de técnicas de anonimização depende por isso da consciência da informação explícita e implícita contida nos dados e a quantidade ou o tipo de informação que se pretende anonimizar.

e. **Competência no processo de anonimização e as suas técnicas**: A anonimização é complexa. Além de ter conhecimento do assunto em questão (como explicado em cima), as organizações que pretendam partilhar dados anonimizados devem também assegurar que o processo de anonimização em si é realizado por pessoas familiarizadas com técnicas e princípios de anonimização. Se o conhecimento necessário não for encontrado dentro da organização, deve-se procurar ajuda externa.

f. **O receptor**: Factores como o conhecimento do receptor sobre o assunto em questão, controlos implementados para limitar os receptores e para evitar que os dados sejam partilhados com partes não autorizadas, desempenham um papel importante na escola de técnica de anonimização. Em particular, o uso expectável dos dados anonimizados pelo receptor

pode impor limitações nas técnicas aplicadas, porque a utilidade dos dados pode ser perdida além de limites aceitáveis. É necessário usar muita precaução ao fazer uma publicação dos dados, e o mesmo requer uma forma muito mais forte de anonimização quando comparada com dados partilhados sob um formato contratual.

g. **Ferramentas**: Devido à complexidade e computação necessária, as ferramentas de *software* podem ser muito úteis como apoio na execução de técnicas de anonimização. Existem algumas ferramentas[4] dedicadas disponíveis, mas a Guia não fornece qualquer avaliação ou recomendação de ferramentas de anonimização e re-identificação. Nota-se que mesmo as melhores ferramentas precisam de entradas adequadas (ex. parâmetros adequados para ser utlizados), ou podem conter limitações, daí que a supervisão e a familiaridade do elemento humano com as ferramentas e dados continuam a ser necessárias.

5. Riscos de divulgação

5.1 Existem vários riscos de divulgação. Esta secção explica alguns que são fundamentais, para facilitar uma discussão aprofundada na anonimização de dados.

- Divulgação de identidades (re-identificação): determinar, com alto nível de confiança, a identidade de um indivíduo descrita por um registo específico. Isto pode surgir de cenários tais como anonimização insuficiente, re-identificação por ligação, ou inversão de pseudónimos. Ex. um processo de anonimização que cria pseudónimos com base num algoritmo fácil de adivinhar e reversível tais como substituir '1' por 'a', '2' por 'b' e daí em diante.

- Divulgação de atributo: determinar, com alto nível de confiança, que um atributo descrito no conjunto de dados pertence a um individuo específico, mesmo que o registo do individuo não possa ser distinguido. Ex. um conjunto de dados contendo registos de clientes anonimizados de um cirurgião plástico específico revela que todos os seus clientes abaixo dos 30 anos realizaram um certo procedimento. Se se sabe que um indivíduo particular tem 28 anos e é cliente deste cirurgião, sabemos que este realizou aquele procedimento, mesmo que o seu registo não possa ser distinguido dos outros nos dados anonimizados.

- Divulgação de inferências: realizar uma inferência, com alto nível de confiança, sobre um indivíduo mesmo que este não esteja no conjunto de dados, pelas propriedades estatísticas do mesmo. Ex. se um conjunto de

[4] Ferramentas de anonimização incluem ARGUS, sdcMicro, ARX, *Privacy Analytics Eclipse*, Arcad DOT-Anonymizer

dados libertado por um investigador médico revela que 70% dos indivíduos com idade acima de 75 tem uma certa condição médica, esta informação pode ser inferida sobre um individuo que nem sequer consta no conjunto de dados.

5.2 Em geral, a maioria das técnicas de anonimização pretende proteger contra a divulgação de identidade e não necessariamente outros tipos de divulgação.

PARTE 3: TÉCNICAS BÁSICAS DE ANONIMIZAÇÃO

6. Supressão de atributos

6.1 **Descrição**: A supressão de atributos refere-se à remoção de uma secção inteira dos dados (também denominada de "coluna" em bases de dados e folhas de cálculo) no conjunto de dados.

6.2 **Quando usar**: Quando um atributo não é necessário no conjunto de dados anonimizado, ou quando o atributo não pode, de outra forma, ser anonimizado adequadamente com outra técnica. Esta técnica deve ser aplicada no início do processo de anonimização, dado que é uma forma fácil de diminuir a identificabilidade nesta fase.

6.3 **Como usar**: Apagar (ou seja, remover) o(s) atributo(s), ou se a estrutura do conjunto de dados tiver de ser mantida, limpa os dados (e possivelmente o cabeçalho). Nota-se que a supressão deve ser de facto remoção (por exemplo, permanente), e não apenas "esconder a coluna"[5]. Do mesmo modo, "rescrever" pode não ser suficiente se os dados subjacentes permaneçam relativamente acessíveis.

Outras observações:

6.4 Esta é a técnica de anonimização mais forte, porque não há forma de recuperar informação de um atributo sujeito à mesma.

6.5 Em certos casos, pode ser possível criar um "atributo derivado" que fornece utilidade, mas ainda assim é menos sensível que o(s) atributo(s) original(is) que podem deste modo ser suprimidos. Ex. criar um atributo "duração nas instalações", com base nos atributos "data & hora de entrada" e "data e hora de saída".

6.6 **Exemplo**

Neste exemplo, o conjunto de dados consiste em pontuações de teste. Pelo que o receptor apenas precisa de analisar as pontuações obtidas pelos estudantes relativamente aos seus vários treinadores mas sem a análise dos estudantes em

[5] Encontrado em *Software* de folha de cálculo.

si, o atributo "estudante" foi removido.

Antes da anonimização:

Estudante	Treinador	Pontuação
John	Tina	87
Yong	Tina	56
Ming	Tina	92
Poh	Huang	83
Linnie	Huang	45
Jake	Huang	67

Após a supressão do atributo "estudante":

Treinador	Pontuação
Tina	87
Tina	56
Tina	92
Huang	83
Huang	45
Huang	67

7. Supressão do registo

7.1 **Descrição**: A supressão de registo refere-se à remoção de um registo inteiro do conjunto de dados. Em contraste com a maioria das técnicas, esta afecta vários atributos ao mesmo tempo.

7.2 **Quando usar**: Para remover registos nos extremos que são únicos ou não satisfazem outros critérios como *k*-anonimato, e não os manter no conjunto de dados anonimizado. Os extremos podem levar a uma fácil re-identificação. Pode ser aplicada antes ou depois de outras técnicas (ex. generalização) terem sido aplicadas.

7.3 **Como usar**: Apagar um registo na íntegra. Nota-se que a supressão deve ser permanente e não simplesmente a função "esconder linha"[6]; do mesmo modo, "rescrever" pode não ser suficiente se os dados subjacentes permaneçam relativamente acessíveis.

Outras observações:

[6] Encontrado em *Software* de folha de cálculo.

7.4 *Vide* o exemplo na secção em relação a generalização sobre a ilustração de como a supressão de registos é usada.

7.5 Nota-se que a remoção de um registo pode influenciar o conjunto de dados, ex. em termos de estatísticas tais como média e mediana.

8. Encobrimento de Caracteres

8.1 **Descrição**: O encobrimento de caracteres é uma alteração de caracteres num valor dos dados, ex. usando um símbolo constante ("*" ou "x"). O encobrimento é tipicamente parcial, ou seja, aplicado somente a alguns caracteres no atributo.

8.2 **Quando usar**: No caso de o valor do dado ser uma cadeia de caracteres e ocultar parte dos mesmos ser suficiente para fornecer o grau de anonimidade necessário.

8.3 **Como usar**: Dependendo da natureza do atributo, substituir os caracteres adequados com o símbolo escolhido. Dependendo do tipo de atributo, é possível decidir substituir um número fixo de caracteres (ex. em números de cartão de crédito), ou um número variável de caracteres (ex. para endereços de e-mail)

Outras observações:

8.4 Nota-se que o encobrimento pode ter em conta se o comprimento dos dados originais fornece informação sobre os mesmos. É importante ter conhecimento do assunto em causa, sobretudo para que o encobrimento parcial garanta que os caracteres certos são encobertos. Pode ser necessário dar consideração especial às somas de controlo dentro dos dados; por vezes a mesma pode ser utilizada para recuperar (outras partes de) os dados encobertos. Para um encobrimento completo, o atributo pode alternativamente ser suprimido, a não ser que o comprimento dos dados seja de alguma relevância.

8.5 O cenário de encobrir dados de tal modo que se pretenda que os sujeitos dos dados reconheçam os seus próprios dados é um caso especial, e não se enquadra nos objectivos habituais de anonimização de dados. Um exemplo disto é a publicação de resultados de sorteios, onde tipicamente são publicados os nomes e os números de identidade nacional parcialmente encobertos para que os indivíduos possam reconhecer que são eles os vencedores. Nota-se que geralmente, os dados anonimizados *não* devem ser reconhecíveis para os próprios sujeitos.

8.6 **Exemplo**

Este exemplo mostra uma mercearia *online* a realizar um estudo sobre procura através de suas entregas de dados históricos, com vista a melhorar a eficiência

operacional. A empresa encobriu os últimos 4 dígitos dos códigos postais, deixando os primeiros 2 dígitos, os quais correspondem ao "código de sector" dentro de Singapura.

Antes da anonimização:

Código Postal	Horário preferido de entrega	N.º médio de encomendas por mês
100111	20h00 a 21h00	2
200222	11h00 a 12h00	8
300333	14h00 a 15h00	1

Após encobrimento parcial do código postal:

Código Postal	Horário preferido de entrega	N.º médio de encomendas por mês
10xxxx	20h00 a 21h00	2
20xxxx	11h00 a 12h00	8
30xxxx	14h00 a 15h00	1

9. Pseudonimização

Descrição:

9.1 A substituição de dados identificadores com valores inventados. A pseudonimização também é denominada de codificação. Os pseudónimos podem ser irreversíveis, onde os valores originais são adequadamente removidos e a pseudonimização foi feita de forma não-repetível, ou reversível (pelo dono dos dados originais), no qual os valores originais são mantidos de forma segura, mas podem ser recolhidos e ligados de volta ao pseudónimo, quando a necessidade existir[7].

9.2 Pseudónimos persistentes permitem a ligação utilizando os mesmos valores de pseudónimo para apresentar o mesmo individuo por conjuntos de dados diferentes. Por outro lado, pseudónimos diferentes podem ser usados para apresentar o mesmo individuo em conjuntos de dados diferentes para prevenir a ligação de conjuntos de dados diferentes.

9.3 Os pseudónimos podem ser gerados aleatoriamente ou de forma determinista.

9.4 **Quando usar**: Quando os valores dos dados precisam de ser distinguidos e onde nenhum caracter ou outra informação implícita do original deve

[7] Por exemplo, na eventualidade de um estudo de investigação, os resultados teriam a capacidade de dar um aviso útil a um sujeito.

ser mantida.

9.5 **Como usar**: Substituir os respectivos valores dos atributos com valores inventados. Uma maneira de o fazer é pré-gerar uma lista de valores inventados e de seguida escolher um da lista aleatoriamente para substituir cada um dos valores originais. Os valores inventados devem ser únicos e não devem ter relação com os valores originais (de tal modo que não seja possível derivar os valores originais dos pseudónimos).

Outras observações:

9.6 Ao distribuir pseudónimos, é garantido que não se reutiliza pseudónimos que já tenham sido utilizados (especialmente quando são gerados ao acaso). Evitar usar o mesmo gerador de pseudónimos sobre vários atributos, sem uma alteração (ex. pelo menos utilizar uma lista aleatória diferente).

9.7 Pseudónimos persistentes tipicamente fornecem melhor utilidade ao manter a integridade referencial entre conjuntos de dados.

9.8 Para pseudónimos reversíveis, a base de dados de identidade não pode ser partilhada com o receptor, deve ser mantida em segurança e apenas pode ser utilizada pela organização para resolver quaisquer questões específicas (porém, o número destas questões deve ser controlado, senão as mesmas podem ser utilizadas para "descodificar" o pseudónimo).

9.9 Do mesmo modo, se for utilizada encriptação, a chave de encriptação não deve ser partilhada, e de facto deve ser protegida de acessos não autorizados, dado que uma fuga deste tipo de chave pode resultar numa violação de dados que permite a reversão da encriptação. O mesmo aplica-se a geradores de números pseudoaleatórios, os quais requerem um algoritmo. A segurança de qualquer chave utilizada deve ser garantida como com qualquer processo reversível de encriptação[8].

9.10 Se for utilizada encriptação, rever o método de encriptação (ex. algoritmo e comprimento da chave) periodicamente, para assegurar que é reconhecida pela indústria como sendo relevante e segura.

9.11 Em alguns casos, os pseudónimos podem ter de seguir uma estrutura ou tipo de dados do valor original (ex. para pseudónimos serem utilizáveis em aplicações de *software* ou simplesmente para parecerem semelhantes ao atributo original); nestes casos, geradores especiais de pseudónimos podem ser necessários para criar bases de dados sintéticas, ou em alguns casos pode se considerar "encriptação preservadora do formato" , a qual

[8] Nota-se que confiar num "segredo" particular de processo de reversão (com ou sem chave) é provavelmente mais sujeito a descodificação e sob risco de ser quebrado do que confiando numa chave padronizada de encriptação.

cria pseudónimos que têm o mesmo formato dos dados originais.

9.12 **Exemplo**

Este exemplo mosta a pseudonimização a ser aplicada aos nomes de pessoas que obtiveram as suas cartas de condução e a alguma informação sobre eles. Neste exemplo, os nomes foram substituidos por pseudónimos em vez do atributo ser suprimido, porque a organização pretendia ter a capacidade de reverter a pseudonimização se for necessário.

Antes da anonimização:

Pessoa	Resultado na pré-avaliação	Horas de aulas antes de aprovação
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

Após pseudonimização do atributo "pessoa":

Pessoa	Resultado na pré-avaliação	Horas de aulas antes de aprovação
416765	A	20
562396	B	26
964825	C	30
873892	D	29
239976	B	32
943145	A	25

Para pseudonimização reversível, a base de dados da identidade é guardada em segurança no caso de haver uma razão legitima no futuro para identificar indivíduos. Os controlos de segurança (incluindo os técnicos e administrativos) também devem ser usados para proteger a base de dados de identidade.

Base de dados de identidade (codificação única):

Pseudónimo	Pessoa

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

9.13 **Exemplo**

Para segurança adicional relativamente à base de dados de identidades, a dupla codificação pode ser utilizada. Continuando do exemplo anterior, este exemplo mostra a base de dados de ligação adicional, a qual é alojada com um terceiro de confiança. Com dupla codificação, a identidade dos indivíduos apenas pode ser descoberta quando o terceiro de confiança (estando na posse da base de dados de ligação) e a organização (tendo a base de dados de identidade) juntam as duas bases de dados.

Após anonimização:

Pessoa	Resultado na pré-avaliação	Horas de aulas antes de aprovação
373666	A	20
594824	B	26
839933	C	30
280074	D	29
746791	B	32
785282	A	25

A base de dados de ligação (mantida em segurança por um terceiro de confiança; mesmo a organização irá removê-la, eventualmente. O terceiro não recebe qualquer informação adicional.)

Pseudónimo	Pseudónimo intermédio
373666	OQCPBL
594824	ALGKTY
839933	CGFFNF
280074	BZMHCP
746791	RTJYGR

785282	RCNVJD

Base de dados de identificação(mantida em segurança pela organização)

Pseudónimo intermédio	Pessoa
OQCPBL	Joe Phang
ALGKTY	Zack Lim
CGFFNF	Eu Cheng San
BZMHCP	Linnie Mok
RTJYGR	Jeslyn Tan
RCNVJD	Chan Siew Lee

Nota: tanto na base de dados de ligação como na de identidade, é boa prática trocar a orgem dos registos em vez de os manter na mesma ordem que o conjunto de dados. Neste exemplo, os dois foram deixados na mesma ordem para facilidade de visualização.

10. Generalização

10.1 **Descrição**: uma redução deliberada na precisão dos dados. Ex. converter a idade de uma pessoa numa faixa etária, ou de um ponto preciso para um ponto menos preciso. Esta técnica também é denominada de recodificação.

10.2 **Quando aplicar**: para valores que possam ser generalizados e ainda assim sejam uteis para o objectivo pretendido.

10.3 **Como aplicar**: Desenhar categorias de dados apropriadas e regras para a tradução de dados. Considerar suprimir registos que se destaquem depois da tradução (ou seja, a generalização)

Outras observações:

10.4 Conceber as faixas etárias em dimensões apropriadas. Faixas etárias que sejam demasiado grandes podem implicar que os dados podem ser muito "modificados", enquanto que as faixas muito estreitas podem significar que os dados estão pouco alterados e por isso ainda fáceis de re-identificar. Se o *k*-anonimato for utilizado, o valor *k* escolhido irá afectar os alcances de dados também. Nota-se que a primeira e a última categoria podem ter um alcance maior para acomodar o número tipicamente mais baixo de registos nestes pontos; isto é frequentemente referido de codificação do topo/base.

10.5 **Exemplo**

Neste exemplo, este conjunto de dados contém nome de pessoa (o qual já foi

pseudónimizado), idade e endereço de residencia.

Antes da anonimização:

S/n	Pessoa	Idade	Endereço
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

Em relação à idade, a abordagem escolhida é generalizar para as seguintes faixas etárias.

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

Quanto ao endereço, uma abordagem possível é remover o bloco/número de porta e deixar apenas o nome da rua.

Após a generalização de Idade e Endereço:

S/n	Pessoa	Idade	Endereço
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 Payo 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

É suposto que exista, de facto, apenas uma unidade residencial em *Stonehenge Road*, como um exemplo, o endereço exacto pode assim ser deduzido, ainda que os dados tenham passado pela generalização. Isto pode ser considerado como sendo ainda "demasiado único".

Por isso, como o passo seguinte da generalização, o registo n.º 6 pode ser removido (ou seja, utilizando a técnica de supressão), pelo que o endereço ainda é demasiado único após remover número da unidade. Em alternativa, todos os endereços podem ser generalizados por uma extensão maior (ex cidade ou distrito), de tal modo que a supressão não seja ncessária, mas isto pode afectar a utilidade dos dados muito mais que suprimir alguns registos do conjunto de dados.

11. Troca

11.1 **Descrição**: O objectivo de trocar é reorganizar os dados no conjunto de dados de tal forma que os valores dos atributos individuais ainda estejam apresentados no conjunto, mas geralmente não correspondem ao registo original. Esta técnica também é denominada de baralhamento e permutação.

11.2 **Quando usar**: Quando uma análise subsequente apenas necessita de olhar para os dados no agregado, ou a análise se dá ao nível do atributo; por outras palavras, não é necessária a análise das relações entre atributos a nível do registo.

11.3 **Como usar**: Primeiro, identificar quais atributos serem alvo de trocar. Depois, para cada um, trocar ou reatribuir os valores do atributo para qualquer registo no conjunto de dados.

11.4 **Outras observações**: Avaliar e decidir quais atributos (colunas) precisam de ser trocadas. Dependendo da situação, as organizações podem decidir que, por exemplo, apenas atributos (colunas) que contém valores relativamente identificativos precisam de ser trocados.

11.5 **Exemplo**

Neste exemplo, o conjunto de dados contém informação sobre registos de clients para uma organização de negócios.

Antes da Anonimização:

Pessoa	Título	Data de Nascimento	Tipo de Membro	Visitas Mensais em Média
A	Reitor de universidade	3 Jan 1970	Prata	0
B	Vendedor	5 Fev 1972	Platina	5
C	Advogado	7 Mar 1985	Ouro	2
D	Informático	10 Abr 1990	Prata	1
E	Enfermeira	13 Mai 1995	Prata	2

Neste exemplo, todos os valores de todos os atributos foram trocados.

Pessoa	Título	Data de Nascimento	Tipo de Membro	Visitas mensais em Média
A	Advogado	10 Abr 1990	Prata	1
B	Enfermeira	7 Mar 1985	Prata	2
C	Vendedor	13 Mai 1995	Platina	5
D	Informático	3 Jan 1970	Prata	2
E	Reitor de universidade	5 Fev 1972	Ouro	0

Nota: Por outro lado, se o objectivo do conjunto de dados anonimizado for estudar as relações entre o perfil laboral e os padrões de consume, outros métodos de anonimização serão mais adequados, ex. por via da generalização dos títulos de emprego, o que pode resultar em que "Reitor de Universidade" seja alterado para "educador".

12. Perturbação de Dados

12.1 **Descrição**: os valores do conjunto de dados são modificados ligeiramente.

12.2 **Quando usar**: Para quási-identificadores (tipicamente números e datas), os quais podem ser potencialmente identificadores quando combinados com outras fontes de dados, e ligeiras mudanças nos valores sejam aceitáveis. Esta técnica não deve ser usada em casos que a precisão dos dados é crucial.

12.3 **Como usar**: Depende da técnica de perturbação a ser utlizada. Estas incluem arredondamento e adição de ruído aleatório. O exemplo nesta secção mostra arredondamento de base x.

Outras observações:

12.4 O grau de perturbação deve ser proporcional ao âmbito dos valores no atributo. Se a base for muito pequena, o efeito de anonimização será mais fraco; por outro lado, se a base for muito longa, os valores finais serão muito diferentes dos originais e a utilidade do conjunto de dados será provavelmente reduzida.

12.5 Nota-se que onde a computação for aplicada nos valores de atributos que tenham sido perturbados, o valor resultante pode sofrer perturbação numa escala ainda maior.

12.6 **Exemplo**

Neste exemplo, o conjunto de dados contem informação a ser usada para investigação sobre a possível ligação entre a altura, o peso, e a idade de uma pessoa, se a mesma fuma, e se tem a "doença A" e/ou a "doença B". O nome da pessoa já foi pseudonimizado.

Aplica-se o arredondamento seguinte:

Atributo	Técnica de anonimização
Altura (em *cm*)	Arredondamento de base 5 (5 é o número escolhido por ser relativamente proporcional aos valores típicos de altura, de, ex. 120 a 190 cm)
Peso (em *kg*)	Arredondamento de base 3 (3 é o número escolhido por ser relativamente proporcional aos valores típicos de peso, de, ex. 40 a 100 cm)
Idade (em anos)	Arredondamento de base 3 (3 é o número escolhido por ser relativamente proporcional aos valores típicos de idade, de, ex. 10 a 100 cm)
(atributos remanescentes)	Zero, devido a não serem numéricos e difíceis de modificar sem uma alteração substancial no valor

Conjunto de dados antes da anonimização:

Pessoa	Altura (cm)	Peso (kg)	Idade (anos)	Fumador?	Doença A?	Doença B?
198740	160	50	30	Não	Não	Não
287402	177	70	36	Não	Não	Sim
398747	158	46	20	Sim	Sim	Não
498732	173	75	22	Não	Não	Não
598772	169	82	44	Sim	Sim	Sim

Após anonimização (as colunas sombreadas representam os valores afectados):

Pessoa	Altura (cm)	Peso (kg)	Idade (anos)	Fumador?	Doença A?	Doença B?
198740	160	51	30	Não	Não	Não
287402	175	69	36	Não	Não	Sim
398747	160	45	18	Sim	Sim	Não
498732	175	75	21	Não	Não	Não
598772	170	81	42	Sim	Sim	Sim

Nota: para arredondamento de base-*x, os valores do atributo a ser arredondados são arredondados para o multiplo mais próximo de x.*

13. Dados sintéticos

13.1 **Descrição**: esta técnica é ligeiramente diferente quando comparada com as outras técnicas descritas nesta Guia, pelo que é principalmente usada para gerar conjuntos de dados sintéticos directamente e separá-los dos dados originais, em vez de modificar o conjunto original.

13.2 **Quando usar**: tipicamente, quando um grande volume de dados é necessário para o teste do sistema, mas os dados em concreto não podem ser utilizados e ainda assim os dados devem ser "realistas" em certos aspectos, como formato, relação entre atributos, etc.

13.3 **Como usar:** estudar os padrões do conjunto de dados original (os dados em si) e aplicar os padrões ao criar um conjunto de dados anonimizado (os dados sintéticos). O grau ao qual os padrões do conjunto original precisam de ser replicados depende de como o conjunto de dados anonimizado será usado.

Outras observações:

13.4 Dependendo do âmbito do teste e dos controlos administrativos, podem-se gerar dados total ou parcialmente sintéticos; ou seja, onde os testes são

realizados, e estes precisem de referenciar outros conjuntos de dados, então os poucos itens que estejam a ser testados têm de permanecer no seu formato original, mas o resto da informação pode ser sintético.

13.5 Enquanto que noutras técnicas os dados anonimizados são tipicamente do mesmo ou aproximadamente do mesmo volume que os dados originais (ex. quando a supressão ou a agregação são aplicadas), os dados sintéticos podem ser gerados em qualquer volume, conforme necessário.

13.6 Ao aplicar esta técnica, os extremos podem necessitar de atenção redobrada. Para finalidades de teste, as extremidades são frequentemente importantes, mas as extremidades em dados sintéticos também podem indicar extremidades específicas no conjunto de dados original. É por isso aconselhável criar extremidades nos dados sintéticos intencionalmente e independentes dos dados originais.

13.7 Esta técnica é de fraca utilidade para análise de dados, porque os dados não são "reais" e os dados foram criados com base num modelo pré-concebido.

13.8 **Exemplo**

Neste exemplo uma instalação de escritórios que se especializa em escritórios multifuncionais mantém o registo do tempo que os utilizadores começam e terminam a utilização das suas instalações. O objectivo é criar um conjunto de dados sintéticos para realizar teste de simulações sobre um algoritmo de alocação de novas instalações.

Uma discussão detalhada das medidas estatísticas vai além do âmbito da Guia, porém, neste exemplo, algumas medidas possíveis podem ser o valor médio ou mediano de utilizadores durante cada hora do dia.

Conjunto de dados original:

Utilizador	**Data**	**Entrada**	**Saída**
Utilizador A	1-Mar-17	8:27	18:04
Utilizador A	2-Mar-17	8:20	18:10
Utilizador B	1-Mar-17	8:45	17:17
Utilizador B	2-Mar-17	8:55	17:54
Utilizador C	1-Mar-17	13:18	15:48
Utilizador C	2-Mar-17	13:02	16:02
Utilizador D	1-Mar-17	17:55	7:31
Utilizador D	2-Mar-17	18:04	7:39

(etc.)	(etc.)	(etc.)	(etc.)

Estatísticas obtidas do conjunto de dados original

Início	**Fim**	**N.º Médio de Utilizadores**
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

Conjunto de dados sintético (para 1 dia):

Utilizador	**Data**	**Entrada**	**Saída**
100001	3-Abr-17	8:25	17:53
100002	3-Abr-17	8:00	18:04
100003	3-Abr-17	8:12	18:48
100004	3-Abr-17	8:49	18:02
100005	3-Abr-17	8:33	18:11
100006	3-Abr-17	8:37	18:05
100007	3-Abr-17	8:55	20:05
100008	3-Abr-17	8:23	18:34
100009	3-Abr-17	13:16	15:48
100010	3-Abr-17	13:03	15:11
100011	3-Abr-17	13:28	15:25
100012	3-Abr-17	13:18	15:32
100013	3-Abr-17	17:55	7:38

100014	3-Abr-17	18:04	7:32
100015	3-Abr-17	17:57	7:02
(etc.)	(etc.)	(etc.)	(etc.)

Nota: basciamente, o conjunto de dados sintético é criado com base em estatísticas derivadas do conjunto original, ex. o número médio de utilizadores no escritório em diferentes períodos temporais do dia.

14. Agregação de dados

14.1 **Descrição**: converter um conjunto de dados de uma lista para valores resumidos.

14.2 **Quando usar**: quando os registos individuais não são necessários e os dados agregados são suficientes para a finalidade.

14.3 **Como usar**: uma discussão detalhada das medidas estatísticas está fora do âmbito desta Guia, porém, os métodos típicos incluem a utilização de totais ou médias, etc. Pode também ser útil discutir com o receptor dos dados sobre a utilidade esperada para encontrar um compromisso adequado.

Outras observações:

14.4 Onde aplicável, deve ser exercida precaução para grupos que tenham poucos registos após realizar a agregação. Ex. no exemplo abaixo, se os dados agregados incluem um único registo em cada uma das categorias, pode ser fácil para alguém com algum conhecimento adicional identificar um doador.

14.5 Por isso, a agregação pode ter de ser aplicada em combinação com a supressão. Algum atributo pode ter de ser removido, pelo que este contém pormenores que não podem ser agregados, e os novos atributos podem ter de ser adicionados, ex. para conter valores agregados calculados recentemente.

14.6 **Exemplo**

Neste exemplo, uma organização de caridade tem o registo das doações feitas, bem como algumas informações sobre os doadores.

A organização de caridade avaliou que os dados agregados são suficientes para um consultor externo realizar a análise dos dados, e por isso realizada a agregação dos dados no conjunto de dados original.

Conjunto de dados original:

Doador	**Salário mensal ($)**	**Quantia doada em 2016 ($)**

Doador A	4000	210
Doador B	4900	420
Doador C	2200	150
Doador D	4200	110
Doador E	5500	260
Doador F	2600	40
Doador G	3300	130
Doador H	5500	210
Doador I	1600	380
Doador J	3200	80
Doador K	2000	440
Doador L	5800	400
Doador M	4600	390
Doador N	1900	480
Doador O	1700	320
Doador P	2400	330
Doador Q	4300	390
Doador R	2300	260
Doador S	3500	80
Doador T	1700	290

Conjunto de dados anonimizado:

Salário mensal ($)	**N.º de donativos recebidos (2016)**	**Soma da quantia doada em 2016 ($)**
1000-1999	4	1470
2000-2999	5	1220
3000-3999	3	290
4000-4999	5	1520
5000-6000	3	870
Total	**20**	**5370**

PARTE 4: ELABORAÇÃO

15. Metodologia de Anonimização

15.1 Enquanto a Parte 3 desta Guia se focou nas várias técnicas básicas de

anonimização, esta requer mais do que simplesmente aplicar a(s) técnica(s) apropriada(s). A Parte 4 enquadra o panorama geral e discute considerações adicionais. É favor notar que esta descrição se foca apenas em divulgação não-pública; um modelo de divulgação pública pode necessitar de considerações adicionais e mais detalhadas.

15.2 O seguinte é uma metodologia sugerida para a realização de anonimização:

1) Determinar o modelo de divulgação.

Isto refere-se a como o conjunto de dados anonimizado será divulgado. *Público* refere-se a disponibilizar o mesmo para essencialmente qualquer pessoa. *Não-público* refere-se a uma divulgação controlada para receptores limitados (e frequentemente num número fixo). O modelo de divulgação a público impõe inerentemente mais desafios às técnicas de anonimização.

2) Determinar o nível de risco de re-identificação aceitável bem como a utilidade esperada e o nível de risco pretendido ou necessário.

Vide a secção 17 para mais pormenores. Nota-se que o limite de risco definido nesta fase deve ser claramente distinguido se os controlos adicionais forem tomados em conta ou se apenas reflectir o risco dos dados.

3) Classificar os atributos dos dados.

Isto prende-se com classificar os atributos no conjunto de dados como identificadores directos, indirectos ou não-identificadores, o que afecta como os mesmos serão subsequentemente processados.

4) Remover atributos de dados que não se utilizaram.

No processo de anonimização, a maioria dos atributos, quer identificador directo quer indirecto, requer tipicamente processamento ou pelo menos consideração, para que se torne em menos identificador. Por isso, qualquer atributo que não seja claramente necessário no conjunto de dados anonimizado deve ser suprimido.

5) Anonimizar identificadores directos e indirectos.

Isto é feito pela aplicação de técnicas tais como as descritas nesta Guia. As técnicas diferentes são aplicáveis para tipos de identificadores. Algumas técnicas podem (e muitas vezes devem) ser usadas em conjunto. Os registos extremos devem ser considerados para supressão.

6) Determinar o risco real e comparar o mesmo com o limite.

Vide a secção 17 para mais pormenores.

7) Realizar mais anonimização, se for necessário.

Se o risco real for mais alto que o limite, é necessário aplicar anonimização mais "forte" e os passos 5 a 7 devem ser realizados novamente com os ajustes necessários, até que o risco real seja mais baixo que o limite.

8) Avaliar a solução.

Isto inclui examinar o conjunto de dados anonimizado para avaliar se a utilidade vai ao encontro do objectivo. Se a utilidade for insuficiente, o processo de anonimização pode ter de ser redesenhado, ou pode ter de se considerar se a anonimização é exequível para este conjunto de dados.

9) Determinar os controlos necessários. ◦

Os controlos incluem tanto os técnicos como os não-técnicos (ex. medidas legais e organizacionais). Os controlos técnicos são descritos mais em pormenor na secção 18.

10) Documentar o processo de anonimização.

Os detalhes do processo de anonimização, os parâmetros e controlos usados devem ser claramente registados para referência futura. Tal documentação facilita a revisão, manutenção, aperfeiçoamento e auditoria. Nota-se que esta documentação deve ser mantida na maior segurança pelo que a divulgação dos parâmetros pode facilitar a re-identificação.

16. *K*-anonimato – uma medida do risco

16.1 *O K*-anonimato (e outras extensões semelhantes do mesmo como *L*-diversidade e *T*-proximidade) é muitas vezes considerado como técnica de anonimização, mas é mais uma medida utilizada para assegurar que a barreira do risco não é ultrapassada, como parte da metodologia de anonimização (ver em particular o passo 6).

16.2 O *K*-anonimato não é a única medida disponível nem está livre de limitações, mas é relativamente bem compreendido e fácil de aplicar. Os métodos alternativos como privacidade diferencial[9] tem vindo a emergir nos últimos anos.

16.3 **Descrição**: O modelo do *k*-anonimato é usado como orientação antes de e para a verificação, depois de técnicas de anonimização (ex. generalização) terem sido aplicadas, para assegurar que os identificadores directos e/ou indirectos do registo são partilhados por pelo menos *k*-1 outros registos.

Esta é a principal protecção fornecida pelo *k*-anonimato contra ataques de ligação, porque os registos *k* (ou pelo menos identificadores directos e

[9] Privacidade diferencial envolve vários conceitos, incluindo responder a consultas em vez de fornecer os dados anonimizados, adicionando ruído aleatório para proteger os registos individuais, providenciar garantias matemáticas para que o "orçamento de privacidade" pré-definido não seja excedido, etc.

indirectos diferentes) são idênticos nos atributos identificativos (e por isso criam uma "classe de equivalência" com os membros *k*), e por isso não é possível ligar ou isolar o registo de um indivíduo; existem sempre atributos *k* idênticos.

Um conjunto de dados anonimizado pode ter níveis de k-anonimidade para diferentes conjuntos de identificadores indirectos, mas para avaliar a protecção contra a ligação, o *k* mais baixo é utilizado para comparar com o limite.

16.4 **Quando usar**: para confirmar que as medidas de anonimização em efeito atingem o limite desejado contra ataques de ligação.

16.5 **Como usar**: Primeiro, é necessário decidir um valor para *k* (o que essencialmente é igual a ou maior que o inverso do tamanho da classe de equivalência), o que promove que o *k* mais baixo seja atingido entre todas as classes. Geralmente, quanto mais alto o valor de *k*, mais difícil é para os sujeitos dos dados serem identificados; porém, a utilidade pode-se tornar mais baixa à medida que o *k* aumenta e mais registos terem de ser suprimidos. Apos outras técnicas de anonimização serem aplicadas, deve-se verificar que cada registo tem pelo menos *k*-1 outros registos com os mesmos atributos abordados pelo k-anonimato. Os registos em classes de equivalência com menos de *k* registos devem ser considerados para supressão; em alternativa, pode-se realizar mais anonimização.

Outras observações:

16.6 Além da generalização e supressão, os dados sintéticos também podem ser criados (ex. perto dos extremos) para atingir o *k*-anonimato. Estas técnicas (e outras) podem por vezes ser usadas em conjunto, mas se nota que o método exacto escolhido pode afectar a utilidade dos dados. Deve-se considerar os compromissos entre eliminar os extremos ou inserir dados sintéticos.

16.7 *O K-anonimato* presume que cada registo se prenda com um individuo diferente. Se o mesmo individuo tiver registos diferentes (ex. visitas ao hospital em ocasiões diferentes), então o *k*-anonimato terá de ser mais alto do que os registos repetidos, caso contrário os mesmos podem ser associáveis, e também ser re-identificáveis do registo, apesar de cumprirem aparentemente as "classes de equivalência *k*".

16.8 **Exemplo**

Neste exemplo, o conjunto de dados contém informações sobre pessoas que apanham táxis.

É utilizado *K*=2, ou seja, cada registo deve eventualmente partilhar os mesmos

atributos com 1 outro registo, após a anonimização. Nota: *K*=2 é utilizado para simplificar o exemplo, mas é um valor provavelmente muito baixo para dados reais, porque isto significa que o risco de identificação é de 50%.

As seguintes técnicas de anonimização são utilizadas em combinação e o nível de resolução é um exemplo que permite atingir o nível *K* necessário.

Atributo	Técnica de anonimização
Idade	Generalização (intervalos de 10 anos)
Ocupação	Generalização – ex. Tanto "Administrador de base de dados" como "programador" são generalizados para "*IT*"
Supressão de Registo	Registos que não vão de encontro ao *k*-anonimato de após aplicação das técnicas de anonimização terem sido aplicadas (neste caso, generalização), são removidos. Ex. o banqueiro que é o único do seu tipo nos dados.

Conjunto de dados antes da anonimização:

Idade	Género	Ocupação	Número médio de viagens por semana
21	Feminino	Jurista	15
38	Masculino	Oficial de Privacidade de Dados	2
25	Feminino	Banqueiro	8
44	Feminino	Administrador de Base de Dados	3
25	Feminino	Assistente Administrativo	1
31	Masculino	Oficial de Privacidade de Dados	5
42	Feminino	Programador	3
22	Feminino	Assistente Administrativo	4
30	Feminino	Jurista	2

Conjunto de dados após a da idade e ocupação e supressão da extremidade (as respectivas classes de equivalência estão marcadas a cores diferentes):

Idade	Género	Ocupação	Número médio de viagens por semana
21 a 30	Feminino	Jurista	15
31 a 40	Masculino	Oficial de Privacidade de Dados	2
~~21 a 30~~	Feminino	Banqueiro	~~8~~
41 a 50	Feminino	Administrador de Base de Dados	3
21 a 30	Feminino	Assistente Administrativo	1
31 a 40	Masculino	Oficial de Privacidade de Dados	5
41 a 50	Feminino	Programador	3
21 a 30	Feminino	Assistente Administrativo	4
21 a 30	Feminino	Jurista	2

Nota: O número médio de viagens por semana é tomado aqui como exemplo para um não-identificador, sem a necessidade de se anonimizar mais este atributo. Também se nota que um conjunto de dados que siga k-anonimato sem outros não identificadores ou outros atributos pode ser simplificado ao remover os duplicados e simplesmente indicar o valor de k.

17. Avaliar o Risco de Re-identificação

17.1 Existem várias maneiras de avaliar o risco de re-identificação, e estes podem envolver cálculos relativamente complexos envolvendo cálculo probabilístico. *Vide* as publicações de referência no Anexo B para informação detalhada.

17.2 Esta secção descreve um modelo simplificado, utilizando *k*-anonimato[10], e garantindo certos pressupostos. Um dos pressupostos é que o modelo de divulgação não é público. O segundo pressuposto é que os ataques tentam ligar um individuo ao conjunto de dados anonimizado. O terceiro pressuposto é que o conteúdo dos dados anonimizados não é tido em

[10] Os cálculos seriam diferentes se fossem feitos utilizando, por exemplo, privacidade diferencial ou controlos de divulgação estatística tradicionais.

conta e que o risco é calculado independentemente de que tipo de informação o aggressor tem de facto disponível.

17.3 Primeiro, o limite de risco deve ser estabelecido. Este valor, reflectindo uma probabilidade, vai de 0 a 1. Reflecte o nível de risco que uma organização está disposta a aceitar. Os factores principais que afectam esta decisão devem incluir o dano que pode ser causado ao sujeito dos dados, bem como os danos à organização, no caso de a re-identificação suceder; mas também tem em conta quais outros controlos foram colocados a postos para mitigar o risco noutras formas além de anonimização. Quando mais alto o risco potencial, mais alto o limite de risco deve ser. Não devem existir regras demasiado rígidas para o que devem ser os valores do risco a utilizar; os valores seguintes servem apenas como exemplo:

Dano Potencial	Limite de Risco
Baixo	0,2
Médio	0,1
Alto	0,01

17.4 Ao calcular o risco real, esta Guia explica como ver o "Risco do Pesquisador", que presume que o adversário conheça uma pessoa no conjunto de dados e está a tentar estabelecer qual dos registos no conjunto se refere a essa pessoa.

17.5 A regra simples para calcular a probabilidade de re-identificação para um registo único no conjunto de dados é de tomar o inverso do tamanho da classe de equivalência do registo, ou seja

P (ligação individual a um registo único) = 1 / tamanho da classe de equivalência do registo

17.6 Agora, para calcular a probabilidade de re-identificação de qualquer registo no conjunto de dados na integra, mais uma vez, dado que existe uma tentativa de re-identificação, uma abordagem conservadora seria de equipará-la à probabilidade máxima de re-identificação entre todos os registos do conjunto de dados.

P (re-ID qualquer registo no conjunto) = 1 / tamanho mínimo da classe de equivalência do registo

Nota: se o conjunto está *k*-anonimizado,

P (re-ID qualquer registo no conjunto) <= 1 / k

17.7 Podemos considerar 3 cenários de re-identificação: (1) ataque interno deliberado; (2) reconhecimento inadvertido por um conhecido, e (3) fuga

de dados.

P (re-ID) = P (re-ID | tentative de re-ID) x P (tentative de re-ID)

Onde P (re-ID | tenativa de re-ID) refere-se à probabilidade de uma re-identificação bem sucedida, dado que existe uma tentativa de re-identificação. Como foi discutido antes, podemos tomar P (re-ID | tentativa de re-ID) como sendo (1 / tamanho mínimo da classe de equivalência do registo

Assim, P (re-ID) = (1 / tamanho mínimo da classe de equivalência do registo) x P (tentative de re-ID)

17.8 Para o cenário #1 – o ataque interno deliberado, presumemos que a parte que receba os dados tenta a re-identificação. Para estimar P (tentativa de re-ID), ou seja, a probabilidade de sucesso da tentativa de re-identificação, factores que podem ser considerados incluem a extensão dos controlos mitigantes em acção bem como os motivos e a capacidade do adversário. A tabela seguinte apresenta valores exemplo; mais uma vez é necessário para a parte anonimizadora decidir os valores adequados a utilizar.

P (tentativa de re-ID) para o cenário #1 – ataque interno deliberado		Motivação e Recursos do Adversário		
		Baixo	Médio	Baixo
Extensão dos Controlos Mitigantes	Alto	0,03	0,05	0,1
	Médio	0,2	0,25	0,3
	Baixo	0,4	0,5	0,6
	Nenhum	1,0	1,0	1,0

Factores que afectem a motivação e os recursos do adversário podem incluir:

- Vontade de violar o contrato (presumindo que um contrato que previna a re-identificação está em vigor)
- Restrições de tempo e financeiras
- Inclusão de personalidades de renome (ex. celebridades) nos dados

Factores que afectem a extensão dos controlos mitigantes inclui:

- Estruturas organizacionais
- Controlos administrativos (e.g. contratos)
- Medidas técnicas e físicas (vide secção 18)

17.9 Para o cenário #2 reconhecimento inadvertido por um conhecido, assumimos que a parte que recebe os dados re-identifica inadvertidamente o sujeito de dados ao examinar o conjunto dos dados.

Isto é possível porque a parte tem conhecimento adicional sobre o sujeito dos dados devido à sua relação (ex. amigo, vizinho, parente, colega, etc.). Para estimar P (tentativa de re-ID), ou seja, a probabilidade de uma tentativa de re-identificação, o factor principal a ser considerado é a probabilidade de que o receptor conheça alguém no conjunto de dados.

17.10 Para o cenário #3 – Uma fuga de dados que ocorra no sistema ICT do receptor dos dados, a probabilidade pode ser estimada com base nas estatísticas disponíveis da prevalência de fugas de dados na indústria do receptor. Isto baseia-se no pressuposto que os atacantes que obtém o conjunto de dados irão tentar a re-identificação.

Cenário #3 – fuga de dados P (tentativa de re-ID) = P (fuga de dados na indústria do receptor dos dados)

17.11 A probabilidade mais alta entre os 3 cenários deve ser usada como P (tentativa de re-ID).

P (tentativa de re-ID) = Max (P (ataque interno deliberado), P (reconhecimento inadvertido por um conhecido), P (fuga de dados))

17.12 Ao montar tudo,

P (re-ID) = (1 / Min. tamanho da classe de equivalência no conjunto) x P (tentative de re-ID) = (1 / k) x P (tentative de re-ID) *para um conjunto k-anonimizado* Onde P (tentativa de re-ID) = Max (P (ataque interno deliberado), P (reconhecimento inadvertido por um conhecido), P(fuga de dados))

18. Controlos Técnicos

18.1 Esta secção discute os controlos técnicos que podem ser implementados para reduzir ainda mais o risco de re-identificação após anonimização. Os controlos podem ou não ser adequados para a situação, dependendo da política de acesso ao conjunto de dados anonimizado. Onde for relevante, estes controlos podem geralmente ser implementados em combinação uns com os outros. Nota-se que alguns destes apenas são eficazes desde que o conjunto anonimizado com alto risco residual não seja passado para o receptor, já que assim que isto acontecer, os controlos técnicos deixam de ser possíveis, habitualmente. Nota-se ainda que uma abordagem com base no risco também deve ser tomada; logo, os controlos discutidos nesta secção são para consideração e não são de adopção obrigatória.

18.2 Acesso revogável – com registos dos acessos concedidos, pode ser possível, dependendo do tipo de controlo técnico utilizado, de revogar o

acesso onde for necessário. Habitualmente, isto é mais fácil de implementar onde o acesso ao conjunto de dados em questão é apenas feito *online*.

18.3 Só através de consulta – permitir a realização de consultas em vez de fornecer acesso directo aos dados. Um modo ainda mais seguro é que cada consulta seja votada por um curador que avalia se a consulta em si deve ser concedida.

18.4 Limite de receptores – Isto é frequentemente feito através de implementação de autenticação de utilizadores e autorização onde o acesso ao conjunto de dados ou à consulta é feito *online*, ou protecção por *password* ou encriptação onde o acesso é feito *offline*.

18.5 Controlos de Gestão dos Direitos Digitais (Digital Rights Management, DRM) – Isto é habitualmente feito proporcionando acesso online, mas implementando controlos adicionais tais como não permitir ao utilizador guardar ou imprimir os dados. Nota-se que existem limitações tais como não ser possível prevenir que os dados em exibição sejam fotografados.

18.6 Acesso *in loco* – Exigir que o utilizador esteja fisicamente presente no local onde o acesso ao conjunto de dados é disponibilizado, ou onde o acesso para realizar consultas é disponibilizado. A segurança adicional vem de ser possível controlar o que o utilizador faz com os dados, ex. impedir até que fotografias sejam tiradas dos dados em exibição. Medidas adicionais de segurança a ser tomadas no local podem incluir o não-fornecimento de rede/ligação à Internet, proibição de telefones ou computadores externos, videovigilância, etc.

18.7 Disponibilizar apenas um subconjunto do conjunto anonimizado – este subconjunto pode ser seleccionado ao acaso e/ou perturbado.

18.8 Medidas físicas – as medidas acima são sobretudo relacionadas com o controlo de acesso dos dados anonimizados em formato digital. As medidas físicas também se aplicam; exemplos destas incluem restringir o acesso físico a dispositivos ou dispositivos de armazenamento contendo ou com a possibilidade de aceder dados anonimizados, bem como restringir o acesso a impressões contendo os dados anonimizados.

19. Governança

19.1 A metodologia oferecida na secção 15 descreve os passos necessários para anonimizar de forma metódica um conjunto de dados. Porém, a anonimização responsável não acaba aí. Nota-se que uma abordagem com base no risco deve ser levada a cabo; por isso, as sugestões discutidas nesta secção são apenas para consideração e não são de adopção obrigatória.

19.2 Após a divulgação de um conjunto de dados anonimizado, é necessária uma governança adequada relativa ao conjunto anonimizado, até para divulgação privada. Isto pode incluir o seguinte:

- Manter o registo de conjuntos de dados anonimizados para divulgação pela organização. Os detalhes incluem os receptores e o método de acesso (ex. fornecer uma cópia do conjunto de dados anonimizado, ou acesso online, ou acesso físico, ou por pedido, etc.) Isto inclui as diferentes variantes/subconjuntos do mesmo, bem como conjuntos publicados por diferentes partes da organização; nos ambos os cenários, a combinação de diferentes conjuntos de dados pode levar a re-identificação.
- Gestão de chaves e mapeamento de tabelas – algumas técnicas de anonimização, incluindo pseudonimização, requerem o uso de chaves de encriptação, tabelas de mapeamento, etc. É crucial que estas sejam adequadamente geridas e mantidas em segurança, pelo que qualquer entidade que consiga aceder a estas pode imediatamente inverter a anonimização.
- Rever com regularidade o risco de re-identificação e os controlos que estão em efeito.
- Conduzir auditorias sobre os receptores dos dados, para garantir que estes cumprem com os requisitos contratuais.
- Notificação das partes relevantes se ocorrer algum tipo de fuga[11].
- Manter o registo dos requisitos de conformidade e as melhores práticas relativamente à anonimização de dados.

20. Agradecimentos

20.1 No desenvolvimento desta Guia, as melhores práticas de organizações da protecção de dados pessoais e outras autoridades de outros países foram consideradas. Ver Anexo B para as Guias que foram referenciados.

20.2 Gostaríamos de expressar o nosso agradecimento às seguintes organizações pela sua valiosa contribuição no desenvolvimento desta Guia:

- *Agency of Integrated Care (AIC)*
- *Changi General Hospital (CGH)*
- *Cyber Security Agency of Singapore (CSA)*
- *Government Technology Agency (GovTech)*

[11] *Vide* a Guia para a gestão de violação de dado, elaborada pela PDPC.

- *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*
- *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)*

Fim do Documento

Anexo A: Resumo das técnicas de anonimização

Nome da técnica	**Quando usar**	**Tipo de atributo**
Supressão de atributos	Atributo não requerido no conjunto de dados anonimizado	Todos
Supressão de registo	Precisa de registos nos extremos	N.D. (aplica-se por todo o registo, por isso todos os atributos são afectados)
Encobrimento de caracteres	Ocultar alguns caracteres num atributo fornece anonimização suficiente	Identificadores directos
Pseudonimização	Os registos ainda precisam de ser distinguidos entre si nos dados anonimizados, mas nenhuma parte do valor do atributo original pode ser retida	Identificadores directos
Generalização	Os atributos podem ser modificados para ser menos precisos, mas ainda assim úteis	Todos
Troca	Não é necessária uma análise de relações entre os atributos a nível do registo	Todos
Perturbação de dados	É aceitável uma ligeira modificação dos atributos	Identificadores indirectos
Dados sintéticos	Grandes quantidades de dados inventados semelhantes em natureza aos dados originais são necessárias, ex. para teste do sistema	Todos
Agregação de dados	Registos individuais não são necessários e os dados em agregado são suficientes	Identificadores indirectos

Anexo B: Referências principais

- "Advisory Guidelines on Key Concepts In the PDPA" (Chapter 5 – Personal Data). https://www.pdpc.gov.sg/AG. Personal Data Protection Commission (Singapore), revistas em 27 de Julho de 2017.
- "Advisory Guidelines on the PDPA for Selected Topics" (Chapter 3 – Anonymisation). https://www.pdpc.gov.sg/AG. Personal Data Protection Commission (Singapore), revistas em 28 de Março de 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, em Junho de 2016.
- "Guide to Managing Data Breaches". https://www.pdpc.gov.sg/OG. Personal Data Protection Commission (Singapore), em 8 de Maio de 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), em 10 de Abril de 2014.
- "Personal Data Protection Act 2012". Government Gazette. https://sso.agc.gov.sg/Act/PDPA2012. Republic of Singapore, 7 de Dezembro de 2017.
- 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), em Outubro de 2015.