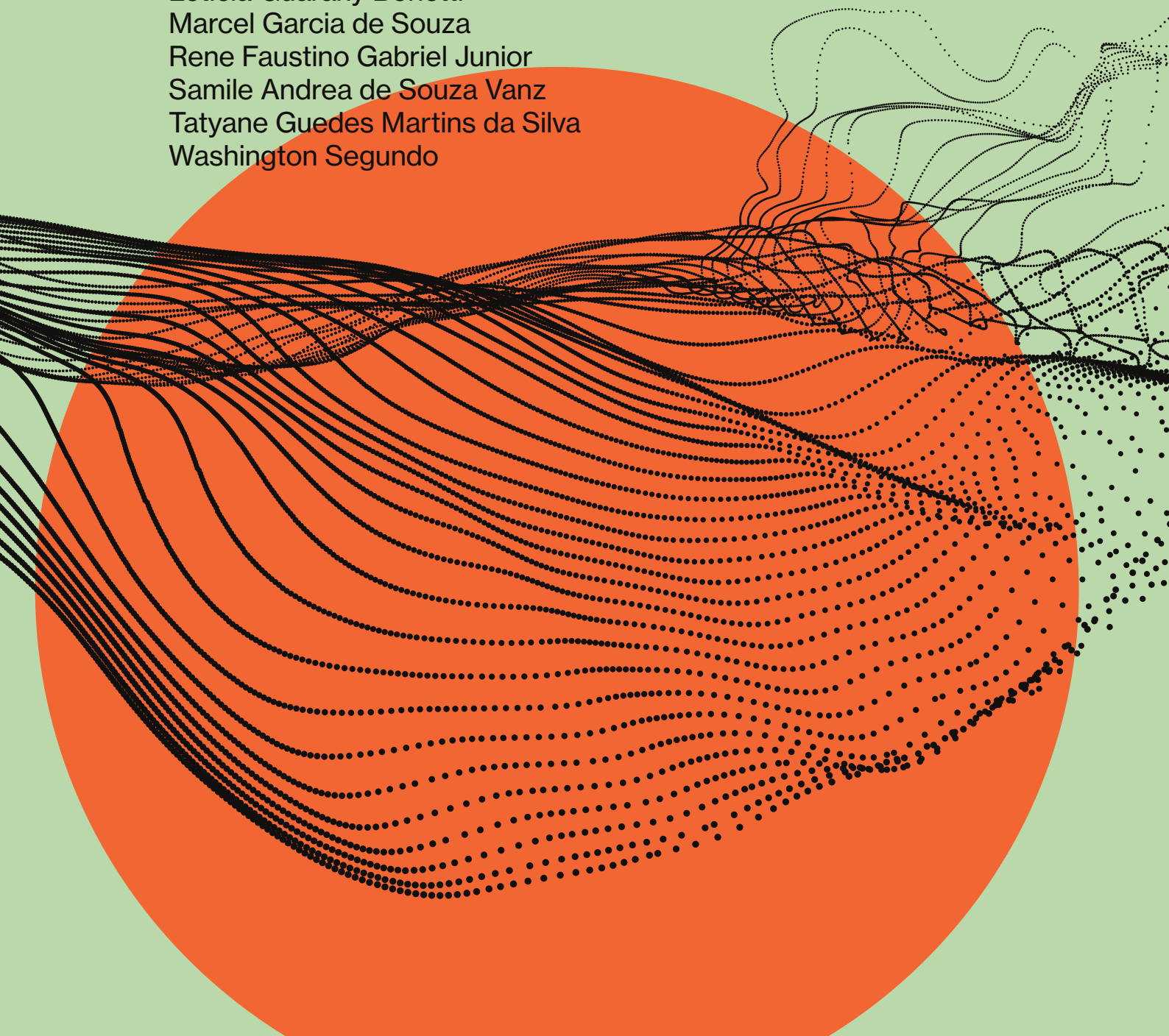
Editora
Ibict

Brasília 2026

GUIA DE BOAS PRÁTICAS
Anonimização de dados de pesquisa

Caterina Groposo Pavão
Letícia Guarany Bonetti
Marcel Garcia de Souza
Rene Faustino Gabriel Junior
Samile Andrea de Souza Vanz
Tatyane Guedes Martins da Silva
Washington Segundo



Ministério da Ciência, Tecnologia e Inovação
Instituto Brasileiro de Informação em Ciência e Tecnologia

GUIA DE BOAS PRÁTICAS PARA ANONIMIZAÇÃO DE DADOS DE PESQUISA

Caterina Groposo Pavão
Letícia Guarany Bonetti
Marcel Garcia de Souza
Rene Faustino Gabriel Junior
Samile Andrea de Souza Vanz
Tatyane Guedes Martins da Silva
Washington Luis Ribeiro de Carvalho Segundo

Editora Ibict

Brasília
2026

PRESIDÊNCIA DA REPÚBLICA

Luiz Inácio Lula da Silva
PRESIDENTE DA REPÚBLICA

Geraldo José Rodrigues Alckmin Filho
VICE-PRESIDENTE DA REPÚBLICA

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO

Luciana Santos
Ministra da Ciência, Tecnologia e Inovação

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA

Tiago Emmanuel Nunes Braga
Diretoria
Carlos André Amaral de Freitas
Coordenação de Administração - COADM
Ricardo Medeiros Pimenta
Coordenação de Ensino e Pesquisa em Informação para a Ciência e Tecnologia - COEPI
Henrique Denes Hilgenberg Fernandes
Coordenação de Planejamento, Acompanhamento e Avaliação - COPAV
Cecília Leite Oliveira
Coordenação-Geral de Informação Tecnológica e Informação para a Sociedade - CGIT
Washington Luís Ribeiro de Carvalho Segundo
Coordenação-Geral de Informação Científica e Técnica - CGIC
Hugo Valadares Siqueira
Coordenador-Geral de Tecnologias de Informação e Informática - CGTI
Milton Shintaku
Coordenador de tecnologias para informação - COTEC

Ministério da Ciência, Tecnologia e Inovação
Instituto Brasileiro de Informação em Ciência e Tecnologia

GUIA DE BOAS PRÁTICAS PARA ANONIMIZAÇÃO DE DADOS DE PESQUISA

Caterina Groposo Pavão
Letícia Guarany Bonetti
Marcel Garcia de Souza
Rene Faustino Gabriel Junior
Samile Andrea de Souza Vanz
Tatyane Guedes Martins da Silva
Washington Luis Ribeiro de Carvalho Segundo

Editora Ibict

Brasília - DF
2026

EDITORA Ibict

Conselho Editorial
Gustavo Silva Saldanha
Milton Shintaku
Luana Farias Sales
Franciele Garcês
Leyde Klébia Rodrigues da Silva
Stella Moreira Dourado
Daniel Strauch
Walisson Oliveira

Comitê Editorial
Tiago Braga
Milton Shintaku
Henrique Denes
Cecília Leite Oliveira
Ricardo Pimenta
Leda Cardoso Sampson Pinto
Carlos André Amaral de Freitas
Marcel Souza
Alexandre Oliveira
Washington Segundo
Emanuelle Torino
Alexandre Faria de Oliveira

Comitê Científico
Ania Rosa Hérnandez Quintana
Fernanda do Valle
María Arminda Damus
Martha Sabelli
Natalia Duque Cardona
Vinicios Meneses

EQUIPE TÉCNICA

Normalização
Caterina Groposo Pavão
Samile Andrea de Souza Vanz

Diagramação e projeto gráfico
Rafael Fernandez Gomes

Revisão
Caterina Groposo Pavão
Samile Andrea de Souza Vanz

Coordenador do Projeto
Marcel Garcia de Souza

Autorias
Caterina Groposo Pavão
Letícia Guarany Bonetti
Marcel Garcia de Souza
Rene Faustino Gabriel Junior
Samile Andrea de Souza Vanz
Tatyane Guedes Martins da Silva
Washington Luís Ribeiro de Carvalho Segundo

G943 Guia de boas práticas para anonimização de dados de pesquisa [recurso eletrônico] / Caterina Groposo Pavão...[et al.]. -- Brasília: Editora Ibict, 2026.

1 recurso online [55 p.] : il.

Modo de acesso: WWW
Publicação digital (e-book) no formato PDF. [3,36 MB]
ISBN: 978-85-7013-224-6
DOI: 10.22477/ 9788570132246

1. Anonimização de dados. 2. Dados de pesquisa. 3. Ciência Aberta. 4. Proteção de dados pessoais. I. Bonetti, Letícia Guarany. II. Souza, Marcel Garcia de. III. Gabriel Junior, Rene Faustino. IV. Vanz, Samile Andrea de Souza. V. Carvalho Segundo, Washington Luis Ribeiro de. VI. Título.

CDU 001.89:004.62

Ficha catalográfica elaborada por Bernardo Dionízio Vechi - CRB 1/2775

Como referenciar este livro:

PAVÃO, Caterina Groposo et al. **Guia de boas práticas para anonimização de dados de pesquisa.** Brasília, DF: Editora Ibict, 2026.

Endereço: Ibict - Instituto Brasileiro de Informação em Ciência e Tecnologia Setor de Autarquias Sul (SAUS), Quadra 05, Lote 06, Bloco H – 5o. andar CEP: 70.070-912 - Brasília.

Após escrever “HeLa” de Henrietta e Lacks, em letras pretas e grandes na superfície de cada tubo, Mary levou-os até a sala do incubador...

[...]

As células de Henrietta eram preciosas porque permitiam aos cientistas realizar experiências que seriam impossíveis com um ser humano vivo. Eles retalhavam as células HeLa e as expunham a inúmeras toxinas, radiação e infecções. Bombardeavam-nas com remédios, esperando encontrar um que matasse as células malignas sem destruir as normais. Estudavam a imunossupressão e o crescimento do câncer injetando as células HeLa em ratos imunocomprometidos, que desenvolviam tumores malignos semelhantes ao de Henrietta. Se as células morressem no processo, não importava – os cientistas podiam recorrer ao seu estoque em eterno crescimento de células HeLa e recomeçar.

[...]

Tudo é sempre sobre as células, mas nem se preocupam com seu nome ou se HeLa foi realmente uma pessoa.

[...]

Apesar de todos os outros processos e da cobertura feita pela imprensa, a família Lacks nunca tentou processar ninguém por causa das células HeLa. [...] como a esta altura não é mais possível manter as células HeLa anônimas, as pesquisas com elas deveriam estar cobertas pela Common Rule. E, como parte do DNA presente nas células de Henrietta também está presente em seus filhos, é possível argumentar que, ao fazerem pesquisas com células HeLa, os cientistas também estão fazendo pesquisas com os filhos de Henrietta. Uma vez que a Common Rule afirma que as cobaias devem ter o direito de abandonar a pesquisa a qualquer momento, esses especialistas me disseram que, em tese, a família Lacks poderia remover as células HeLa de todas as pesquisas no mundo inteiro. [...] Todos os pesquisadores a quem apresento essa ideia se arrepiam só de pensar nessa possibilidade.

Rebecca Skloot em A vida imortal de Henrietta Lacks.

SUMÁRIO

PREFÁCIO

Os dados de pesquisa são elementos centrais para o fazer científico. Nas orientações sobre uma ciência mais aberta, transparente, reprodutível e cidadã, os dados são reconhecidos como ativos estratégicos para inovação, construção de políticas públicas baseadas em evidências, assim como para soberania e garantia da democracia. Neste ponto, entende-se que há um conjunto de questões e procedimentos que precisam ser observados nos processos de abertura de dados de pesquisas, incluindo aqueles que envolvem aspectos éticos e jurídicos, visando a conformidade com as normativas vigentes no país.

Nesse contexto, o **Guia de Boas Práticas para Anonimização de Dados de Pesquisas**, elaborado pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), representa um avanço concreto na materialização dessas agendas. O documento enfrenta um dos desafios centrais para a Ciência Aberta: compatibilizar o uso e o compartilhamento de dados para pesquisas com a proteção dos direitos fundamentais, como o de liberdade, privacidade e o livre desenvolvimento da personalidade da pessoa natural.

O arcabouço normativo brasileiro reconhece a legitimidade do uso de dados para fins científicos, considerando evidente interesse público ou geral, como previsto na Lei de Acesso à Informação (LAI) e na Política Nacional de Dados Abertos. A Lei Geral de Proteção de Dados Pessoais (LGPD), assim como a mais recente Lei que regula a pesquisas com seres humanos e institui o Sistema Nacional de Ética em Pesquisa com Seres Humanos, também mencionam o uso de dados pessoais para pesquisas científicas, mas ressaltam os aspectos de consentimento e segurança. Ainda assim, persiste uma lacuna entre esses marcos e a operacionalização prática do compartilhamento e da abertura de dados, especialmente no que se refere à definição de técnicas, critérios e níveis de acesso.

É nesse ponto que este Guia se afirma como uma contribuição estratégica. Ao oferecer orientações claras sobre anonimização de dados para pesquisa, o documento apoia pesquisadores e instituições na adoção de práticas responsáveis, alinhadas aos princípios da Ciência Aberta e ao interesse público. Reafirma-se, assim, que a abertura de dados não é um fim em si mesma, mas um processo cuidadoso, ético e situado.

O lançamento deste Guia reforça o papel do Ibict na articulação entre políticas, infraestruturas e práticas no campo da informação científica e contribui para o fortalecimento de uma ciência mais aberta, sem perder o compromisso ético, a conformidade jurídica e com foco no acesso e na democratização do conhecimento.

Vanessa Arruda Jorge - Fiocruz

1 INTRODUÇÃO

A pesquisa científica constitui uma atividade eminentemente social, repleta de processos cíclicos realizados por equipes de pesquisadores e encadeados em uma série sucessiva, que inclui a coleta de dados, a aplicação do método científico, a análise e discussão dos dados, a apresentação de resultados preliminares em conferências e eventos, a publicação dos resultados em revistas científicas, até que estes resultados finalmente sejam acessados por outros pesquisadores, citados e posteriormente fomentem ideias para novos projetos de pesquisa (Meadows, 1999; Hurd, 2000; Targino, 2000).

O contexto atual de fazer ciência pressupõe a colaboração científica, independentemente da localização física das equipes de pesquisadores, bem como o acesso a instrumentos, dados, informações e recursos computacionais, e a bibliotecas digitais. Há uma tendência à publicação dos artigos que apresentam os resultados de pesquisa em *open journals* (periódicos de acesso aberto), assim como ao arquivamento destes artigos em repositórios abertos e à demanda em torno do compartilhamento dos dados de pesquisa.

Este formato mais colaborativo de fazer ciência insere-se num movimento denominado Ciência Aberta, que busca tornar o conhecimento científico acessível a todos, promovendo a transparência e a colaboração em todas as etapas da pesquisa. Isso inclui o acesso aberto a artigos científicos, dados e software, bem como o uso de práticas transparentes e colaborativas em todas as etapas do processo de pesquisa (Albagli; Maciel; Adbo, 2015).

A disponibilização de dados de pesquisa é um dos pilares da Ciência Aberta e vem crescendo cada vez mais nos últimos anos. Com isso, surge a necessidade de compartilhá-los de forma segura e responsável nos repositórios de dados (Curty, 2019; Gabriel Junior, et al. 2019). Dependendo do tipo de informação coletada na pesquisa, é essencial que os dados sejam anonimizados antes de serem compartilhados, protegendo a privacidade e cumprindo os requisitos legais.

Este Guia foi desenvolvido para oferecer orientações sobre como realizar a anonimização de dados de pesquisa, com a finalidade de capacitar pesquisadores a compreender o significado e a importância da anonimização, apresentar técnicas de proteção de informações pessoais e contribuir para a Ciência Aberta, em conformidade com a Lei Geral de Proteção de Dados (Lei 13.709/2018). A Lei Federal n. 13.709, a Lei Geral de Proteção de Dados Pessoais (LGPD), foi publicada em 14 de agosto de 2018 e entrou em vigor em 18 de setembro de 2020, data em que sua observância passou a ser obrigatória em todo o território nacional (Brasil, 2018). A Lei Geral de Proteção de Dados (LGPD) se relaciona ao tema dos dados abertos de pesquisa em razão da identificação de dados pessoais.

Conforme a Autoridade Nacional de Proteção de Dados (ANPD), em seu Guia Orientativo Tratamento de dados pessoais para fins acadêmicos e para a realização de estudos e pesquisas (Vargas, 2023, p. 5), a LGPD procurou estabelecer uma relação de equilíbrio entre, de um lado, a proteção de dados pessoais e as garantias da privacidade e da autodeterminação informativa e, de outro, a liberdade acadêmica e o livre fluxo de informações necessário para a realização de estudos e pesquisas nas mais diversas áreas do saber. De acordo com a ANPD, o artigo 13 da LGPD ratifica a autorização para disponibilização de acesso a dados pessoais para fins de realização de estudos e pesquisas, estipulando, em acréscimo, medidas específicas de prevenção e segurança a serem observadas no campo dos estudos de saúde pública. Os dados pessoais devem ser tratados exclusivamente no órgão de pesquisa e para o atendimento ao protocolo de pesquisa ou conforme o consentimento do participante. Além disso, devem ser armazenados em ambiente controlado e seguro, com a sua anonimização ou pseudonimização sempre que possível. Por fim, devem ser observados os padrões éticos, não se admitindo a revelação de informações pessoais por ocasião da publicação do resultado do estudo.

Tendo ciência das orientações da ANPD quanto ao atendimento às disposições da LGPD por pesquisadores envolvidos com dados de pesquisa acadêmica e científica, o Guia de Boas Práticas para Anonimização de Dados de Pesquisa, criado pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), é voltado a orientar pesquisadores que desejam compartilhar seus dados em repositórios de dados de pesquisa. Os repositórios "[...] buscam organizar, estruturar, permitir acesso, disseminar e preservar todos os dados gerados por meio de pesquisas realizadas em sua maioria por Instituições de Ensino e Pesquisa" (Sanchez; Vidotti; Vechiato, 2017, p. 3).

Um exemplo é o Deposita Dados, um repositório de dados em acesso aberto que tem como objetivos arquivar, publicar, disseminar e preservar conjuntos de dados de pesquisa de pesquisadores brasileiros vinculados a instituições científicas que ainda não possuem seus repositórios de dados de pesquisa e/ou de pesquisadores brasileiros que executaram seus conjuntos de dados por meio de colaboração científica em instituições estrangeiras de ensino e pesquisa. Sendo assim, qualquer pesquisador brasileiro pode depositar, gratuitamente, dados de pesquisa no Deposita Dados. Outro exemplo de repositório de dados de pesquisa é o *Aleia*, que tem como objetivo arquivar, publicar, disseminar e preservar conjuntos de dados de pesquisa da comunidade científica do Ibict, sendo, portanto, um repositório institucional. Ambos os repositórios citados são geridos pela Coordenação-Geral de Informação Científica e Técnica (CGIC), que faz parte do Ibict.

O objetivo do Guia de Boas Práticas para Anonimização de Dados de Pesquisa é reunir algumas informações, técnicas e exemplos de anonimização. Destaca-se que o Guia não tem a pretensão de ser exaustivo e está organizado em seis seções. A seção 2 reúne conceitos importantes relativos à anonimização de dados. A seção 3 apresenta o passo a passo para a anonimização de dados. A seção 4 reúne métodos para anonimização, incluindo dados alfanuméricos e arquivos em imagem e vídeo. A seção 5 apresenta alguns exemplos de anonimização de dados alfanuméricos, seguida pelos exemplos de desidentificação de dados em imagem e vídeo, apresentados na seção 6. Ao final do Guia são apresentadas as referências e as informações sobre os autores.

2 CONCEITOS RELATIVOS À ANONIMIZAÇÃO DE DADOS

A Lei Geral de Proteção de Dados Pessoais busca garantir a proteção das informações pessoais dos indivíduos, estabelecendo princípios, direitos e responsabilidades que promovem a segurança e a privacidade no uso de dados no Brasil. A LGPD define dados pessoais como informações relacionadas a uma pessoa identificada ou identificável. Isso inclui não apenas as informações mais evidentes, como nomes e números de identificação, mas também informações que, quando combinadas ou analisadas, podem levar à identificação de uma pessoa.

Dado pessoal é a "informação relacionada a pessoa natural identificada ou identificável", enquanto dado pessoal sensível é o dado pessoal relacionado a aspectos da personalidade do titular, "sobre origem racial ou étnica, convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político, dado referente à saúde ou à vida sexual, dado genético ou biométrico, quando vinculado a uma pessoa natural" (Brasil, 2018, art. 5º, I, II, III).

Dados pessoais são aqueles que se relacionam com um indivíduo vivo que pode ser identificado: a) a partir desses dados ou b) a partir de dados e outras informações que tenham sido registradas pelo criador dos dados. Portanto, informações ou uma combinação de informações que não se relacionam nem identificam um indivíduo não são dados pessoais (Newton; Sweeney; Malin, 2005).

Uma pessoa identificável é aquela que pode ser identificada, direta ou indiretamente, em particular por referência a um número de identificação ou a um ou mais fatores específicos da sua identidade física, fisiológica, mental, econômica, cultural ou social (TransCelerate BioPharma, 2016).

Informações pessoais identificáveis referem-se a informações que podem ser usadas para distinguir ou rastrear a identidade de um indivíduo, como seu nome, número de previdência social, registros biométricos etc., isoladamente ou quando combinadas com outras informações pessoais ou de identificação que estejam vinculadas ou possam ser vinculadas a um indivíduo específico, como data e local de nascimento, nome de solteira da mãe etc. (União Europeia, 2016)..

Dado anonimizado é o dado "relativo a titular que não possa ser identificado, considerando a utilização de meios técnicos razoáveis e disponíveis na ocasião de seu tratamento" (Brasil, 2018, art. 5º, I, II, III).

A anonimização é definida como uma etapa subsequente à desidentificação, que envolve a destruição irreversível de todos os vínculos entre os conjuntos de dados desidentificados e os conjuntos de dados originais (União Europeia, 2016). A técnica tem como objetivo proteger a privacidade dos indivíduos ao remover ou alterar informações pessoais de forma que não seja mais possível identificar diretamente ou indiretamente um indivíduo.

Anonimizar dados consiste em remover ou modificar as variáveis de identificação, ou seja, informações que descrevem uma característica observável de uma pessoa, registradas (números de identificação etc.) ou, de modo geral, que podem ser conhecidas por outras pessoas. As variáveis de identificação são classificadas em diretas e indiretas (International Household Survey Network, [2025]).

Identificadores diretos são variáveis como nomes, endereço ou número de documento de identidade. Eles permitem a identificação direta de um respondente, mas não são necessários para fins estatísticos ou de pesquisa e, portanto, devem ser removidos do conjunto de dados publicado.

Identificadores indiretos são características que podem ser compartilhadas por vários respondentes e cuja combinação pode levar à reidentificação de um deles. Por exemplo, a combinação de variáveis como estado de residência, idade, sexo e profissão identificaria se apenas um indivíduo de determinado sexo, idade e profissão vivesse naquele estado específico. Essas variáveis são necessárias para fins estatísticos e não devem ser removidas dos arquivos de dados publicados. Também são conhecidos como quase-identificadores (Personal Data Protection Commission Singapore, 2018).

A anonimização dos dados envolve determinar quais variáveis são identificadores potenciais (com base em julgamento pessoal) e ajustar a especificidade dessas variáveis para reduzir o risco de reidentificação a um nível aceitável. O desafio é maximizar a segurança e, ao mesmo tempo, minimizar a perda de informações resultante. Vale ressaltar que as técnicas de anonimização podem reduzir o risco de identificação, mas deve haver uma avaliação separada para determinar se o risco reduzido é aceitável em circunstâncias específicas e se o processo técnico constitui “anonimização eficaz”, caso seja necessário. Também deve ser observado que, mesmo quando for legalmente permitido o uso de informações que identifiquem um indivíduo, seu uso deve ser minimizado, por exemplo, enviando apenas um subconjunto relevante das informações ou utilizando uma ou mais técnicas de anonimização (A Guide to Confidentiality in Health and Social Care, 2013).

Diferente da anonimização, a pseudonimização é uma técnica que substitui informações que identificam diretamente as pessoas ou desacopla essas informações do conjunto de dados resultante. Por exemplo, pode envolver a substituição de nomes ou de outros identificadores (facilmente atribuídos a pessoas) por um número de referência. Isso é semelhante a como o termo “de-identificado” é usado em outros contextos. Por exemplo, remover ou mascarar identificadores diretos em um conjunto de dados.

Os “dados pseudonimizados” são dados sobre pessoas que não podem ser identificadas a partir dessas informações por si sós, mas podem ser identificadas a partir de informações adicionais mantidas separadamente. A pseudonimização reduz as ligações entre as pessoas e os dados que as relacionam, mas não as remove completamente. Já a anonimização evita que haja uma ligação entre a informação e a pessoa em questão.

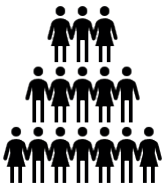
A LGPD enfatiza a importância da anonimização e pseudonimização dos dados pessoais em pesquisas, tornando a identificação de indivíduos impossível ou altamente improvável. Além disso, destaca a necessidade de consentimento informado para a coleta, o processamento e o compartilhamento de dados pessoais em pesquisas, garantindo que os participantes compreendam claramente como seus dados serão utilizados.

Após esse entendimento, é essencial compreender quais dados precisam ser anonimizados para garantir o cumprimento da LGPD.

3 PASSO A PASSO PARA ANONIMIZAÇÃO DE DADOS

Apresentam-se a seguir os procedimentos para identificar a presença de dados pessoais dentre os dados da pesquisa, e a consequente necessidade de anonimização.

Verifique se os dados envolvem informações que podem identificar indivíduos (ex.: nome, CPF, e-mail);



Classifique os dados como pessoais ou sensíveis. Dados sensíveis (ex.: saúde, religião, orientação sexual) exigem cuidados extras;

Escolha o método apropriado: determine o método de anonimização (supressão, generalização etc.) adequado ao tipo de dado e à finalidade;

Certifique-se de que o método escolhido não permite a reidentificação do indivíduo.

De acordo com a LGPD, dados pessoais são aqueles que identificam ou podem identificar um indivíduo, como por exemplo:

Nome completo

CPF ou RG

Número de conta bancária

Número de telefone

E-mail

Endereço residencial

Endereço IP

Dados pessoais sensíveis são aqueles que contêm informações pessoais que podem levar à discriminação, prejuízo ou danos à privacidade do indivíduo, por exemplo:

Origem racial ou étnica

Orientação sexual

Dados de saúde (exames médicos, diagnósticos, prontuários)

Crenças religiosas

Opiniões políticas

Portanto, antes que esses dados sejam compartilhados, é necessário anonimizá-los. O processo de anonimização não compromete a privacidade dos indivíduos envolvidos na pesquisa e coloca os dados em conformidade com a legislação brasileira.

De acordo com a *Personal Data Protection Commission Singapore* (2018), ao realizar a anonimização é necessário:

1. Determinar a forma como o conjunto de dados anonimizado será divulgado. **Público** refere-se à disponibilização a praticamente qualquer pessoa. **Não público** refere-se à divulgação controlada a destinatários conhecidos limitados (e frequentemente, a um número fixo de destinatários);

2. Determinar o limite aceitável de risco de reidentificação, bem como o risco esperado, utilidade e limite de risco pretendido ou exigido;

3. Classificar os atributos dos dados no conjunto de dados como identificadores diretos, identificadores indiretos, ou não identificadores, que afetam como os atributos serão posteriormente processados;

4. Remover atributos de dados não utilizados no processo de anonimização. Geralmente, a maioria dos atributos, sejam identificadores diretos ou indiretos, requer processamento ou pelo menos consideração, a fim de diminuir sua capacidade de identificação. Portanto, deve ser suprimido qualquer atributo que claramente não seja necessário no conjunto de dados anonimizado;

5. Anonimizar identificadores diretos e indiretos. Diferentes técnicas são aplicáveis a diferentes tipos de identificadores. Algumas técnicas podem (e frequentemente devem) ser usadas em combinação;

6. Determinar o risco real e comparar com o limite;

7. Realizar mais anonimização, se necessário. Se o risco real for maior que o limite, será necessária uma anonimização "mais forte" e as etapas 5 a 7 deverão ser executadas novamente com os ajustes necessários, até que o risco real seja menor que o limite;

8. Avaliar a solução, inclui examinar o conjunto de dados anonimizados para ponderar se a utilidade atende à meta. Se a utilidade for insuficiente, o processo de anonimização pode precisar ser reformulado ou pode-se considerar se a anonimização é viável para esse conjunto de dados;

9. Determinar os controles necessários, tanto técnicos quanto não técnicos (por exemplo, medidas legais e organizacionais);

10. Documentar o processo de anonimização. Os detalhes do processo de anonimização, os parâmetros utilizados e os controles devem ser registrados de forma clara para referência futura. Essa documentação facilita a revisão, a manutenção, os ajustes finos e as auditorias. Observe que essa documentação deve ser mantida em segurança, pois a divulgação dos parâmetros pode facilitar a reidentificação.

A próxima seção apresenta alguns métodos para anonimização de dados de pesquisa, seguidos de exemplos de aplicação prática.

4 MÉTODOS DE ANONIMIZAÇÃO DE DADOS DE PESQUISA

Os dados de pesquisa abrangem uma ampla gama de formatos, incluindo imagens, áudio, dados não estruturados, entre outros. Existem diferentes métodos que visam a anonimização de dados pessoais, porém o método que será utilizado depende do tipo de dado e do contexto da pesquisa.

4.1 DADOS ALFANUMÉRICOS

Apresentam-se a seguir os principais métodos básicos de anonimização para dados alfanuméricos que podem ser aplicados de forma simples pelas equipes responsáveis pela pesquisa e coleta de dados:

Quadro 1 - Principais técnicas de anonimização aplicáveis à dados de pesquisa alfanuméricos

Técnica	Definição	Utilização	Observação
Supressão de atributos	Remoção de informações como, por exemplo, uma coluna inteira de dados	Nomes, endereços, CPFs	A sua utilização pode levar à perda de informações relevantes para a pesquisa
Supressão de registro	Remoção de um registro (caso) inteiro, usado para casos únicos (outliers) facilmente identificáveis	Indivíduos portadores de respostas/características que diferem do restante do grupo, como o mais velho, o mais obeso etc.	A utilização altera cálculos de média e mediana
Pseudonimização	Desidentificação de dados para que uma referência codificada ou pseudônimo seja anexado a um registro para permitir que os dados sejam associados a um indivíduo específico sem que o indivíduo seja identificado Técnica que substitui um identificador (nome ou CPF, por exemplo) por um valor não relacionado, mas único (como um numeral, por exemplo)	Permite a vinculação usando os mesmos valores de pseudônimo para representar o mesmo indivíduo em diferentes conjuntos de dados	É uma técnica de risco relativamente alto semelhante ao mascaramento É uma técnica reversível quando os valores originais são mantidos em segurança

Técnica	Definição	Utilização	Observação
Generalização	Redução deliberada da precisão dos dados. Por exemplo, converter a idade de uma pessoa em uma faixa etária, ou uma localização precisa em uma localização menos precisa, de nível regional ou continental. Essa técnica também é conhecida como recodificação	Para valores que podem ser generalizados e ainda assim serem úteis para o propósito pretendido. Útil para data, idade, localizações geográficas, faixas de renda	Intervalos de dados muito pequenos permitem a fácil reidentificação
Perturbação de Dados ou Adição de ruído	Modificação dos valores do conjunto de dados original para serem ligeiramente diferentes. Por exemplo, é feito um arredondamento	Para dados quase-identificadores (números e datas) que podem ser potencialmente identificadores quando combinados com outras fontes de dados. Pequenas alterações no valor são aceitáveis	
Agregação de dados	Conversão de um conjunto de dados de uma lista de registros em valores resumidos	Quando registros individuais não são necessários e dados agregados são suficientes para o propósito	
Mascaramento	Ocultação de partes importantes /únicas dos dados com caracteres aleatórios ou outros caracteres como por exemplo "*" ou "x"	Quando o valor dos dados é uma sequência de caracteres e ocultar parte dele é suficiente para fornecer o grau de anonimato necessário	
Embaçamento	Criação de uma aproximação aos valores de dados para tornar seu significado obsoleto e/ou impossibilitar a identificação de indivíduos		
Permutação (*swapping*)	Reorganização dos dados no conjunto de forma que os valores dos atributos individuais ainda sejam representados no conjunto de dados, mas geralmente não correspondem aos registros originais. Essa técnica também é conhecida como embaralhamento e permutação		Esta técnica não deve ser usada quando a precisão dos dados é importante. Somente é indicada quando não há necessidade de análise das características dos sujeitos da pesquisa

Fonte: elaboração própria (2026).

Destaca-se que os dados anonimizados não devem ser reconhecíveis nem mesmo pelo próprio pesquisador titular dos dados, tampouco pela equipe responsável pela pesquisa.

Os métodos de anonimização mais sofisticados exigem técnicas computacionais e conhecimento especializado, por isso não são abordados neste Guia. Entre estes métodos citamos os Dados agregados, a Criptografia e a Tokenização. Determinados tipos de dados, dependendo da área do conhecimento onde a pesquisa se desenvolve, podem exigir o emprego destas técnicas. Dada a necessidade de amplo conhecimento para aplicação destas técnicas, sugere-se a contratação de um profissional especialista dedicado ao processo e controle da anonimização dos *datasets*.

4.2 DADOS EM IMAGEM E VÍDEO

No contexto do tratamento de dados de pesquisa de imagens e vídeos, pode-se falar em anonimização ou desidentificação. A desidentificação é utilizada quando se refere ao processo de ocultar, suprimir ou modificar elementos visuais ou sonoros que permitam reconhecer uma pessoa, como o rosto, características corporais específicas, voz ou outros marcadores individuais. Diferentemente, a anonimização é definida pela LGPD como um processo irreversível que impossibilita qualquer forma de reidentificação.

A desidentificação não elimina completamente o risco de identificação, mas reduz significativamente sua probabilidade. Dessa forma, a desidentificação preserva a utilidade do material para fins de pesquisa, ensino ou divulgação, ao mesmo tempo em que assegura a proteção dos direitos dos titulares e reduz os riscos associados ao tratamento de dados pessoais (Brasil, 2018).

A desidentificação de imagens é um campo essencial para a preservação da privacidade, especialmente diante do aumento do uso de câmeras de vigilância e de algoritmos de reconhecimento facial. Embora algumas técnicas simples possam ser facilmente contornadas por *softwares* avançados, existem métodos que oferecem garantias mais robustas contra identificação, equilibrando privacidade e utilidade das imagens desidentificadas. O objetivo da desidentificação é modificar as imagens para impedir o reconhecimento automático por *softwares* de reconhecimento facial, preservando, ao mesmo tempo, alguns detalhes visuais úteis para análise (Newton; Sweeney; Malin, 2005).

Em muitos países, a coleta e o processamento de imagens faciais são regulados por leis de proteção de dados, como o *General Data Protection Regulation* (GDPR) da União Europeia. A desidentificação pode ser usada para cumprir exigências legais ao evitar o armazenamento de dados biométricos identificáveis sem o consentimento adequado. No entanto, se um método de desidentificação puder ser revertido, as imagens podem ainda ser consideradas dados pessoais, estando sujeitas a regulamentações mais rigorosas (Newton; Sweeney; Malin, 2005).

A eficácia dos métodos de desidentificação pode impactar as pesquisas. Se uma técnica não for robusta o suficiente, pode gerar problemas relacionados à reidentificação indevida. A desidentificação facial é uma ferramenta útil para proteger a privacidade em diversas aplicações, mas sua implementação deve levar em conta as exigências legais e o risco de reidentificação. Os pesquisadores que trabalham com identificação de imagens, especialmente no contexto de reconhecimento facial e

desidentificação, têm várias responsabilidades éticas, legais e técnicas. As principais responsabilidades, de acordo com Newton, Sweeney e Malin (2005), dizem respeito a:

- **Proteção da Privacidade**

  - » Anonimização eficiente: os pesquisadores devem garantir que as técnicas de desidentificação realmente protejam a identidade dos indivíduos, evitando a reidentificação indesejada.

  - » Conformidade com Leis de Proteção de Dados: devem seguir regulamentações como o GDPR (Europa) e a LGPD (Brasil), garantindo que dados biométricos sejam manipulados de maneira ética e legal.

  - » Evitar riscos de abuso: qualquer tecnologia desenvolvida deve impedir usos indevidos, como vigilância massiva sem consentimento ou discriminação de grupos específicos.

- **Transparência e consentimento**

  - » Informação clara aos participantes: se imagens faciais forem coletadas de pessoas, é essencial garantir que os sujeitos saibam como seus dados serão usados e tenham a opção de consentir ou recusar.

  - » Divulgação de métodos e impactos: publicações científicas devem apresentar claramente as limitações das técnicas de desidentificação e os riscos de reidentificação.

- **Segurança dos dados**

  - » Armazenamento seguro: dados sensíveis, como imagens faciais, devem ser protegidos contra acessos não autorizados e ataques cibernéticos.

  - » Técnicas de criptografia: métodos de proteção, como pseudonimização e criptografia, devem ser considerados para reforçar a segurança dos dados armazenados.

- **Desenvolvimento responsável de tecnologia**

  - » Minimizar viés algorítmico: os pesquisadores devem testar suas técnicas com conjuntos de dados diversos para evitar desigualdades na identificação facial, especialmente para diferentes grupos raciais e de gênero.

  - » Equilíbrio entre privacidade e utilidade: as técnicas de desidentificação devem manter um nível adequado de utilidade para segurança pública e pesquisa médica, sem comprometer direitos individuais.

- **Responsabilidade social e ética**

  » Impacto na sociedade: a pesquisa deve considerar como novas tecnologias de identificação e desidentificação podem influenciar questões como direitos civis e liberdade individual.

  » Colaboração com reguladores: trabalhar em conjunto com legisladores para criar normas claras e diretrizes sobre o uso dessas tecnologias.

Os pesquisadores têm um papel fundamental na criação de tecnologias que equilibram privacidade e segurança. Garantir conformidade legal, transparência e ética na manipulação de imagens faciais é essencial para evitar riscos de reidentificação e proteger os direitos individuais. As principais técnicas utilizadas na desidentificação de imagens e vídeos foram resumidas no Quadro 2.

Quadro 2 - Principais técnicas de desidentificação para dados de pesquisa em imagem e vídeo

Técnica	Objetivo	Limitações
BlackOut (Máscara Completa)	Consiste em cobrir completamente o rosto com uma cor sólida.	Apesar de eficaz para impedir o reconhecimento, elimina todas as características faciais, tornando a imagem pouco útil para análises de características úteis, como expressões e movimentos.
Máscaras Parciais	Consiste em cobrir os olhos (*Bar Mask*) ou nariz e olhos (*T Mask*) para ocultar apenas partes da face.	*Softwares* avançados ainda conseguem identificar indivíduos com base nas características remanescentes.
Pixelização	Reduz a resolução da imagem, substituindo blocos de pixels por uma média de valores.	Embora seja uma técnica popular, experimentos mostram que a maioria dos sistemas de reconhecimento facial ainda conseguem identificar rostos pixelados com alta precisão. Pode ser revertida parcialmente usando algoritmos avançados de restauração de imagem. Isso significa que, se uma pessoa tiver acesso a técnicas aprimoradas, pode recuperar informações faciais suficientes para identificação.
Ruído aleatório	Introdução de distorções aleatórias nos pixels da imagem para dificultar a identificação.	O efeito depende da quantidade de ruído aplicado; pequenas alterações não impedem o reconhecimento, mas alterações significativas podem distorcer demais a imagem. Pode ser revertida parcialmente usando algoritmos avançados de restauração de imagem. Isso significa que, se uma pessoa tiver acesso a técnicas aprimoradas, pode recuperar informações faciais suficientes para identificação.
Limiarização (*Thresholding*)	Redução da quantidade de tons de cinza para apenas dois valores (preto e branco).	O impacto sobre o reconhecimento varia com o nível de limiar aplicado.

Técnica	Objetivo	Limitações
k-Same (Média de Vários Rostos)	Algoritmo que agrupa rostos semelhantes e substitui cada rosto por uma imagem média do grupo. Um rosto desidentificado pode representar vários indivíduos. » *k-Same-Pixel* utiliza a média dos pixels das imagens originais. » *k-Same-Eigen* aplica a média em um espaço reduzido de características principais (*eigenfaces*), causando um efeito de desfoque.	Pode tornar a identificação automática imprecisa. Tenta preservar algumas características visuais, mas ainda pode distorcer a identidade original. Pode ser vulnerável a modelos de aprendizado de máquina que identificam padrões sutis na imagem desidentificada. Isso pode permitir que algoritmos reconstruam rostos ou associem imagens a indivíduos com base em outros fatores, como estrutura facial ou estilo de cabelo.

Fonte: elaboração própria (2026).

Todas as técnicas apresentam limitação na proteção contra identificação contextual. Newton, Sweeney e Malin (2005) colocam que a desidentificação de rostos impede o reconhecimento facial direto, mas não protege contra identificação baseada em contexto. Por exemplo, roupas, postura corporal ou objetos próximos podem revelar a identidade de uma pessoa, tornando a proteção da privacidade incompleta em imagens e vídeos. Por outro lado, os mesmos autores acrescentam que, por exemplo, em estudos médicos, a desidentificação pode impedir que pesquisadores obtenham informações cruciais para suas análises. Portanto, é necessário equilibrar privacidade e usabilidade dependendo do contexto.

Embora as técnicas de desidentificação sejam eficazes na proteção contra reconhecimento facial, elas não garantem anonimato absoluto. A escolha do método depende do nível de privacidade desejado e do contexto de uso. É fundamental considerar essas limitações ao implementar soluções de proteção de identidade em imagens e vídeos.

5 EXEMPLOS DE ANONIMIZAÇÃO DE DADOS ALFA NUMÉRICOS

Nesta seção, serão apresentados exemplos práticos de técnicas de anonimização de dados alfa-numéricos para demonstrar como e quando podem ser aplicadas em diferentes contextos de uma pesquisa. Como base para a elaboração desta seção, foi utilizado o *Guide to Basic Data Anonymisation Techniques*, publicado em 2018 pelo *Personal Data Protection Commission Singapore*, que tem como objetivo fornecer uma introdução geral aos aspectos técnicos da anonimização.

Antes de iniciar a descrição e apresentar exemplos de cada uma das técnicas, é importante abordar um modelo de proteção da privacidade que se aplica a várias técnicas que serão apresentadas a seguir: o *k*-anonimato (k-Anonymity). É um modelo de proteção da privacidade que busca reduzir o risco de reidentificação ao garantir que cada registro em um conjunto de dados seja indistinguível de, no mínimo, outros (*k*-1) registros a partir de seus quasi-identificadores. Para isso, aplica técnicas como generalização, supressão e agregação para formar grupos de indivíduos com perfis equivalentes. Embora efetiva para diminuir os riscos baseados em vinculação de atributos, a *k*-Anonimização apresenta limites diante de cenários de homogeneidade ou de conhecimento prévio dos dados, motivando o uso de métodos complementares, como *l*-Diversidade e *t*-Proximidade (Sweeney, 2002). Pode ser considerada uma técnica de anonimização, mas é mais uma medida aplicada para garantir que o limiar de risco não foi ultrapassado, como parte da metodologia de anonimização. É usado como uma diretriz para verificação, depois que técnicas de anonimização (por exemplo, generalização) foram aplicadas (PDPC, 2018).

A *l*-Diversidade (l-Diversity) e a *t*-Proximidade (tCloseness) constituem extensões importantes do modelo de *k*-Anonimização, desenvolvidas para diminuir suas vulnerabilidades a ataques de reidentificação. A *l*-Diversidade propõe que, dentro de cada grupo de equivalência, haja pelo menos *l*-valores distintos e bem representados para o atributo sensível, reduzindo o risco de inferência quando os grupos são homogêneos. A *t*-Proximidade, por sua vez, avança essa proteção ao exigir que a distribuição dos valores sensíveis em cada grupo esteja a uma distância máxima *t* da distribuição global desses valores no conjunto de dados, impedindo que diferenças estatísticas revelem informações privadas. Juntas, essas técnicas buscam fortalecer a privacidade ao evitar tanto a homogeneidade quanto a fuga significativa de informação estatística nos dados anonimizados (Li; Li; Venkatasubramanian, 2007).

5.1 GENERALIZAÇÃO

Descrição: a técnica agrupa os dados com características em comum em um nível de granularidade maior. Os valores dos atributos são substituídos pelos valores do grupo. Por exemplo, converter a idade de uma pessoa em uma faixa etária ou uma localização precisa em uma localização menos precisa. Essa técnica também é conhecida como recodificação (Figura 1).

Quando usar: para valores que podem ser generalizados e ainda assim serem úteis para o propósito pretendido.

Como usar: crie categorias de dados e regras apropriadas para traduzir dados. Considere suprimir quaisquer registros que ainda se destaquem após a generalização.

Figura 1 - Exemplo da anonimização por generalização

Dados originais		
Nome	**Cidade de nascimento**	**Idade**
FM	Rio Branco	79
AFB	Macaíba	63
LB	Natal	91
MTL	Xapuri	85
CGG	Macaé	34
RJ	Rio de Janeiro	66

Dados anonimizados por generalização		
Nome	**Cidade de nascimento**	**Idade**
FM	Acre	70-79
AFB	Rio Grande do Norte	60-69
LB	Rio Grande do Norte	90-99
MTL	Acre	80-89
CGG	Rio de Janeiro	30-39
RJ	Rio de Janeiro	60-69

Fonte: Adaptado de PDPC (2018).

5.2 MASCARAMENTO

Descrição: a técnica consiste em substituir uma parte dos caracteres dos dados por um símbolo (por exemplo, * ou x). O mascaramento é normalmente parcial, ou seja, aplicado apenas a alguns caracteres no atributo (Figura 2).

Quando usar: quando o valor dos dados é uma sequência de caracteres e ocultar parte dela é suficiente para fornecer o grau de anonimato necessário.

Como usar: dependendo da natureza do atributo, substitua os caracteres apropriados por um símbolo escolhido. Dependendo do tipo de atributo, você pode optar por substituir um número fixo de caracteres (por exemplo, para números de cartão de crédito) ou um número variável de caracteres (por exemplo, para endereço de e-mail).

Figura 2 - Exemplo da anonimização por mascaramento

Dados originais		
Nome	**CPF**	**Idade**
FM	111.111.111-11	79
AFB	222.222.222-22	63
LB	333.333.333-33	91
MTL	444.444.444-44	85
CGG	555.555.555-55	34
RJ	666.666.666-66	66

Dados anonimizados por mascaramento		
Nome	**CPF**	**Idade**
FM	111.***.***-11	79
AFB	222.***.***-22	63
LB	333.***.***-33	91
MTL	444.***.***-44	85
CGG	555.***.***-55	34
RJ	666.***.***-66	66

Fonte: adaptado de PDPC (2018).

5.3 PSEUDONIMIZAÇÃO

Descrição: esta técnica consiste na substituição de dados identificadores por valores fictícios. A pseudonimização também é chamada de codificação ou tokenização, consiste na substituição de dados por informações aleatórias denominadas *tokens*. Esse procedimento pode assumir caráter irreversível, quando os valores originais são devidamente descartados e a pseudonimização é realizada de maneira não repetível; ou reversível (pelo detentor dos dados originais), quando os valores originais são guardados de forma segura, mas podem ser recuperados e vinculados novamente ao pseudônimo, caso seja necessário (Figura 3).

Quando usar: quando os valores de dados precisam ser diferenciados de forma única e quando nenhum caractere ou qualquer outra informação implícita do atributo original deve ser mantida.

Como usar: gere uma lista de valores e selecione aleatoriamente dessa lista os valores fictícios para substituir cada um dos valores originais. Os valores fictícios devem ser únicos e não devem ter relação com os valores originais (de modo que se possa derivar os valores originais dos pseudônimos). Se for utilizada criptografia, revise o método de criptografia (por exemplo, algoritmo e comprimento da chave) periodicamente para garantir que seja reconhecido pelo setor como relevante e seguro.

Figura 3 - Exemplo da anonimização por pseudonimização

Dados originais		
Aluno	**Pré-avaliação**	**Horas de aulas necessárias para aprovação**
Camila Duarte	B	20
Henrique Amaral	C	26
Larissa Monteiro	D	03
Rafael Tavares	B	30
Bianca Rodrigues	A	32
Rosa D. Domingues	A	25

Dados anonimizados por pseudonimização		
Aluno	**Pré-avaliação**	**Horas de aulas necessárias para aprovação**
416765	B	20
562396	C	26
964825	D	03
873892	B	30
239976	A	32
943145	A	25

Fonte: adaptado de PDPC (2018).

Para pseudonimização reversível, o banco de dados de identidades é mantido em segurança caso haja uma necessidade legítima futura de identificar indivíduos (Figura 4).

Figura 4 – Dados originais mantidos e codificados

Banco de dados de codificação única	
Camila Duarte	416765
Henrique Amaral	562396
Larissa Monteiro	964825
Rafael Tavares	873892
Bianca Rodrigues	239976
Rosa D. Domingues	943145

Fonte: adaptado de PDPC (2018).

A pseudonimização permite reidentificar e recuperar os dados originais, já a anonimização, em tese, não permitiria essa reidentificação.

5.4 PERMUTA OU TROCA DE VALORES (SWAPPING)

Descrição: o objetivo da troca de valores é reorganizar os dados do conjunto de dados de forma que os valores individuais dos atributos ainda sejam representados pelo conjunto de dados, mas não correspondam aos registros originais (Figura 5).

Quando usar: quando a análise subsequente precisa apenas olhar para dados agregados ou a análise é no nível intra-atributo; em outras palavras, não há necessidade de análise de relacionamentos entre atributos no nível de registro.

Como usar: identifique quais atributos trocar. Em seguida, para cada um, troque ou reatribua os valores dos atributos para qualquer registro no conjunto de dados.

Figura 5 - Exemplo da anonimização por permuta ou troca de valores (swapping)

Dados originais				
Pessoa	**Cargo**	**Nascimento**	**Tipo de associação**	**Visitas mensais**
A	Reitor	03/01/1970	Prata	0
B	Vendedor	05/02/1972	Platina	5
C	Advogado	07/03/1985	Ouro	2
D	Profissional de TI	10/04/1990	Prata	1
E	Enfermeira	13/05/1995	Prata	2

Dados anonimizados por permuta ou troca de valores (swapping)				
Pessoa	**Cargo**	**Nascimento**	**Tipo de associação**	**Visitas mensais**
A	Vendedor	07/03/1985	Prata	0
B	Advogado	03/01/1970	Platina	5
C	Enfermeira	10/04/1990	Prata	2
D	Reitor	13/05/1995	Ouro	1
E	Profissional de TI	05/02/1972	Prata	2

Fonte: adaptado de PDPC (2018) e Guedes, Machado e Costa (2023).

Observação: se o propósito do conjunto de dados anonimizados for estudar as relações entre perfil profissional e padrões de consumo, outros métodos de anonimização podem ser mais adequados.

5.5 PERTURBAÇÃO

Descrição: os valores do conjunto de dados original são modificados para serem ligeiramente diferentes (Figura 6).

Quando usar: para quase-identificadores (normalmente números e datas) que podem ser identificadores quando combinados com outras fontes de dados, e pequenas variações de valor são aceitáveis. Esta técnica não deve ser usada quando a precisão dos dados é crucial.

Como usar: depende da técnica exata de perturbação de dados utilizada. Isso inclui arredondamento e adição de ruído aleatório. No exemplo a seguir (Figura 6), o conjunto de dados contém informações a serem utilizadas em pesquisas sobre possíveis vínculos entre altura, peso, idade, tabagismo e a presença da "doença A" e/ou da "doença B" na pessoa. O nome da pessoa já foi pseudonimizado.

Figura 6 - Exemplo da anonimização por perturbação

Técnica de anonimização aplicada	
Atributo	**Técnica aplicada**
Altura (em cm)	Arredondamento de base 5 (5 é escolhido para ser um pouco proporcional ao valor típico de altura de, por exemplo, 120 a 190 cm)
Peso (em kg)	Arredondamento de base 3 (3 é escolhido para ser um pouco proporcional ao valor de peso típico de, por exemplo, 40 a 100 kg)
Idade (em anos)	Arredondamento de base 3 (3 é escolhido para ser um pouco proporcional ao valor típico de idade de, por exemplo, 10 a 100 anos)
Demais atributos	Nulo, por não ser numérico e difícil de modificar sem alteração substancial de valor

Dados originais						
Pessoa	**Altura**	**Peso**	**Idade**	**Fuma?**	**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

Dados anonimizados por perturbação de dados						
Pessoa	**Altura**	**Peso**	**Idade**	**Fuma?**	**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

Fonte: adaptado de PDPC (2018) e Guedes, Machado e Costa (2023).

Observação: as colunas sombreadas representam os atributos afetados no processo de anonimização.

5.6 AGREGAÇÃO

Descrição: conversão de um conjunto de dados de uma lista de registros em valores resumidos. A agregação pode precisar ser aplicada em combinação com a supressão (Figura 7).

Quando usar: quando registros individuais não são necessários e dados agregados são suficientes para o propósito da pesquisa.

Como usar: use totais, médias etc.

Figura 7 - Exemplo da anonimização por agregação

Dados originais		
Doador	**Renda mensal (R$)**	**Valor doado (R$)**
Doador A		210.00
Doador B	4900,00	420,00
Doador C	2200,00	150,00
Doador D	4200,00	110,00
Doador E	5500,00	260,00
Doador F	2600,00	40,00
Doador G	3300,00	130,00

Doador H	5500,00	210,00
Doador I	1600,00	380,00
Doador J	3200,00	80,00
Doador K	2000,00	440,00
Doador L	5800,00	400,00
Doador M	4600,00	390,00
Doador N	1900,00	480,00
Doador O	1700,00	320,00
Doador P	2400,00	330,00
Doador Q	4300,00	390,00
Doador R	2300,00	260,00
Doador S	3500,00	80,00
Doador T	1700,00	290,00

Dados anonimizados por agregação		
Renda mensal (R$)	**Nº de doações em 2016**	**Soma do valor doado em 2016 (R$)**
1000-1999	4	1470
2000-2999	5	1220
3000-3999	3	290
4000-4999	5	1520
5000-6000	3	870
Total geral	**20**	**5370**

Fonte: adaptado de PDPC (2018).

Observação: detalhes sobre medidas estatísticas não são o objetivo deste Guia.

5.7 SUPRESSÃO DE ATRIBUTOS

Descrição: esta técnica refere-se à remoção de uma parte inteira de dados (também chamada de "coluna" em bancos de dados e planilhas) de um conjunto de dados. Este é o tipo mais forte de técnica de anonimização, porque não há como recuperar qualquer informação do atributo (Figura 8).

Quando usar: quando um atributo não é necessário no conjunto de dados anonimizado ou quando não pode ser adequadamente anonimizado por outra técnica. Esta técnica deve ser aplicada no início do processo de anonimização, pois é uma maneira fácil de diminuir a identificabilidade.

Como usar: remova o(s) atributo(s) ou, se a estrutura do conjunto de dados precisa ser mantida, limpe os dados (não esquecer do cabeçalho). Observe que a supressão deve ser uma remoção real (ou seja, permanente), e não apenas "ocultar a coluna". A supressão de um atributo pode não ser suficiente se os dados subjacentes permanecerem acessíveis.

Figura 8 - Exemplo de anonimização por supressão de atributos

Dados originais		
Estudante	**Treinador**	**Pontuação do teste**
João	Bruna	87
Pedro	Bruna	56
Mirta	Bruna	92
Paulo	Hugo	83
Lucia	Hugo	45
Jaqueline	Hugo	67

Dados anonimizados por supressão de atributos	
Treinador	**Pontuação do teste**
Bruna	87
Bruna	56
Bruna	92
Hugo	83
Hugo	45
Hugo	67

Fonte: adaptado de PDPC (2018).

Observação: como o objetivo era apenas analisar as notas obtidas pelos alunos em relação aos seus diversos instrutores, sem analisar os próprios alunos, o atributo "Aluno" foi removido.

5.8 SUPRESSÃO DE REGISTRO

Descrição: esta técnica se refere à remoção de um registro inteiro de um conjunto de dados. Ao contrário da maioria das outras técnicas, esta afeta **múltiplos atributos ao mesmo tempo** (Figura 9).

Quando usar: para remover registros discrepantes que são únicos ou que não atendem a outros critérios das técnicas de anonimização apresentadas até aqui e não devem ser mantidos no conjunto de dados. Valores discrepantes podem permitir a reidentificação. Esse método pode ser aplicado antes ou depois de outras técnicas (por exemplo, generalização) terem sido utilizadas.

Como usar: exclua o registro inteiro (Figura 9). Observe que a supressão deve ser permanente, e não apenas uma função de "ocultar linha". A supressão de registro pode não ser suficiente se os dados subjacentes permanecerem acessíveis. A remoção de um registro pode impactar o conjunto de dados, por exemplo, nas estatísticas como a média e a mediana.

Figura 9 - Exemplo de anonimização por supressão de registro

Dados originais						
Pessoa	**Altura**	**Peso**	**Idade**	**Fuma?**	**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

Dados anonimizados por supressão de registro						
Pessoa	**Altura**	**Peso**	**Idade**	**Fuma?**	**Doença A**	**Doença B**
198740	160	50	30	Não	Não	Não
398747	158	46	20	Sim	Sim	Não
498732	173	75	22	Não	Não	Não
598772	169	82	44	Sim	Sim	Sim

Fonte: adaptado de PDPC (2018).

6 EXEMPLOS DE DESIDENTIFICAÇÃO DE DADOS EM IMAGEM E VÍDEO

É importante destacar que existem normas e fundamentos legais, como a LGPD, que impõem restrições ao uso de imagens de pessoas, não apenas as captadas em contextos de pesquisa, mas também as de gravações, transmissões e fotografias realizadas em auditórios e outros ambientes públicos ou institucionais.

A imagem de uma pessoa é considerada um dado pessoal. Assim, filmar, fotografar ou transmitir uma palestra, reunião, apresentação etc. em que as pessoas possam ser identificadas configura tratamento de dados, exigindo: base legal, informação prévia aos participantes e finalidade específica (Brasil, 2018). Para divulgação pública (*site*, redes sociais, *YouTube* etc.), o mais seguro é solicitar consentimento explícito ou utilizar técnicas de desidentificação[1].

O *Código Civil Brasileiro* também proíbe o uso público da imagem de pessoas sem autorização quando houver: finalidade comercial, prejuízo à honra, exposição indevida e/ou falta de interesse público (Brasil, 2002).

Na prática, em auditórios e outros ambientes públicos ou institucionais, é necessário: informar previamente que o evento será gravado ou fotografado, colocar avisos visíveis na entrada, coletar consentimento quando houver possibilidade de identificação clara dos participantes e restringir o enquadramento às áreas autorizadas (preferir o palestrante e o palco). Recomenda-se evitar: *close* no público, divulgação ampla do vídeo sem autorização, exposição de pessoas que não foram informadas.

Para proteger a privacidade de pessoas em imagens e vídeos, de acordo com a LGPD, o Código Civil Brasileiro e o GDPR, pode-se utilizar *softwares* e ferramentas de código aberto para desidentificação de imagens e vídeos, alguns exemplos são:

1 A diferença entre desidentificação e anonimização de imagens e vídeos na seção 4.2

- Para Imagens:

  - OpenCV + Python[2]: biblioteca de visão computacional com funções para blur, pixelização, detecção facial (usando Haar Cascades ou modelos DNN como YOLO).

  - TensorFlow / PyTorch[3]: permite treinar ou usar modelos pré-treinados (MTCNN, FaceNet, RetinaFace) para detecção e desfocagem.

  - DeepPrivacy[4]: usa redes generativas (GANs) para substituir rostos por versões sintéticas.

  - Microsoft Presidio[5]: framework para anonimização de dados (incluindo imagens e textos).

  - Dlib[6]: biblioteca C++/Python com detecção facial avançada (HOG + SVM ou CNN).

  - PULSE (Self-Supervised Photo Upsampling)[7] pode ser adaptado para gerar rostos anônimos em baixa resolução.

- Para Vídeos:

  - FFmpeg[8]: ferramenta CLI para edição de vídeo, podendo aplicar filtros de blur (boxblur, gblur) ou ser combinada com OpenCV.

  - NVIDIA Video Processing Framework (VPF)[9]: Aceleração por GPU para processamento de vídeo com Python.

  - OpenFace[10]: Toolkit para análise facial que pode ser usado para detecção antes da anonimização.

2 https://opencv.org/

3 https://www.tensorflow.org/

4 https://github.com/hukkelas/DeepPrivacy

5 https://github.com/microsoft/presidio

6 http://dlib.net/

7 https://github.com/adamian98/pulse

8 https://ffmpeg.org/

9 https://github.com/NVIDIA/VideoProcessingFramework

10 https://github.com/TadasBaltrusaitis/OpenFace

- Para dados médicos/sensíveis:

  - DICOM Anonymizer[11] (via PyDICOM): remove metadados e pixels identificáveis em imagens médicas.

  - AutoRedact[12]: redação automática de informações sensíveis em documentos/imagens.

  - Ferramentas com GUI (Open Source)

  - ObscuraCam[13] (Android): App para anonimização de fotos em dispositivos móveis. ImageJ + plugins[14]: usado em pesquisas científicas para processamento de imagens (pode ser adaptado para desidentificação).

6.1 TÉCNICAS DE DESIDENTIFICAÇÃO FACIAL

A identificação facial ocorre quando uma imagem facial é devidamente associada a identificadores explícitos, como nome e endereço do sujeito da imagem facial. A identificação explícita é uma grave preocupação de privacidade. A Figura 10 mostra os resultados da identificação facial em que (a) e (b) são identificadas como “John Smith” (c).

O reconhecimento facial, em oposição à identificação facial, relaciona uma imagem facial a outra “conhecida”, que pode ou não ser explicitamente identificada. Reconhecer um rosto não é necessariamente o mesmo que identificar uma pessoa, pois as identidades dos sujeitos de um conjunto de rostos podem não ser conhecidas. Alterar imagens faciais para ocultar a identidade é chamado de desidentificação facial.

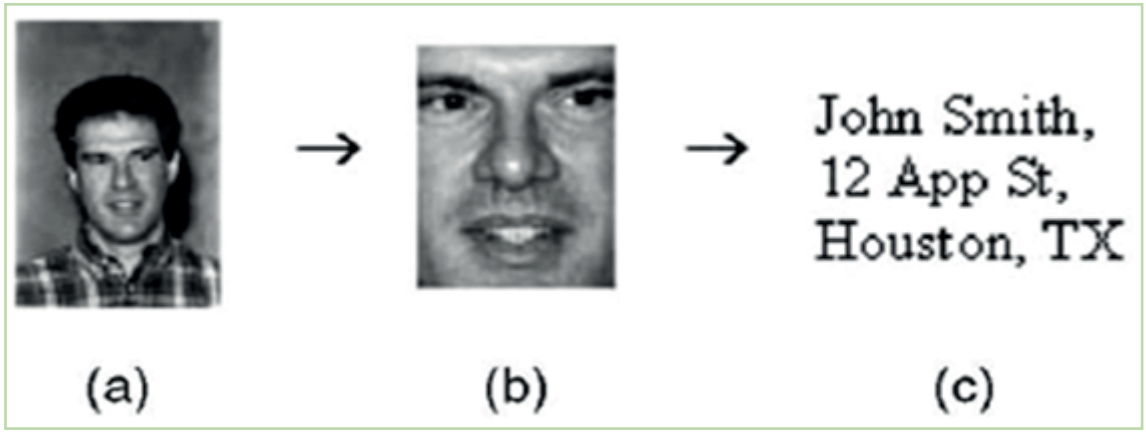
11 https://dicomanonymizer.com/

12 https://nebulahelp.ediscovery.com/Content/Topics/Review/AutoRedact/AutoRedactJobs.htm

13 https://guardianproject.info/apps/obscuracam/

14 https://imagej.net/

Figura 10 - Identificação facial



Fonte: adaptado de Newton, Sweeney e Malin (2005).

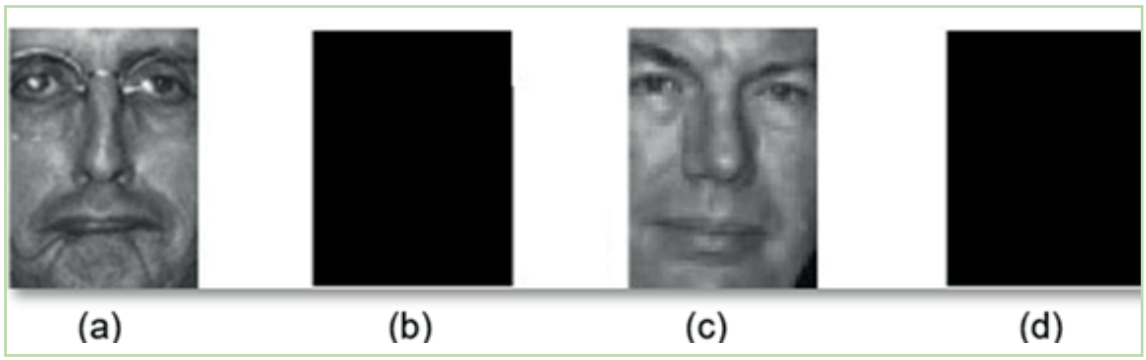
Desidentificar uma imagem facial pode oferecer alguma proteção de privacidade, mas o próprio ato de desidentificação não oferece garantia de privacidade total, pois outros detalhes podem permanecer na imagem e permitir reidentificar o sujeito.

A seguir, serão apresentadas as técnicas mais comuns de desidentificação de imagens faciais. Os exemplos baseiam-se principalmente nas obras de Sweeney (2002) e Newton, Sweeney e Malin (2005).

6.1.1 *BlackOut* (máscara completa)

Esta técnica garante uma proteção de privacidade eficaz em contextos em que nenhum detalhe facial deve ser fornecido. A técnica de máscara completa ou *BlackOut* colore toda a imagem do rosto com uma única cor ou padrão. No exemplo da Figura 11, nenhum humano ou máquina pode determinar a identidade do sujeito das imagens (a) ou (c) porque (b) e (d) são idênticas.

Figura 11 - Desidentificação facial por *BlackOut*

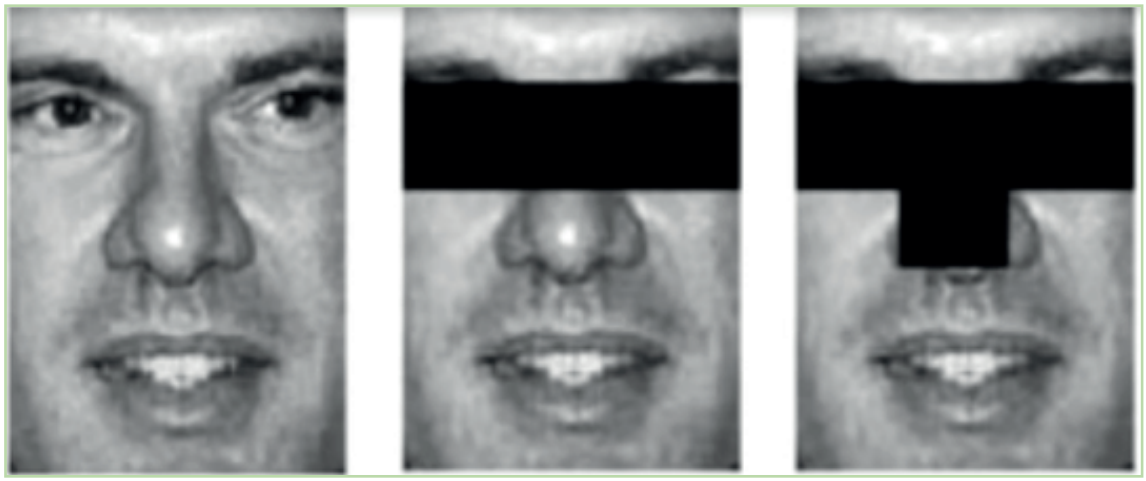


Fonte: adaptado de Newton, Sweeney e Malin (2005).

6.1.2 Máscara parcial

É a técnica de desidentificação conhecida como máscara em barra e máscara em T; remove ou oculta regiões críticas do rosto, elimina detalhes essenciais para diferenciar uma pessoa de outra e quebra a correspondência direta entre a imagem original e a imagem alterada. As regiões ocultadas geralmente incluem olhos (muito importantes para reconhecimento), ponte do nariz, área da boca e as proporções faciais centrais (Figura 12).

Figura 12 - Desidentificação facial por Máscara de barra e Máscara T



Fonte: adaptado de Newton, Sweeney e Malin (2005).

Muitos métodos simples de desidentificação conseguem enganar um sistema de reconhecimento facial básico que apenas compara imagens originais com imagens alteradas, sem necessitar de técnicas avançadas. Nesse tipo de "reconhecimento ingênuo[15]", de acordo com os testes de Newton, Sweeney e Malin (2005), a taxa de acertos foi muito baixa: 2% para a máscara em barra, 1% para a máscara em T e 0% para o apagamento total da face por *BlackOut*. Em contrapartida, quando imagens originais foram comparadas entre si, o reconhecimento foi 100% correto.

As técnicas de desidentificação enganam o "reconhecimento ingênuo" porque exploram justamente as limitações desse tipo de reconhecimento, que é simples e depende muito da aparência visual direta do rosto. Este tipo de reconhecimento compara uma imagem com outra (galeria x prova), baseia-se em semelhança visual ou em características básicas do rosto e assume que os olhos, o nariz, a boca e a estrutura facial estão claramente visíveis. O "reconhecimento ingênuo" não é preparado para tratar de partes do rosto escondidas, perda de informação facial e alterações artificiais introduzidas de propósito.

15 Reconhecimento ingênuo (*naïve recognition*) é um método simples de reconhecimento facial, usado como referência experimental, que falha facilmente quando a imagem do rosto sofre alterações ou oclusões (Newton; Sweeney; Malin, 2005).

6.1.3 Pixelização

Reduz o número de valores de pixel distintos em uma imagem facial, substituindo um bloco quadrado de valores de pixel por seu valor médio. Na Figura 13, a pixelização foi realizada em blocos de 15, 20 e 30 pixels.

Figura 13 - Desidentificação parcial por pixelização

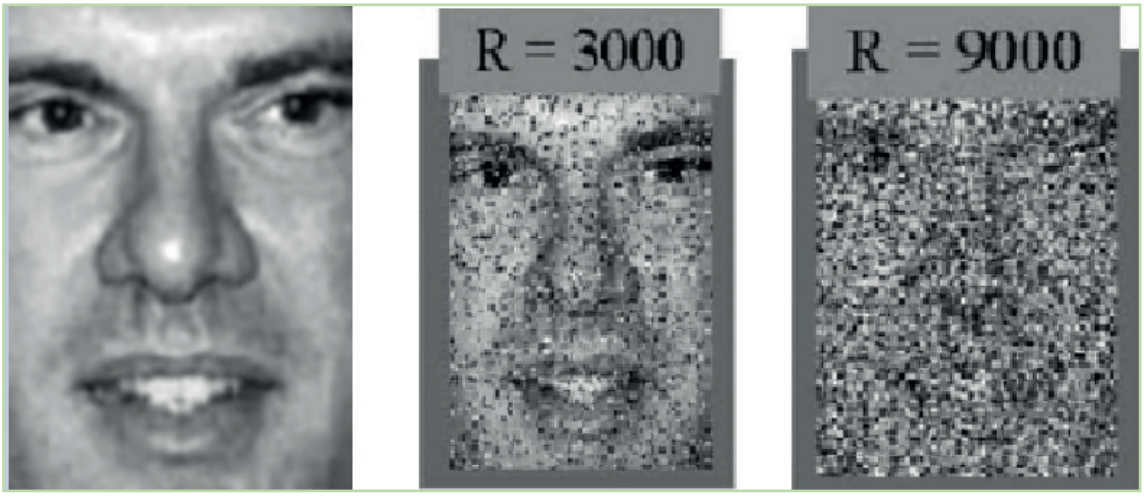
Fonte: adaptado de Newton, Sweeney e Malin (2005).

Em testes realizados por Newton, Sweeney e Malin (2005), a pixelização permitiu 99% de reconhecimento ao combinar imagens originais com imagens pixeladas. Apesar de a imagem parecer um pouco desidentificada para os humanos e apesar do uso comum da pixelização na televisão para esconder rostos durante entrevistas, essa técnica praticamente não tem efeito em impedir o reconhecimento facial por meio de *software* simples.

6.1.4 Perturbação por ruído aleatório

A desidentificação por ruído aleatório é uma técnica de desidentificação facial no nível de pixels que consiste em alterar aleatoriamente os valores dos pixels de uma imagem, com o objetivo de dificultar ou impedir o reconhecimento da identidade da pessoa por sistemas automáticos de reconhecimento facial (Figura 14).

Figura 14 - Imagem facial desidentificada por ruído aleatório



Fonte: adaptado de Newton; Sweeney; Malin (2005).

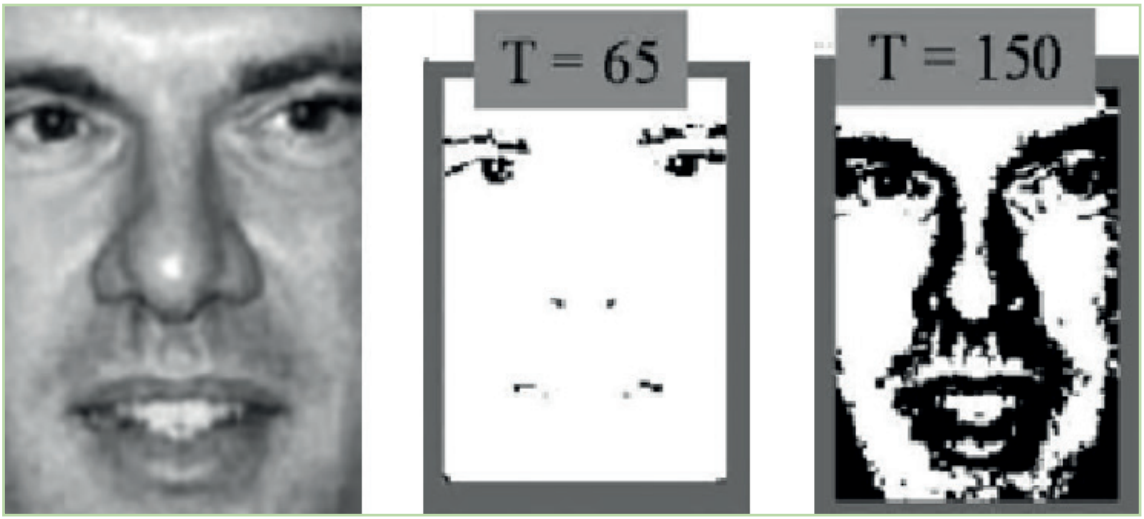
Quando, em 9.000 de 13.266 pixels (68%), foi acrescentado ruído aleatório, o reconhecimento pode ser de apenas 1%; entretanto, em 3.000 pixels perturbados, o reconhecimento pode chegar a 23%. Em imagens em preto e branco, o ruído aleatório reduz o reconhecimento após aproximadamente metade dos valores de pixel serem perturbados (Newton; Sweeney; Malin, 2005).

6.1.5 Limiarização (*thresholding*)

A limiarização (*thresholding*) é uma técnica de desidentificação facial que converte imagens em tons de cinza em imagens binárias (Figura 15), reduzindo drasticamente os detalhes faciais ao preservar apenas contrastes fortes, o que pode dificultar o "reconhecimento ingênuo", mas apresenta eficácia limitada e a substancial perda de informação. Quando aplicada a imagens faciais, remove texturas, sombras e gradações sutis da face, simplifica a aparência do rosto e preserva apenas os contornos grosseiros e áreas de contraste elevado.

O reconhecimento por sistemas ingênuos (simples) depende fortemente de padrões de intensidade e de textura. A limiarização altera esses padrões ao reduzir a imagem a dois valores possíveis, dificultando a correspondência direta entre imagens originais e limiarizadas. Entretanto, resultados experimentais mostram que certos valores de limiar ainda preservam estrutura suficiente para permitir reconhecimento considerável e, em alguns casos, o reconhecimento pode ser alto, indicando que a limiarização nem sempre é uma técnica eficaz para a desidentificação facial.

Figura 15 - desidentificação facial por limiarização (*thresholding*)



Fonte: adaptado de Newton, Sweeney e Malin (2005).

Para aplicar esta técnica de desidentificação facial, costuma-se testar vários valores para avaliar o impacto no reconhecimento. O limiar (T) é um número entre 0 e 255 que define a conversão dos pixels, em que pixel ≤ T, preto (0), e pixel > T, branco (255). Exemplos:

- T baixo (ex.: 60–80) - imagem mais clara, preserva contornos fortes,
- T médio (ex.: 100–130) - equilíbrio entre áreas claras e escuras,
- T alto (ex.: 150–180) - imagem mais escura, muitos detalhes perdidos.

Ao aplicar limiarização para desidentificação facial, é necessário considerar a sensibilidade ao limiar porque pequenas mudanças em T alteram drasticamente o resultado e, consequentemente, a perda de utilidade da imagem, que pode se tornar inadequada para outras análises, já que esta técnica de desidentificação pode falhar se aplicado o reconhecimento reverso.

6.1.6 *k-Same*

O algoritmo *k*-Same é uma técnica clássica de desidentificação facial baseada em *k*-anonimato, cujo objetivo é garantir que cada face desidentificada seja indistinguível de pelo menos outras *k - 1* faces dentro de um conjunto conhecido (Figura 16). Este algoritmo foi proposto para oferecer garantias de privacidade contra sistemas automáticos de reconhecimento facial. Após a desidentificação, uma face não pode ser associada unicamente a um indivíduo, mas a um grupo de *k* indivíduos possíveis. Isso significa que a probabilidade de reconhecimento correto é limitada a, no máximo, *1/k*.

Figura 16 - Desidentificação facial usando técnica *k-Same*

Fonte: adaptado de Newton, Sweeney e Malin (2005).

Meng e Shenoy (2017) descrevem variações do algoritmo, como *k-Same Pixel*, em que a face desidentificada é a média pixel das *k*-faces mais próximas, e *k-Same Eigen*, em que a média é calculada no espaço de componentes principais (eigenfaces) e depois reconvertida para imagem. Essas variações mantêm o princípio de *k*-Anonimato, mas diferem na qualidade visual e na preservação de informações.

De acordo com as mesmas autoras o *k-Same* não foi projetado para preservar a utilidade dos dados e apresenta limitações como:

- Perda de utilidade dos dados: a média tende a apagar características importantes como gênero, idade e expressão facial;
- Artefatos visuais: as faces médias podem apresentar efeito "fantasma";
- Conjunto fechado: as garantias de privacidade valem apenas se o ataque for restrito ao conjunto conhecido de faces.

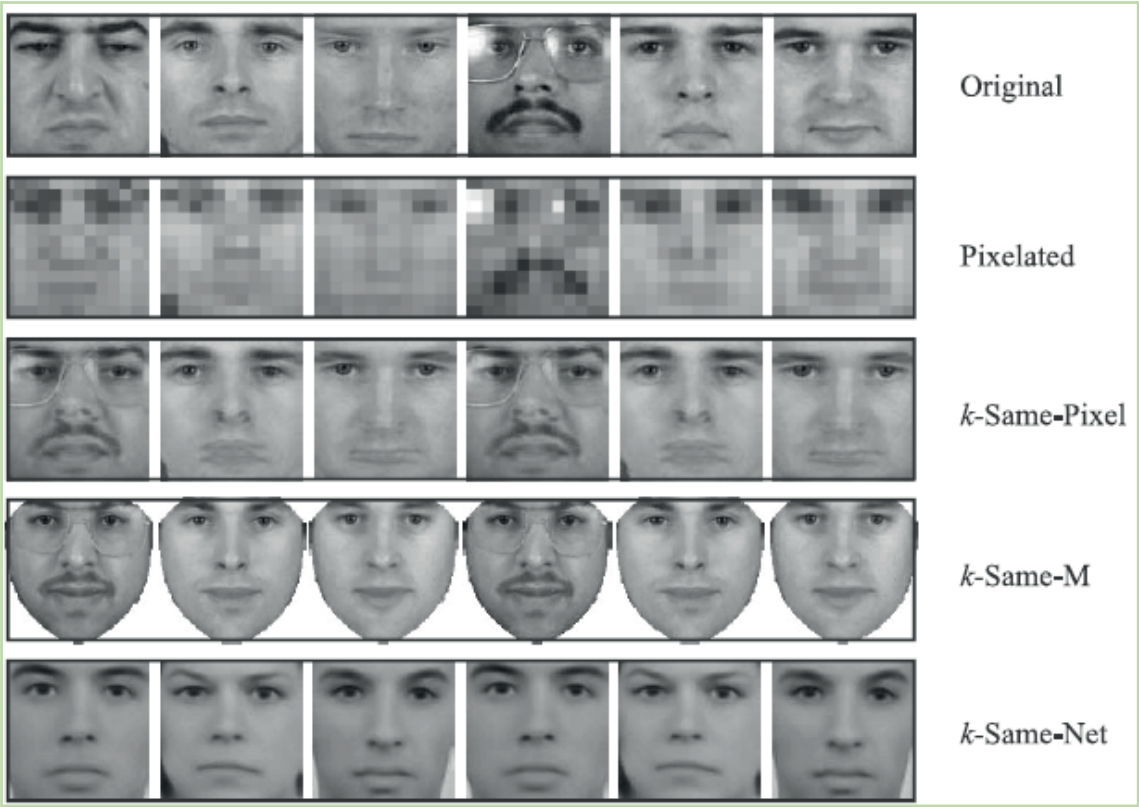
A versão desidentificada da Figura 17 mostra um rosto masculino que pode parecer feminino e um rosto neutro que pode exibir um sorriso, evidenciando algumas das limitações apontadas acima.

Figura 17 - Limitações apresentadas pela técnica *K-Same*

Fonte: adaptado de Meng e Shenoy (2017).

Essas limitações motivaram extensões modernas do método, usando redes generativas para melhorar a qualidade visual, mantendo o *k*-Anonimato, como mostra a Figura 18, onde cada conjunto de imagens faciais desidentificadas difere na qualidade visual e na quantidade de conteúdo de informação preservado (Meden *et al*., 2018).

Figura 18 - Imagens faciais originais e desidentificadas por várias técnicas *K-Same*



Fonte: Meden *et al.* (2018).

Conceitualmente, o *k-Same* segue os seguintes passos:

- Representação das faces: as imagens faciais são convertidas para uma representação comum (por exemplo, vetores de pixels ou espaço de componentes principais);
- Busca dos *k* vizinhos mais próximos: para cada face original, o algoritmo identifica as *k* faces semelhantes dentro de um conjunto público (galeria);
- Geração da face desidentificada: as *k* faces selecionadas são combinadas (normalmente por média), produzindo uma única face sintética;
- Substituição: todas as *k* faces originais são substituídas por essa mesma face média, tornando-as indistinguíveis entre si.

6.2 ANONIMIZAÇÃO DE IMAGENS MÉDICAS

Tanto a LGPD, o *General Data Protection Regulation* (GDPR) da União Europeia e a *Health Insurance Portability and Accountability Act* (HIPAA) dos Estados Unidos impõem a anonimização ou desi-

dentificação de dados de saúde, incluindo imagens médicas (ressonância magnética, tomografia computadorizada e radiografias), como condição para o uso secundário.

No entanto, a anonimização de imagens médicas apresenta desafios específicos, uma vez que identificadores podem estar presentes tanto nos metadados quanto no conteúdo visual.

Existem inúmeras maneiras diferentes de armazenar os dados médicos digitais, devido à ampla gama de técnicas, os mais populares são: padrão *Digital Imaging and Communications in Medicine (DICOM), Neuroimaging Informatics Technology Initiative* (NIfTI), *Medical Imaging* (MINC), *Joint Photographic Experts Group* (JPEG) *e Portable Network Graphics* (PNG) (Larobina; Murino, 2014).

O DICOM representa o principal formato para armazenamento e intercâmbio de imagens médicas, combinando dados de imagem e metadados. Em ambientes clínicos, imagens JPEG costumam ser encapsuladas dentro de DICOM, mantendo o JPEG original, mas com cabeçalho DICOM.

A anonimização nesse contexto envolve a aplicação de perfis de confidencialidade, a remoção de identificadores diretos e a mitigação de riscos de reidentificação associados a textos embutidos e estruturas anatômicas visíveis, especialmente em exames do crânio (Rachel *et al*. 2025).

A anonimização de imagens médicas, de acordo com Rempe *et al*. (2025), consiste em várias tarefas separadas:

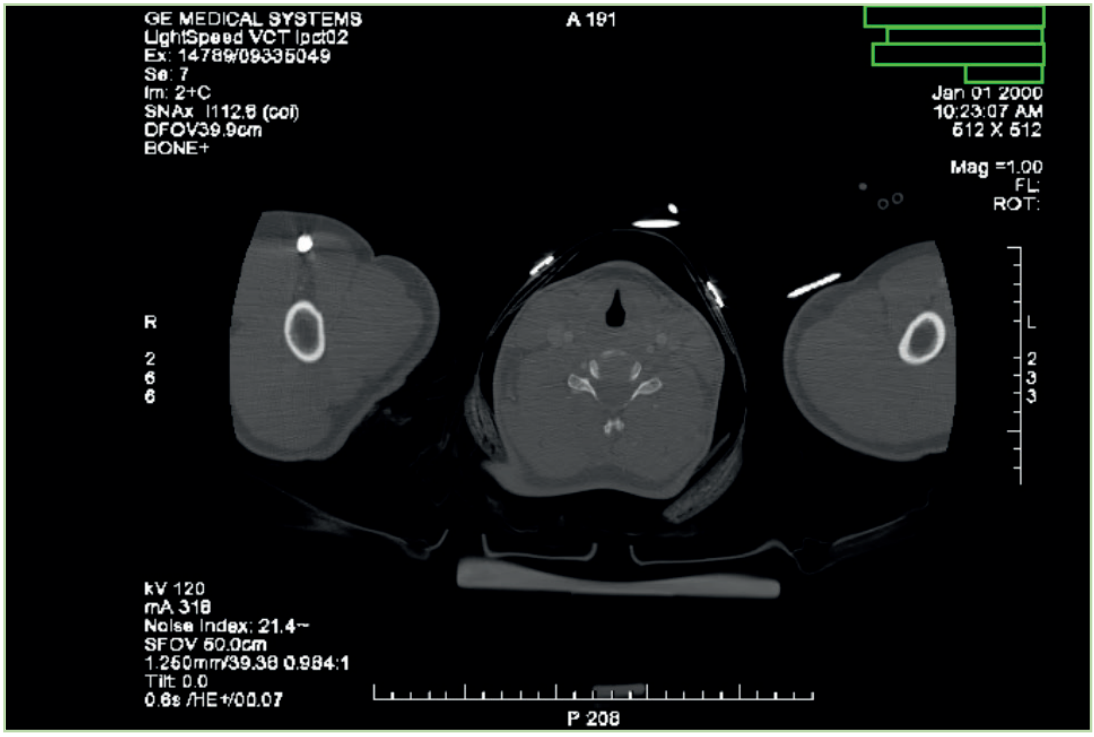
- anonimização de metadados;
- remoção de elementos visuais (desfiguração);
- remoção do crânio;
- remoção de texto e
- anonimização de imagens de lâminas inteiras, incluindo anotações embutidas (burn-in annotations) potencialmente reidentificáveis.

Em muitas situações, utilizar ferramentas para anonimização de anotações embutidas (burned-in annotations) nas imagens médicas pode ser suficiente para a desidentificação dos dados pessoais do paciente. Na Figura 19, encontram-se as informações do paciente e com informações suprimidas.

Figura 19 - Imagem de ressonância magnética, original

Fonte: Madeleine (2022).

Imagem de ressonância magnética com informações suprimidas

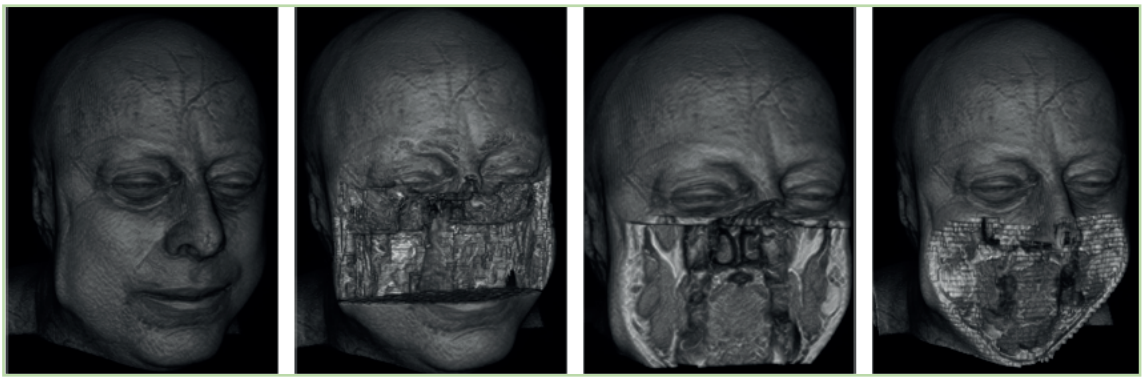


Fonte: Madeleine (2022)

Se a imagem incluir anotações gravadas, a desidentificação também deve ser realizada por meio de uma "limpeza" dos pixels da imagem.

Na Figura 20, encontra-se um modelo de desfiguração de imagem e a comparação dos resultados de desfiguração utilizando diferentes algoritmos. Aplicando o modelo de reconhecimento facial pode-se distinguir entre resultados de desfiguração corretos e falhos. A pontuação de desfiguração é calculada pela quantidade de digitalizações desfiguradas que o modelo de reconhecimento facial não consegue mais classificar como um rosto.

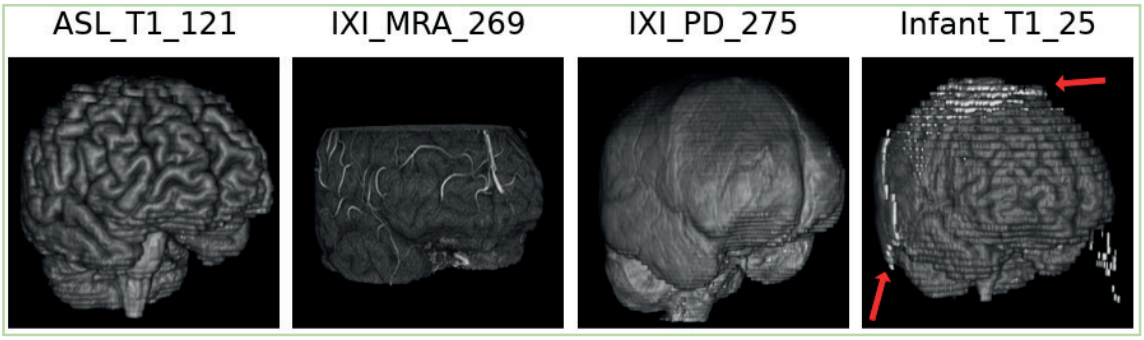
Figura 20 – Desfiguração de imagens da face



Fonte: Rempe *et al*. (2025).

A Figura 21 apresenta o resultado da utilização de um algoritmo de remoção do crânio. O algoritmo proposto por Rempe (2025) é capaz de realizar a tarefa de remoção do crânio em imagens produzidas a partir de ressonância magnética, mas apresenta dificuldades com exames cerebrais de bebês, deixando alguns resíduos de crânio (setas vermelhas). Dependendo do algoritmo utilizado, os resultados podem ser satisfatórios para alguns tipos de exames e apresentar dificuldade em outros.

Figura 21 – Desidentificação de crânio em imagens de ressonân-cia magnética



Fonte: Rempe *et al*. (2025).

Os rótulos do conjunto de dados das imagens de crânio: ASL_T1_121, IXI_MRA_269, IXI_PD_275 e Infant_T1_25 combinam modalidade de ressonância magnética, tipo de contraste e identificador do sujeito e do exame (Quadro 3).

Quadro 3 – Identificação do conjunto de imagens de crâneo

Rótulo	Dataset	Modalidade	Tipo de imagem	Sujeitos
ASL_T1_121	-	ASL + T1	Perfusão + anatomia	121
IXI_MRA_269	IXI	MRA	Angiografia	269
IXI_DP_275	IXI	DP	Densidade de prótons	275
Infant_T1_25	Pedriátrico	T1	Anatomía infantil	25

Fonte: Rempe *et al*. (2024).

Os desafios da anonimização de imagens médicas envolvem questões técnicas, legais e éticas. Durante muitos anos, a anonimização concentrou-se quase exclusivamente na remoção de metadados (nome do paciente, data de nascimento, ID) presentes em arquivos DICOM. No entanto, estudos recentes mostram que informações identificáveis também podem estar embutidas no próprio conteúdo da imagem, em nível de pixel, como padrões anatômicos, marcas de aquisição e características faciais visíveis em exames cranianos (Giouroukou *et al*. 2025). Portanto, mesmo após remover todos os dados sensíveis, o risco de reidentificação permanece.

Atualmente, não existe um consenso técnico universal que defina quando uma imagem médica pode ser considerada completamente anonimizada. Regulamentos fornecem princípios gerais, mas a avaliação de risco ainda depende do contexto de uso e do modelo de ameaça (Yanming, 2024). O desafio é determinar quando o risco é aceitável para compartilhamento ou pesquisa.

REFERÊNCIAS

A GUIDE to Confidentiality in Health and Social Care. 2013. London: NHS England. Disponível em: https://digital.nhs.uk/data-and-information/looking-after-information/data-security-and-information-governance/codes-of-practice-for-handling-information-in-health-and-care/a-guide-to-confidentiality-in-health-and-social-care/a-guide-to-confidentiality. Acesso em: 8 mar. 2026.

ALBAGLI, S.; MACIEL, M. L.; ADBO, A. H. (org.). **Ciência aberta, questões abertas**. Brasília: Ibict; Rio de Janeiro: Unirio, 2015. Disponível em: https://livroaberto.Ibict.br/bitstream/1/1060/1/Ciencia%20aberta_questoes%20abertas_PORTUGUES_DIGITAL%20(5).pdf. Acesso em: 10 jul. 2025.

and on the free movement of such data, and repealing Directive 95/46/EC. Uniao Europeia, 2016.

BRASIL. **Lei nº 10.406, de 10 de janeiro de 2002**. Institui o Código Civil. Diário Oficial da União: seção 1, Brasília, DF, 11 jan. 2002.

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018**. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília: Presidência da República, 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 17 dez. 2024.

CURTY, R. Abordagens de reuso e a questão da reusabilidade dos dados científicos. **Liinc Em Revista**, Rio de Janeiro, v. 15, n. 2, 2019. Disponível em: https://doi.org/10.18617/liinc.v15i2.4777. Acesso em: 5 jul. 2025.

Disponível em: http://data.europa.eu/eli/reg/2016/679/oj. Acesso em: 8 mar. 2026.

GABRIEL JÚNIOR, R. F. *et al*. Acesso aberto a dados de pesquisa no Brasil: mapeamento de repositórios, práticas e percepções dos pesquisadores e tecnologias. **Ciência da Informação**, Brasília, DF, v. 48, n. 3, p. 87-101, set./dez. 2019. Suplemento. Disponível em: https://lume.ufrgs.br/handle/10183/212266. Acesso em: 7 jul. 2025.

GIOUROUKOU, K. *et al*. Rethinking privacy in medical imaging AI: from metadata and pixel-level identification risks to federated learning and synthetic data challenges. **Radiology Artificial Intelligence**, v. 8, n. 1, 2025. DOI: 10.1148/ryai.250273. Disponível em: https://pubmed.ncbi.nlm.nih.gov/41295085/. Acesso em: 1 mar. 2026.

GUEDES, M. S.; MACHADO, D. C.; COSTA, A. F. J. **Estudo técnico sobre anonimização de dados na LGPD**: uma visão de processo baseado em risco e técnicas computacionais versão 1.0. Brasília: ANPD, 2023. Disponível em: https://www.gov.br/anpd/pt-br/centrais-de-conteudo/documentos-tecnicos-orientativos/estudo_tecnico_sobre_anonimizacao_de_dados_na_lgpd_uma_visao_de_processo_baseado_em_risco_e_tecnicas_computacionais.pdf. Acesso em: 7 jul. 2025.

HURD, J. M. The transformation of scientific communication: A model for 2020. **Journal of the American Society for Information Science**, v. 51, n. 14, p. 1279-1283, 2000. Disponível em: http://www.ou.edu/ap/lis5703/sessions/hurd.pdf. Acesso em: 20 ago. 2025.

INTERNATIONAL HOUSEHOLD SURVEY NETWORK. Anonymization Principles. Disponível em: https://ihsn.org/. [2025]. Acesso em: 7 ago. 2025.

LAROBINA, M.; MURINO, L. Medical image file formats. **Journal of Digital Imaging**, v. 27, p. 200-206, 2014. DOI: 10.1007/s10278-013-9657-9. Disponível em: https://pubmed.ncbi.nlm.nih.gov/24338090/. Acesso em: 3 mar 2026.

LI, N.; LI, T.; VENKATASUBRAMANIAN, S. t-Closeness: privacy beyond k-anonymity and ℓ-diversity. In: IEEE 23rd International Conference on Data Engineering. **Anais**... Istanbul: IEEE, 2007. p. 106 115. DOI: 10.1109/ICDE.2007.367856. Disponível em: https://ieeexplore.ieee.org/document/4221659. Acesso em: 3 mar. 2026.

MADELEINE, S. Ferramentas gratuitas para anonimização de imagens médicas. **Blog IMAIOS**. 2022. Disponível em: https://www.imaios.com/br/recursos/blog/5-melhores-ferramentas-de-desidentificacao-dicom. Acesso em: 3 mar. 2026.

MEADOWS, Arthur Jack. **Acomunicação científica**. Brasília: Briquet de Lemos, 1999.

MEDEN, B. et al. K-Same-Net: K-Anonymity with generative deep neural networks for face deidentification. **Entropy**, v. 20, n. 1, 2018. DOI: 10.3390/e20010060. Disponível em: https://www.mdpi.com/1099-4300/20/1/60. Acesso em: 5 mar. 2026.

MENG, L.; SHENOY, A. Retaining expression on De-identified faces. In: Speech and Computer: Lecture Notes in Computer Science book series. **LNCS**, v. 10458, p. 651-661, 2017. DOI: 10.1007/978-3-319-66429-3_65. Disponível em: https://uhra.herts.ac.uk/id/eprint/14040/. Acesso em: 3 mar. 2026.

NEWTON E.; SWEENEY L.; MALIN B. Preserving Privacy by De-identifying Facial Images. **IEEE Transactions on Knowledge and Data Engineering**, 2005. Disponível em: https://www.researchgate.net/publication/3297373_Preserving_privacy_by_de-identifying_face_images. Acesso em: 7 abr. 2025.

PERSONAL DATA PROTECTION COMMISSION SINGAPORE. **Guide to basic data anonymisation techniques**. Singapore: PDPCS, 2018. Disponível em: https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(250118).pdf. Acesso em: 7 abr. 2025.

RACHEL, B. et al. Medical imaging privacy: a systematic scoping review of key parameters in dataset construction and data protection. **Journal of Medical Imaging and Radiation Sciences**, v. 56, n. 5, 2025. DOI: 10.1016/j.jmir.2025.101914. Disponível em: https://pubmed-ncbi-nlm-nih-gov.ez45.periodicos.capes.gov.br/40288182/. Acesso em: 5 mar. 2026.

REMPE, M. et al. De-identification of medical imaging data: a comprehensive tool for ensuring patient privacy. **European Radiology**, v. 5, n. 12, 2025. DOI: 10.1007/s00330-025-11695-x. Disponível em: https://arxiv.org/pdf/2410.12402. Acesso em: 5 mar 2026.

SANCHEZ, F. A.; VIDOTTI, S. A. B. G.; VECHIATO, F. L. A contribuição da curadoria digital em repositórios digitais. **Revista Informação na Sociedade Contemporânea**, Natal, p. 11-17, 2017. Número especial. Disponível em: https://periodicos.ufrn.br/informacao/article/download/12280/8508. Acesso em: 26 set. 2023.

SKLOOT, R. **A vida imortal de Henrietta Lacks**. São Paulo: Companhia das Letras. 2011. 454 p.

SWEENEY, L. k-anonymity: a model for protecting privacy. **International Journal on Uncertainty**, v. 10, n. 5, p. 557-570, 2002. Disponível em: https://epic.org/wp-content/uploads/privacy/reidentification/Sweeney_Article.pdf. Acesso em: 3 mar. 2026.

TARGINO, M. G. Comunicação científica: uma revisão de seus elementos básicos. **Informação e Sociedade Estudos**, João Pessoa, v. 10, n. 2, p. 37-85, 2000. Disponível em: https://periodicos.ufpb.br/ojs/index.php/ies/article/view/326. Acesso em: 7 jul. 2025.

UNIAO EUROPEIA. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data

VARGAS, A. G. *et al*. Tratamento de dados pessoais para fins acadêmicos e para a realização de estudos e pesquisas: guia orientativo. Brasília: ANPD, 2023. 58 p. Disponível em: https://www.gov.br/anpd/pt-br/centrais-de-conteudo/materiais-educativos-e-publicacoes/web-guia-anpd-tratamento-de-dados-para-fins-academicos.pdf. Acesso em: 3 mar 2026.

YANMING, Z. *et al*. Privacy-Preserving in Medical Image Analysis: A Review of Methods and Applications. In: Li, Y., Zhang, Y., Xu, J. (ed.) Parallel and distributed computing, applications and technologies. PDCAT 2024. **Lecture Notes in Computer Science**, v. 15502. Springer, Singapore. DOI: 10.1007/978-981-96-4207-6_15. Disponível em: https://link.springer.com/chapter/10.1007/978-981-96-4207-6_15. Acesso em: 2 mar 2026

SOBRE OS AUTORES

CATERINA GROPOSO PAVÃO

Bacharel em Biblioteconomia pela Universidade Federal do Rio Grande do Sul, mestrado e doutorado em Comunicação e Informação pelo Programa de Pós-Graduação em Comunicação e Informação (PPGCOM UFRGS) com doutorado sanduíche na Universidad Complutense de Madrid. Docente do Departamento de Ciência da Informação da Faculdade de Biblioteconomia e Comunicação da Universidade Federal do Rio Grande do Sul e do Programa de Pós-Graduação em Ciência da Informação (PPGCIN UFRGS). Membro do Grupo de Pesquisa de Comunicação Científica da UFRGS e do Núcleo de Estudos em Ciência, Inovação e Tecnologia (NECIT).

https://orcid.org/0000-0003-3712-7200

https://lattes.cnpq.br/4834791532698069

caterina@cpd.ufrgs.br

LETÍCIA GUARANY BONETTI

Bibliotecária pela Universidade de Brasília (2019) e Mestre em Ciência da Informação pela UFSCar (2023). Doutoranda no Programa de Pós-graduação em Ciência da Informação do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict). Atua como Tecnologista no Ibict desenvolvendo serviços na área de Ciência Aberta e repositórios de dados. Coordenadora adjunta da Rede Brasileira de Repositórios Digitais (RBRD). Pesquisa em Ciência da Informação, com foco em gestão de dados de pesquisa, repositórios, metadados e princípios FAIR. Foi bolsista pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) durante o mestrado.

https://orcid.org/0000-0002-3012-8465

https://lattes.cnpq.br/1895977717955732

leticiabonetti@Ibict.br

MARCEL GARCIA DE SOUZA

Doutorando em Ciência da Informação pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict). Mestre em Educação em Ciências pela Universidade Federal do Rio Grande do Sul (2016). Graduado em Psicologia pela Universidade Católica de Brasília (2005). Servidor público federal; Analista em Ciência e Tecnologia no Instituto Brasileiro de Informação em Ciência e Tecnologia atuando como Coordenador de Tratamento, Análise e Disseminação da Informação Científica, além de coordenar pesquisas aplicadas voltadas à Ciência da Informação, Ciência Aberta, Informação para Sustentabilidade e Informação Tecnológica.

https://orcid.org/0000-0003-2255-199X

https://lattes.cnpq.br/9517728665816047

marcelsouza@Ibict.br

RENE FAUSTINO GABRIEL JUNIOR

Graduação em Biblioteconomia e Documentação pela Pontifícia Universidade Católica do Paraná, mestrado em Ciência, Gestão e Tecnologia da Informação pela Universidade Federal do Paraná e doutorado em Ciência da Informação pela Universidade Estadual Paulista Júlio de Mesquita Filho. Atualmente é professor adjunto da Universidade Federal do Rio Grande do Sul. Tem experiência na área de Ciência da Informação, com ênfase em Biblioteconomia, atuando principalmente nos seguintes temas: bibliometria, BRAPCI, ciência da informação, comunicação científica e produção científica. Implantou e coordena a Base de Dados de Periódicos em Ciência da Informação (BRAPCI). Membro do Grupo de Pesquisa de Comunicação Científica da UFRGS e do Núcleo de Estudos em Ciência, Inovação e Tecnologia (NECIT).

https://orcid.org/0000-0003-1021-3360

https://lattes.cnpq.br/5900345665779424

rene.gabriel@ufrgs.br

SAMILE ANDREA DE SOUZA VANZ

Professora titular do Departamento de Ciências da Informação, do Programa de Pós-graduação em Comunicação (PPGCOM UFRGS) e do Programa de Pós-graduação em Ciência da Informação da Universidade Federal do Rio Grande do Sul (PPGCIN UFRGS). Graduada em Biblioteconomia pela Universidade Federal do Rio Grande do Sul (1999), mestre e doutora em Comunicação e Informação pelo PPGCOM UFRGS (2004 e 2009), com estágio sanduíche na Dalian University of Technology (China, 2007-2008). Pós-doutorado pela Universidad Carlos III de Madrid (Espanha, 2016). Desenvolve pesquisas na área de Comunicação Científica, com ênfase na produção de indicadores científicos, bibliometria, colaboração científica, análise de citação, análise de co-citação, rankings universitários, Ciência Aberta, compartilhamento e reuso de dados. Tem experiência acadêmica e profissional na área de Planejamento, gestão e arquitetura de bibliotecas.

https://orcid.org/0000-0003-0549-4567

https://lattes.cnpq.br/5243732207004083

samile.vanz@ufrgs.br

TATYANE GUEDES MARTINS DA SILVA

Bacharel em Biblioteconomia pela Universidade de Brasília (2019). Tem experiência na área de Ciência da Informação, com ênfase em Biblioteconomia. Foi bolsista do Programa de Capacitação Institucional do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) desenvolvendo serviços na área de Ciência Aberta e repositórios de dados. Possui como principais temas de interesse: softwares livres para bibliotecas, Ciência Aberta, dados abertos, gestão de dados de pesquisa, repositórios, metadados e princípios FAIR.

https://orcid.org/0000-0002-1743-0467

https://lattes.cnpq.br/7310861285054095

tatyanesilva@Ibict.br

WASHINGTON LUÍS RIBEIRO DE CARVALHO SEGUNDO

É doutor em Informática pela Universidade de Brasília (UnB), com período sanduíche no Kings College London, e mestre na mesma área pela UnB. Possui também formação em Matemática (bacharelado e licenciatura) pela mesma instituição. Atualmente, é Coordenador-Geral de Informação Científica e Tecnológica no Instituto Brasileiro de Informação em Ciência e Técnica (Ibict), onde lidera projetos voltados à Ciência Aberta, repositórios digitais, interoperabilidade de sistemas e gestão de dados científicos. É Docente Permanente do Programa de Pós-graduação em Ciência da Informação do Ibict. Entre suas contribuições no Instituto, destaca-se a coordenação de iniciativas como o Oasisbr e a Biblioteca Digital Brasileira de Teses e Dissertações (BDTD). Lidera esforços relacionados à Rede dARK. Sua trajetória inclui o desenvolvimento do BrCris e o projeto Laguna.

https://orcid.org/0000-0003-3635-9384

https://lattes.cnpq.br/9453481318889500

washingtonsegundo@Ibict.br

Brasília 2026

Com a necessidade de dar transparência às pesquisas, torna-se necessário a disponibilização dos dados de pesquisa. Entretanto, em parte, isso traz problemas quanto aos dados sensíveis. Assim, essa obra vem atender a sanar dúvidas quanto à anonimização dos dados de pesquisa, de forma a atender as necessidades de oferecer transparência com segurança. Uma obra atual e necessária aos pesquisadores, em tempos de abertura das ciências

Milton Shintaku
Coordenador de Tecnologias para Informação
Instituto Brasileiro de Informação em Ciência e Tecnologia

ISBN: 978-85-7013-224-6
CDL
9 788570 132246

Editora Ibict