A privacidade e os planos de gerenciamento de dados de repositórios de dados científicos

Elizabete Cristina de Souza Aguiar Monteiro
Universidade Estadual Paulista Júlio Mesquita Filho – UNESP, E-mail: beteaguia@yahoo.com.br

Elaine Parra Affonso
Universidade Estadual Paulista Júlio Mesquita Filho – UNESP, E-mail: elainepff@gmail.com

Victor Ubiracy Borba
Universidade Estadual Paulista Júlio Mesquita Filho – UNESP, E-mail: borba.victor.borba@gmail.com

Ricardo César Gonçalves Sant'Ana
Universidade Estadual Paulista Júlio Mesquita Filho – UNESP, E-mail: ricardosantana@marilia.unesp.br

RESUMO

Repositórios de dados científicos são ambientes digitais implementados nas universidades para auxiliar pesquisadores no gerenciamento, disponibilização e acesso a dados científicos, contribuindo com a sua reutilização. Aspectos sobre privacidade de dados dos sujeitos referenciados nas pesquisas devem estar presentes no Plano de Gerenciamento de Dados (PGD), tanto de pesquisadores quanto nos disponibilizados pelos repositórios. O objetivo deste trabalho foi analisar repositórios de dados de universidades para identificar aspectos de privacidade. Para tanto, foi verificado se aspectos de privacidade são tratados nos PGDs, e, ainda, evidenciadas as medidas de privacidade propostas. A metodologia utilizada foi quantitativa e qualitativa com o método exploratório, analisando os PGDs das universidades. Os resultados demonstraram que a maioria das universidades com repositórios, tratam em seus PGDs de medidas para proporcionar privacidade, tais como: consentimento informado, aderência às normas da *Health Insurance Portability and Accountability Act* (HIPAA) e supressão de identificadores pessoais. Embora haja menção sobre a necessidade de proteger dados pessoais e evitar ameaças à privacidade dos sujeitos referenciados nas pesquisas, técnicas para anonimização nem sempre estão detalhadas nos PGDs, podendo deixar dúvidas sobre como realizar tais procedimentos, visto que, essas técnicas são fundamentais para preservar a identidade dos participantes das pesquisas e garantir aspectos éticos.

Palavras-chave: Repositório de dados. Privacidade de dados. Anonimização de dados.

1 INTRODUÇÃO

Os Repositórios de dados científicos são ambientes implementados nas universidades com infraestrutura para dar suporte aos pesquisadores no gerenciamento e na disponibilização de

dados científicos e, dessa forma, ampliar o acesso para que outros pesquisadores possam reutilizá-los (MONTEIRO, 2017).

O objetivo de repositórios de dados é amplo e tem como núcleo viabilizar que dados coletados sejam armazenados em um suporte que permita sua recuperação, ampliando seu potencial de compartilhamento. As atividades inerentes ao gerenciamento de dados podem ser documentadas no Plano de Gerenciamento de Dados (PGD).

O PGD é um documento ou conjuntos de instruções que orientam àqueles que estão envolvidos com a gestão de dados científicos e descreve como os dados serão tratados durante o projeto e o que acontece com os dados após o término (MICHENER, 2015). Tanto o pesquisador quanto o repositório podem dispor de PGD. O PGD do pesquisador, quando presente, deve ser elaborado já no início de sua pesquisa prevendo como seus dados serão gerenciados. Já os repositórios de dados deveriam disponibilizar PGDs para orientar os pesquisadores que vão depositar seus dados, os usuários que irão acessar estes dados e, ainda, os profissionais responsáveis por sua operacionalização.

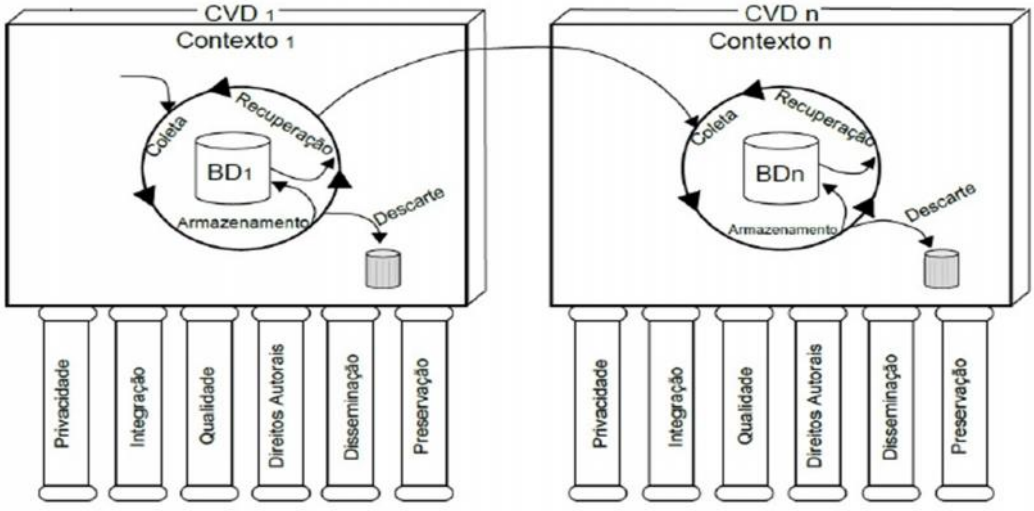
Questões de privacidade são fatores preponderantes a serem tratados no PGD, pois, cada vez mais é exigido pelas instituições de pesquisa e agências de fomento que o próprio pesquisador formalize no PGD seu compromisso sobre questões éticas e de privacidade, como os procedimentos para garantir proteção dos dados pessoais, principalmente em relação ao compartilhamento de dados sensíveis. Sayão e Sales (2016, p. 70) ao falarem sobre curadoria digital e dados de pesquisa, ressaltam que "[...] existe uma preocupação forte com questões éticas, de privacidade e de propriedade intelectual [...]".

Portanto, o conjunto de dados coletados pelo pesquisador, poderá ter dados que contextualizam informações vinculadas a um indivíduo, como dados provenientes dos seus atos, consumo, manifestações e opiniões entre outros. Desta forma, esse conjunto de dados, quando compartilhados sem devidas precauções, pode ameaçar a privacidade dos sujeitos referenciados nesses conjuntos de dados.

Para contextualizar a coleta de dados, este trabalho utilizou o Ciclo de Vida dos Dados (CVD) (SANT'ANA, 2016), considerando o fator privacidade. O CVD é um modelo composto por quatro fases: Coleta, Armazenamento, Recuperação e Descarte, sobre as quais permeiam seis

fatores: Preservação, Disseminação, Direitos Autorais, Qualidade, Integração e Privacidade (Figura 1).

Figura 1 - Ciclo de Vida dos Dados para a Ciência da Informação (CVD-CI)



Fonte: Sant'Ana (2016).

A fase da coleta configura o processo de obtenção dos dados. Nessa fase têm-se as atividades "[...] vinculadas a definição inicial dos dados a serem utilizados, seja na elaboração do planejamento de como serão obtidos, filtrados e organizados, identificando-se a estrutura, formato e meios de descrição que será utilizado." (SANT'ANA, 2013, p. 18).

Nesse cenário, o objetivo deste trabalho foi analisar os repositórios de dados das universidades para identificar aspectos de privacidade na fase de coleta de dados dos repositórios. Para tanto, buscou-se especificamente: Verificar como os aspectos de privacidade têm sido abordados nos PGDs; e evidenciar as medidas de privacidade propostas no PGD de cada repositório identificado.

2 PROCEDIMENTOS METODOLÓGICOS

Utilizou-se a metodologia de natureza qualitativa de tipo documental com levantamento bibliográfico sobre as temáticas. A pesquisa documental consistiu em explorar os planos de gerenciamento de dados dos repositórios. Foi realizada coleta de dados para levantamento dos repositórios de dados das 100 melhores universidades do mundo por meio do ranking

webometrics.info, definindo o escopo com as 100 melhores ranqueadas. A identificação dos repositórios de dados nas universidades foi realizada nos meses de julho a setembro de 2016.

Em seguida foi realizada pesquisa exploratória para levantamento das páginas oficiais das universidades para localização dos repositórios de dados e seus respectivos planos de gerenciamento de dados. Não foram analisados repositórios com acesso restrito ou com link quebrado. O processo de recuperação dos dados foi realizado por meio de coleta dos PGDs dos repositórios de dados encontrados, verificando menção às questões de privacidade de dados.

Para a análise e tratamento dos dados foram condensadas as informações evidenciadas nos PGDs dos repositórios resultando em um quadro com a sistematização do que está explicitamente declarado nos PGDs e, em seguida, feitas as interpretações.

3 REPOSITÓRIO DE DADOS E AS QUESTÕES DE PRIVACIDADE

As instituições de ensino e pesquisa, sobretudo as universidades, viabilizam a implementação de repositórios de dados para fornecer serviços de gerenciamento de dados científicos e auxiliam pesquisadores no ciclo de vida de seus dados para serem publicados e replicados.

Os repositórios de dados científicos são particularmente relevantes no gerenciamento dos conjuntos de dados, pois proporciona a disponibilização e compartilhamento de dados e fornece orientações para a comunidade de pesquisadores, sobretudo relacionadas às questões desafiadoras do compartilhamento legal e ético dos dados. As restrições indicadas no compartilhamento de dados têm o intuito de proteger os participantes humanos e suas identidades e, assim, cumprir as leis e os requisitos sobre privacidade.

Rodrigues et al. (2010, p. 22-23, grifo nosso) contextualizam repositório de dados como uma extensão de repositórios

> [...] “repositório” designa um sistema informático em que existe uma plataforma de armazenamento de objectos representados em ficheiros, capaz de incorporar novos objectos à medida que são produzidos ou submetidos. O repositório oferece serviços que são dirigidos a quem deposita, a quem pesquisa e aos administradores do sistema. Nos **repositórios de dados** pode ir-se muito além desta visão de repositório de objectos, uma vez que cada conjunto de dados tem características próprias e por isso pode requerer um tratamento diferenciado.

As orientações sobre as práticas para viabilizar a reutilização de dados que estão publicados nos repositórios envolve, inclusive, o contexto da coleta e os participantes humanos da pesquisa. O consentimento livre que o pesquisador empregou para a sua coleta com seres humanos deve prever a utilização por outros pesquisadores que poderão replicar os dados. Todas as liberações e implicações que estiverem no consentimento dos sujeitos da pesquisa deverão estar especificados no PGD e aplicados nos dados para proteger a identidade privada dos sujeitos e garantir a sua privacidade.

Privacidade foi definida por Westin em 1967 na obra *Privacy and Freedom* sendo a "[...] reivindicação de indivíduos, grupos ou instituições para determinar, quando, como e em que extensão, informações sobre si próprios devem ser comunicadas a outros" (WESTIN, 1967, p. 5).

Segundo Westin (1967, p. 7) o direito à privacidade não se trata de um direito absoluto, pois muitas vezes é baseado em outros direitos coletivos ou individuais destacando que

> O desejo do indivíduo por privacidade nunca é absoluto, uma vez que a participação em sociedade é igualmente importante. Assim, cada indivíduo está continuamente envolvido em um processo pessoal de equilíbrio entre o desejo de privacidade e o desejo de exposição e comunicação com os outros, à luz de condições do ambiente e de normas sociais na sociedade em que vive. O indivíduo o faz em face das pressões da curiosidade dos outros e dos processos de vigilância que toda sociedade necessita para a implementação de normas sociais.

De acordo com Doneda (2006, p. 23) informações pessoais ou dados pessoais podem ser entendido:

> Uma determinada informação pode possuir um vínculo objetivo com uma pessoa, revelando algo sobre ela. Este vinculo significa que a informação refere-se às características ou ações desta pessoa, que podem ser a ela atribuídas em conformidade com a lei, como no caso do nome civil ou do domicilio, ou então, às informações provenientes de seus atos, como os dados referentes ao seu consumo, informações provenientes de suas manifestações, como as opiniões que manifesta, tantas outras. E importante estabelecer este vínculo, pois ele afasta outras categorias de informações que, embora façam referência a uma pessoa, não seriam consideradas propriamente informações "pessoais no sentido pretendido: as opiniões alheias sobre esta pessoa, por exemplo, a princípio não possuem este vínculo objeto; também a produção intelectual de uma pessoa, em si considerada, não é per se informação pessoal (embora o fato de sua autoria o seja).

Nenhuma outra pessoa, independente dos motivos pelos quais os dados que envolvem seres humanos tenham sido coletados, tem o direito de publicar suas produções sob qualquer

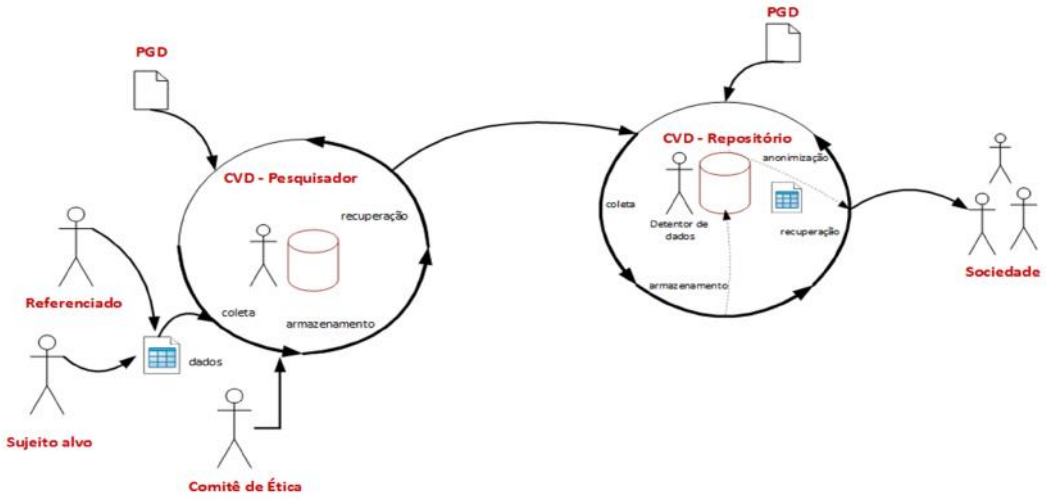
forma, sem o consentimento dos indivíduos que fizerem parte da coleta. Este direito é totalmente independente do material sobre o qual, ou os meios pelos quais, o pensamento, o sentimento ou a emoção são expressos (WARREN; BRANDEIS, 1890, p. 99, tradução nossa).

Tendo em vista a necessidade de proteger dados pessoais quando esses são resultados de pesquisas científicas, torna-se relevante descrever questões e atores envolvidos na fase de coleta de dados, considerando tanto o ciclo de vida dos dados do pesquisador, quanto o ciclo de vida de dados do repositório, incluindo estratégias e verificação dos tipos de dados envolvidos no processo de proteção de dados pessoais.

No contexto da coleta de dados científicos, participam os atores: sujeito alvo (participante da pesquisa), referenciados (aqueles que de alguma forma tem seus dados refletidos nos dados coletados), pesquisador, detentor de dados (profissional responsável pelos dados no repositório), comitê de ética, e sociedade (inclusive pesquisadores que farão coleta nos repositórios).

A fase de coleta acontece tanto no momento da coleta do pesquisador (CVD Pesquisador) quando coleta seus dados para sua pesquisa, quanto no momento do depósito desses dados no repositório (CVD - Repositório). Quando o pesquisador deposita os dados nos repositórios para serem disponibilizados à sociedade, deve-se atentar ao aspecto sobre a privacidade dos dados em que os mesmo podem ser anonimizados para que não ocorra ameaça à privacidade dos sujeitos referenciados no conjunto de dados (Figura 2).

Figura 2 - Processo de Coleta de dados - Pesquisador e Repositório



Fonte: Dados da pesquisa (2017).

Ao coletar e disponibilizar dados nos repositórios é preciso identificar os tipos de dados que compõem o conjunto de dados, assim, em relação aos tipos de dados envolvidos nas questões de proteção da privacidade, esses podem ser classificados em: identificadores, semi-identificadores, atributos sensíveis, e atributos não sensíveis (CIRIANI et al., 2009; DE CAPITANI DI VIMERCATI et al., 2012).

Dados denominados identificadores caracterizam-se por identificar unicamente os indivíduos no conjunto de dados (ex.: CPF, nome, número de identidade, número de Matrícula) (CIRIANI et al., 2009), e são os primeiros as serem evidenciados e protegidos quando a finalidade é garantir a privacidade dos indivíduos referenciados nos conjuntos de dados que serão disponibilizados (SAMARATI; SWEENEY, 1998).

Atributos semi-identificadores são aqueles que caracterizam-se por conterem valores que, quando correlacionados e/ou combinados com dados externos, podem proporcionar a identificação do indivíduo e, desta forma, vincular o indivíduo a seus dados confidenciais. Podem ser considerados dados semi-identificadores: data de nascimento, CEP, cargo, função, dados de localização, sexo, entre outros (CIRIANI et al., 2009).

Para Sweeney (2002) a divulgação de atributos semi-identificadores deve ser realizada de forma cautelosa, pois, por meio deles, é possível a identificação do sujeito no conjunto de dados mediante a combinação de dados.

Os atributos sensíveis são aqueles que representam os dados confidenciais (ex.: doenças, salário, exames médicos, origem racial, opiniões políticas, lançamentos de cartão de crédito), pois quando expostos podem colocar o indivíduo em situações constrangedoras. Os dados que não implicam em danos para o indivíduo quando revelados, não violando o direito à privacidade, são denominados de atributos não sensíveis (DE CAPITANI DI VIMERCATI et al., 2012).

Durante a investigação científica, o pesquisador coleta dados que podem abranger tanto dados identificadores (nome, CPF), semi-identificadores (data de nascimento, endereço, CEP), quanto dados sensíveis (doenças, religião, salário).

Desta forma, no processo da descrição dos procedimentos e diretrizes da gestão dos dados no PGD, o pesquisador pode, se oportuno, solicitar consentimento dos participantes para compartilhamento e uso a longo prazo de dados confidenciais. Logo, é adequado definir e descrever no PGD qual nível de confidencialidade será mantido (MONTEIRO, 2017). Esse consentimento também ajudará o gestor do repositório na disponibilização dos dados.

Além do consentimento do usuário, estratégias como anonimização de dados e criptografia devem ser adotadas pelos profissionais que detém os dados, a fim de possibilitar o compartilhamento de dados e garantir privacidade para indivíduos referenciados em conjunto de dados de pesquisas.

Anonimização é definida no Art. 5, inciso XII do Projeto de Lei 5.276/2016 como "qualquer procedimento por meio do qual um dado perde a possibilidade de associação direta ou indireta, a um indivíduo", desta forma, evita as possíveis correlações de dados semi-identificadores com dados que podem identificar o indivíduo. Ainda, para Skopek (2014) tornar os dados anônimos é esconder a identidade de indivíduos presentes em conjunto de dados, de modo que, os fatos sobre ele podem se tornar conhecidos e disponibilizados sem causar ameaças a privacidade.

Affonso, Oliveira e Sant'Ana (2017) ressaltam que, por meio de técnicas de anonimização, tais como, supressão, generalização, randomização (adição de ruído) ou troca de dados (*swapping*), é possível obter um conjunto de dados anonimizados, que quando disponibilizados, permite acesso aos dados do sujeito, mantendo protegida a sua identidade e minimizando ameaças a privacidade.

A generalização é um meio para tornar atributos de um conjunto de dados com valores menos específicos, conseguindo manter a representação semântica dos dados, sem perder a utilidade (SWEENEY, 2002), por exemplo, o atributo data de nascimento 03/02/2017 ao ser generalizado passa ser representado pelo valor 02/2017, ou apenas o ano 2017. Desta forma, o conjunto de dados que será disponibilizado passa ter mais registros com valores semelhantes, minimizando a identificação do sujeito nesses dados.

Supressão tem a finalidade de remover o valor do atributo que pode ocasionar quebras de privacidade, principalmente dados identificadores (SWENNEY, 2002). As técnicas de generalização e supressão são consideradas técnicas não-perturbativas, devido não alterarem o valor do atributo (CIRIANI, 2007).

Adição de ruído modifica o valor do atributo por meio de um valor aleatório, gerado normalmente pela distribuição Gaussiana, assim, os dados são disponibilizados perturbados aleatoriamente (KARGUPTA et al, 2003), essa técnica ao contrário da generalização ela muda o valor do atributo, desta forma a anonimização ocorre por meio de perturbação (CIRIANI, 2007).

A técnica de *swapping* ou *data swapping* tem a finalidade de modificar os registros de uma tabela de dados trocando os valores de atributos sensíveis, proporcionando incerteza sobre a relação do dado com o indivíduo (CIRIANI, 2007).

Ressalta-se que no processo de coleta de dados, devem-se levar em consideração as mesmas estratégias de anonimização de dados para futura disponibilização, tanto em relação ao pesquisador, quanto em relação ao repositório. Sendo assim, o detentor de dados do repositório deve garantir que os dados que serão disponibilizados estejam sob medidas de privacidade, e nos PGDs devem estar explícitas tais medidas.

4 RESULTADOS E DISCUSSÕES

A análise incluiu a identificação dos repositórios de dados das universidades e a identificação dos PGDs, baseando-se na fase de Coleta com o fator Privacidade do CVD do repositório (FIGURA 2).

As análises demonstraram que: 55 universidades dispõem de repositórios de dados. Dessas, 36 têm PGD. Das universidades que possuem PGD, 30 mencionam aspectos de privacidade nos seus PGDs.

O Quadro 1 identifica na primeira coluna os repositórios das Universidades que possuem PGDs, posteriormente, a abordagem os aspectos de privacidade nos PGDs dos repositórios analisados, e as medidas explícitas para garantir a proteção de dados dos participantes de pesquisas científicos.

Quadro 1 - Menção e medidas para proteção de dados pessoais em repositórios de dados

Universidade	**Abordagem**	**Medidas em relação à privacidade**
Harvard	As informações podem ser divulgadas sem restrições de privacidade se não tiver assuntos pessoais	Consentimento informado; dados anonimizados utilizando protocolo IRB (DATAVERSE PROJECT, c2015).
Massachusetts Institute of Technology	Dados de indivíduos em pesquisas, deve ser verificado questões de privacidade e manter a confidencialidade.	Consulta ao MIT *Committee on the Use of Humans as Experimental Subjects* (COUHES); Consentimento informado; dados de paciente devem estar em aderência com a HIPAA (*Health Insurance Portability and Accountability Act)*; modificação de dados sensíveis; exclusão de identificadores; generalização para semi-identificadores (MASSACHUSETTS INSTITUTE OF

		TECHNOLOGY, [201-]).
Stanford University	Caso a pesquisa envolva assuntos humanos deve considerar problemas de privacidade antes de compartilhar dados	É recomendado modificação dos dados antes de compartilhar. Indicação do regulamento da HIPAA (STANFOR LIBRARIES, [c201-]).
University of California Berkeley	Dados sensíveis e dados humanos	Recomenda utilizar os Padrões de *Berkeley Data Classification Standards (*UNIVERSITY OF CALIFORNIA BERKELEY, c2017)
University of Michigan	Garantia de que o depósito de dados não viola os direitos de qualquer pessoa incluindo a privacidade	Utiliza o University of Michigan's (U-M) Institutional Review Board (IRB) declaração revisada sobre privacidade e confidencialidade (UNIVERSITY OF MICHIGAN, c2017).
University of Pennsylvania	Dados sensíveis	Cita senha e criptografia para os arquivos com dados sensíveis e orientações sobre segurança de dados (UNIVERSITY OF PENNSYLVANIA, c2017).
University of Oxford	Anonimização, consentimento explícito	*Checklist* sobre questões de privacidade: anonimização; uso de pseudônimos (UNIVERSITY OF OXFORD, c2013-2016).
Michigan State University	Mantém os dados privados	Não localizado medidas para proteção de dados (MICHIGAN STATE UNIVERSITY, c1992-2014).
Yale University	Dados que potencialmente violar a confidencialidade dos participantes deve ser imediatamente informada ao ISPS	Não localizado medidas para proteção de dados (YALE UNIVERSITY, c2017).
University of Cambridge	Requisitos de ética e segurança, destruição de dados de participantes humanos; considerar aspectos éticos antes do início da pesquisa	Modelo para consentimento informado; e cita dados anonimizados (UNIVERSITY OF CAMBRIDGE, c2017).
University of Wisconsin Madison	Restringir o acesso a dados sensíveis; segurança dos dados contra ataques de *malware*; atender os requisitos e políticas da universidade para manter privado dados sensíveis; considera a questão de re-identificação dos dados após anonimização	Políticas de dados: *UW-Madison data policies*; FERPA (*Family Education Rights and Privacy Act*) e HIPAA; recomendações para garantir segurança dos dados (UNIVERSITY OF WISCONSIN MADISON, c2011-2017).
University of Texas Austin	Manutenção do anonimato e considerações legais	Requer que os contribuidores removam, substituam ou editem informações confidenciais ou sensíveis de conjuntos de dados antes do upload; solicita que os usuários informem sobre dados que requeiram FERPA, HIPAA ou outros padrões federais de privacidade (TEXAS DATA REPOSITORY, [201-]).

University of California San Diego	Manter a privacidade e segurança dos dados	Orientações sobre o HIPAA (UNIVERSITY OF CALIFORNIA SAN DIEGO, [c201-]).
Pennsylvania State University	Não é permitido Informações pessoais identificáveis ou cobertos pela HIPAA e FERPA	As pesquisas envolvendo o uso de sujeitos humanos devem ser aprovadas pelo Conselho de Revisão Institucional de Penn State (IRB), e o consentimento apropriado deve ser obtido dos participantes para compartilhar suas informações (PENNSYLVANIA STATE UNIVERSITY, c2018).
University of Illinois Urbana Champaign	Os conjuntos de dados devem conter apenas conteúdos irrestritos com **NENHUMA informação** privada, confidencial ou outra legalmente protegida.	Não localizado medidas para proteção de dados (UNIVERSITY OF ILLINOIS URBANA CHAMPAIGN, c2016).
Princeton University	A distribuição da sua apresentação não violará os direitos de privacidade pessoais de qualquer grupo ou indivíduo	Não localizado medidas para proteção de dados
University College London	Dados que contenham informações pessoais ou confidenciais sobre indivíduos, organizações ou negócios	Pesquisa com participantes humanos vivos requer aprovação ética; Pesquisadores devem cumprir padrões éticos e legislação (*Data Protection* Act 1998) e empregar estratégias de consentimento informado, anonimização e controle de acesso a dados; General Data *Protection Regulation* (GDPR); lei de proteção de dados; lei de liberdade de informação; lei de direitos humanos; lei de capacidade mental e lei de serviço de estatísticas e registro); indica a anonimização dos dados e a consulta ao *Handling sensitive and personal information* (UNIVERSITY COLLEGE LONDON, c2018a; c2018b)
University of Virginia	Dados sensíveis e confidenciais	Exige que o pesquisador tenha removido qualquer informação confidencial ou sensível, registros educacionais de alunos protegidos pela FERPA, informações que identificam pessoalmente qualquer pessoa ou informações classificadas como altamente sensíveis leis estaduais ou federais ou a política da universidade (UNIVERSITY OF VIRGINIA, [201-]).
Purdue University	Não localizado evidências sobre privacidade de dados	Não localizado medidas para proteção de dados (PURDUE UNIVERSITY, [201-].
Univerisity of Arizona	Dados sensíveis	Consentimento informado de pesquisa; anonimização de dados; pesquisadores devem identificar dados identificadores e semi-identificadores (combinação de dados); obter uma revisão de confidencialidade (UNIVERISITY OF ARIZONA)

University of Edinburgh	No item permissões e direitos: questiona se o pesquisador anonimizou os dados e obteve consentimento explícito para compartilhamento de dados	Consentimento informado e anonimização (UNIVERSITY OF AMESTERDAM)
Washington University Saint Louis	Aborda anonimização de dados e consentimento informado	Pesquisador deve garantir a privacidade por meio de: anonimização; consentimento informado (WASHINGTON UNIVERSITY SAINT LOUIS)
Simon Fraser University	Dados humanos	Menciona anonimização e encoraja discutir o assunto com o Escritório de ética de pesquisa da universidade (SIMON FRASE UNIVERSITY, [c201-]
Virginia Polytechnic Institute and State University	Não depositar dados sensíveis ou confidenciais	Não depositar dados sensíveis ou confidenciais que violem a Lei, a ética ou a política institucional (VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY)
Tuffs University	Dados sensíveis ou de identificação pessoal devem ser criptografados	Criptografia (TUFFS UNIVERSITY, 2017)
Ruprecht Karls University Heidelberg	Dados pessoais devem ser protegidos de acordo com a política de privacidade	Não localizado medidas para proteção de dados (RUPRECHT KARLS UNIVERSITY HEIDELBERG)
University of Amesterdam	Pesquisador deve garantir privacidade e confidencialidade dos entrevistados	Código de conduta para pesquisa aplicada na educação superior; código de conduta sobre uso de dados pessoais (UNIVERSITY OF AMESTERDAM)
University of California Los Angeles UCLA	Dados sensíveis ou confidenciais	Checklist para verificar questões de privacidade, tais como: se a informação contém dados sensíveis, consentimento informado e anonimização de dados (UNIVERSITY OF CALIFORNIA LOS ANGELES UCLA)
University of Kentucky	Dados de saúde (dados sensíveis) devem ser anonimizados, como perturbação de identificadores pessoais; generalização de datas e randomização de dados; retenção de dados que foram marcados como protegidos	Anonimização, perturbação de dados (randomização) (UNIVERSITY OF KENTUCKY)
Universiteit Utrecht	Todos os dados da pesquisa devem ser gerenciados de acordo com os requisitos da universidade, gerenciando a segurança da informação, proteção de privacidade e transparência	Informa que a política foi baseada Requisitos de protocolo de pesquisa de opinião ética médica de acordo com a pesquisa médica holandesa (Assuntos Humanos) (WMO), conforme avaliado pelo Comitê de Ética em Pesquisa Médica (METC) (UTRECHT UNIVERSITY, c2017)

Fonte: Dados da pesquisa

A partir do Quadro 1, observa-se que as principais medidas para a proteção da privacidade explícita nos PGDs são:

- Consentimento informado (Harvard, Massachusetts Institute of Technology, University of Cambridge, University College London, University of Arizona, University of Edinburgh, Washington University Saint Louis);
- Alinhamento da coleta de dados realizada pelo pesquisador de acordo com a política de dados da HIPAA (Massachusetts Institute of Technology, Stanford University, University of Wisconsin Madison, University of California San Diego, Pennylvania State University, University of Texas Austin);
- Aderência a FERPA; (University of Wisconsin Madison, University of Texas Austin, University of Virginia) ;
- Supressão de dados identificadores e dados sensíveis; (Massachusetts Institute Technology, University of Virginia);
- Higienização dos dados (Michigan State University, Purdue University);
- Generalização de dados semi-identificadores com a finalidade de minimizar a correlação de dados, pois, por meio dessa técnica, é possível tornar os dados menos específicos, aumentando a quantidade de dados similares no conjunto de dados (Massachusetts Institute of Technology, University of Kentucky);
- Uso de técnicas de perturbação para esconder/mascarar dados sensíveis (University of Kentucky);
- Criptografia de dados (Tuffs University; University of Pennsylvania);
- Uso de *checklist* para o pesquisador verificar se realizou anonimização de dados e consentimento informado (University of Oxford);
- Instrução para que o pesquisador não deposite dados sensíveis no repositório (University of Virginia, Purdue University, Virginia Polytechnic Institute and State University);
- Anonimização e anonimização de dados seguindo o protocolo do *Institutional Review Board* (IRB)(Harvard, Simon Fraser University);
- Indicação para anonimização dos dados (University of Oxford, University of Cambridge, University of Wisconsin Madison, University College London, University of Arizona, University of Edinburgh, Washington University Saint Louis, University of Kentucky)
- Disponibilização de termos de uso e código de conduta sobre uso de dados pessoais e segurança de dados (University of Amesterdam);

- Instruções para que pesquisadores identifiquem dados identificadores, semi-identificadores e sensíveis (University of Arizona);
- Armazenamento separado para dados sensíveis (University of California Berkeley);
- Seguir os requisitos do protocolo de pesquisa de opinião ética médica de acordo com a pesquisa médica holandesa (Assuntos Humanos) (WMO) (University Utrecht);
- Randomização de dados como medida para proteger dados confidenciais, muito utilizado para permitir a disponibilização de dados, sem que haja quebras de privacidade no sujeito referenciado nos dados (*University of Kentucky);*
- Substituição de dados sensíveis por códigos (University of California Berkeley).

Embora 30 dos repositórios analisados apresentam em seus PGDs menções às questões de privacidade, observa-se que essas, muitas vezes, não são detalhadas, e não estão explícitas como devem ser realizadas as medidas de proteção da privacidade ou como o pesquisador deverá proceder para realizar anonimização dos dados antes de depositar no repositório.

Verifica-se que as orientações sobre a privacidade explícitas nos PGDs pelos repositórios estão relacionadas ao: "Consentimento Informado", presente em sete repositórios; "Alinhamento da coleta com a HIPAA", cinco repositórios e; "Anonimização de dados", em oito repositórios.

Os repositórios que têm mais orientações sobre a privacidade de dados são: Massachusetts Institute of Technology, University of Texas Austin, University College London, University of Arizona.

Ressalta-se ainda que, cinco repositórios apenas citam a necessidade de proteção de dados pessoais, no entanto, não apresentam políticas ou medidas para proteção de dados pessoais.

Os aspectos sobre manter a privacidade de dados dos sujeitos que fizeram parte da pesquisa no momento da publicação e divulgação dos resultados das pesquisas estão presente na literatura apresentada. Os repositórios de dados são ambientes que armazenam e disponibilizam diversos tipos de dados coletados pelas comunidades científicas e também devem levar em consideração a privacidade dos sujeitos das pesquisas quando disponibilizam os dados para o público e devem orientar a comunidade que vai arquivar os dados sobre as métodos disponíveis para aplicar nos dados para manter a privacidade.

Com a disponibilização de dados científicos em repositórios, a discussão e apresentação de métodos para preservar a privacidade de sujeitos de pesquisa deve estar evidente pois esses fatores previnem problemas éticos e jurídicos tanto para o pesquisador quanto para o repositório.

5 CONSIDERAÇÕES FINAIS

A preocupação dos repositórios de dados científicos em relação à questão da privacidade, principalmente em dados que envolvem humanos, é evidente na maioria deles. As diferentes medidas indicadas em cada repositório são evidenciadas nos PGDs como uma forma de assessorar os pesquisadores na liberação de seus conjuntos de dados envolvendo dados sensíveis, ponderando os diversos aspectos relacionados a manter a privacidade dos envolvidos na pesquisa e assegurando questões éticas.

Os profissionais que atuam nos repositórios de dados devem estar cientes dos vários aspectos descritos nos PGDs para garantir a privacidade dos dados arquivados, considerando as diretrizes elencadas.

Os pesquisadores devem distinguir as diferentes medidas e técnicas necessárias para proteger a privacidade dos indivíduos e deverão ter cautela no momento da disponibilização de dados sensíveis e dados que podem ser correlacionados com outras bases de dados, tal como, os dados semi-identificadores.

As técnicas utilizadas para anonimização dos dados e medidas para proteção de dados pessoais preservam a identidade dos indivíduos participantes da pesquisa, asseguram ao pesquisador os aspectos éticos e direcionam os profissionais dos repositórios na gestão dos dados que ficam disponíveis para sociedade.

Ainda que as questões de privacidade estejam mencionadas nos repositórios, observou-se que em muitos PGDs as medidas para proteger dados pessoais se apresentam vagas, sem muitos detalhes de como proceder para atingir a anonimização de dados antes de depositá-los no repositório, o que pode ocasionar problemas éticos e de exposição dos participantes das pesquisas. Esse cenário revela a importância de estudos dos fatores envolvidos no compartilhamento de dados de pesquisas e as medidas para proteção da privacidade na busca por padronização nos protocolos de ação no gerenciamento de Repositórios de Dados Científicos.

Privacy and the data management plans of scientific data repositories

ABSTRACT

Scientific data repositories are digital environments implemented in universities to help researchers in management, availability and access to scientific data, contributing to their reuse. Aspects about data privacy of the subject referenced in the research should be present in the Data Management Plan (PGD), both from researchers and available by repositories. The purpose of this work was to analyze university data repositories to identify aspects of privacy. For this, it was verified that privacy aspects are treated in PGDs, and the proposed privacy measures are highlighted. The methodology used was quantitative and qualitative with the exploratory method, analyzing the PGDs of the universities. The results shows that most universities with repositories, mention in their PGDs measures to provide data privacy, such as: informed consent, adherence to standards of Health Insurance HIPAA and Accountability Act (HIPAA) and suppression of personal identifiers. Although exists a mention about the necessity to protect personal data and avoid threats to the privacy of the person referenced in research, techniques for anonymization aren't always detailed in PGDs, and may leave doubts about how to do such procedures, since these techniques are fundamental to preserve the research participants identity and to assure ethical aspects.

Keywords: Data repository. Data privacy. Data anonymization.

REFERÊNCIAS

AFFONSO, E. P.; DE OLIVEIRA, S. C; SANT'ANA, R. C. G. Análise do equilíbrio entre privacidade e utilidade no acesso a dados. **Informação & Sociedade**, João Pessoa, v. 27, n. 1, 2017. Disponível em: <http://www.ies.ufpb.br/ojs/index.php/ies/article/view/29422>. Acesso em: 17 jun. de 2017.

BRASIL. **Projeto de Lei nº PL 5276/2016**. Disponível em: <http://www.camara.gov.br/proposicoesWeb/fichadetramitacao?idProposicao=2084378>. Acesso em: 03 dez. 2017.

CIRIANI, V. et al. Theory of privacy and anonymity. **Algorithms and theory of computation handbook**, 2009.

DATAVERSE PROJECT. **Data Management Plan**. c2015. Disponível em: <http://best-practices.dataverse.org/data-management/index.html>. Acesso em: 10 dez. 2017.

DE CAPITANI DI VIMERCATI, S. et al. Data privacy: definitions and techniques. **International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems**, v. 20, n. 06, p. 793-817, 2012. Disponível em: <https://www.semanticscholar.org/paper/Data-Privacy-Definitions-and-Techniques-Vimercati-Foresti/7c6abddbd791dddd281c5764dbe859c55ba2e019/pdf>. Acesso em: 10 de jun. de 2016.

DONEDA, D. **Da privacidade à proteção de dados pessoais**. Rio de Janeiro: Renovar, 2006.

KARGUPTA, H. et al. **On the privacy preserving properties of random data perturbation techniques**. IEEE, 2003. p. 99-106.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY. MIT Libraries. **Confidentiality.** Cambridge, [201-]. Disponível em: <https://libraries.mit.edu/data-management/share/confidentiality/>. Acesso em: 20 dez. 2017.

MICHENER, W. K. Ten simple rules for creating a good data management plan. **Plos Computational Biology**, v. 11, n. 10, Oct. 2015. Disponível em: <http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004525>. Acesso em: 20 dez. 2017.

MICHIGAN STATE UNIVERSITY. **RDP's License Agreement**. Michigan, c1992-2014. Disponível em: <http://rdp.cme.msu.edu/misc/disclaimer.jsp>. Acesso em: 18 dez. 2017.

MONTEIRO, E. C. S. A. **Direitos autorais nos repositórios de dados científicos:** análise sobre os planos de gerenciamento dos dados. 2017. 115 f. Dissertação (Mestrado em Ciência da Informação) – Faculdade de filosofia e Ciências, Universidade Estadual Paulista, Marília, 2017. Disponível em: <http://hdl.handle.net/11449/149748>. Acesso em: 30 abr. 2017.

PENNSYLVANIA STATE UNIVERSITY. **About ScholarSphere.** State College, 2018. Disponível em: <https://scholarsphere.psu.edu/about>. Acesso em: 08 jun. 2018.

PURDUE UNIVERSITY. **Research data management for Purdue**. West Lafayette, [201-]. Disponível em: <https://purr.purdue.edu/>. Acesso em: 18 dez. 2017.

RODRIGUES, E. et al. **Os repositórios de dados científicos:** estado da arte. 2010. Acesso em: 18 dez. 2017. Disponível em: <http://projeto.rcaap.pt/index.php?option=com_remository&Itemid=2&func=startdown&id=271&lang=pt>. Acesso em: 5 jun. 2016.

SAMARATI, P.; SWEENEY, L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. **Technical report, SRI International**, 1998. Disponível em: <https://epic.org/privacy/reidentification/Samarati_Sweeney_paper.pdf>. Acesso em: 20 maio 2017.

SANT'ANA, R. C. G. Ciclo de vida dos dados e o papel da Ciência da Informação. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO (ENANCIB), 14., Florianópolis. **Anais eletrônicos...** Rio de Janeiro: ANCIB, 2013.Disponível em: <http://enancib2013.ufsc.br/index.php/enancib2013/XIVenancib/paper/viewFile/284/319>. Acesso em: 14 jul. 2016.

SANT'ANA, R. C. G. Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. **Informação e informação**, Londrina, v. 21, n. 2, p. 116-142, maio/ago. 2016. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/27940/20124 >. Acesso em: 20 out. 2016.

SAYÃO, L. F.; SALES, L. F. Curadoria digital e dados de pesquisa. **AtoZ**: novas práticas em informação e conhecimento, Curitiba, v. 5, n. 2, p. 67-71, 2016.

SWEENEY, L. k-anonymity: A model for protecting privacy. **International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems**, v. 10, n. 05,p. 557-570, 2002. Disponível em: <http://www.worldscientific.com/doi/abs/10.1142/S0218488502001648>. Acesso em: 14 jun. 2017.

SIMON FRASE UNIVERSITY. **Data management planning**: research data management services. Burnaby, [c201-]. Disponível em: < https://www.lib.sfu.ca/help/publish/research-data-management/data-management-planning>. Acesso em: 10 dez. 2017.

STANFOR LIBRARIES. **Sharing sensitive data**. Stanfor, [c201-]. Disponível em: <https://library.stanford.edu/research/data-management-services/share-and-preserve-research-data/sharing-sensitive-data>. Acesso em: 10 dez. 2017.

TEXAS DATA REPOSITORY. **Policies.** Texas, [201-]. Disponível em: <http://data.tdl.org/policies/#terms-of-use>. Acesso em: 08 jun. 2018.

UNIVERSITY COLLEGE LONDON. **Sharing data**. London, c2018a. Disponível em: <http://www.ucl.ac.uk/library/research-support/research-data/best-practices/guides/sharing/#>. Acesso em: 08 jun. 2018.

UNIVERSITY COLLEGE LONDON. **Handling sensitive & personal information**. London, c2018b. Disponível em: <http://www.ucl.ac.uk/library/research-support/research-data/best-practices/guides/sensitive_information>. Acesso em: 08 jun. 2018.

UNIVERSITY OF CALIFORNIA SAN DIEGO. **HIPAA e segurança.** San Diego, [c201-]. Disponível em: <https://idash.ucsd.edu/hipaa-and-security>. Acesso em: 10 dez. 2017.

UNIVERSITY OF CALIFORNIA BERKELEY. **Keeping sensitive data safe.** Berkeley, c2017. Disponível em: <https://bconnected.berkeley.edu/privacy-security/keeping-sensitive-data-safe >. Acesso em: 10 dez. 2017.

UNIVERSITY OF CAMBRIDGE. **University of Cambridge policy on the ethics of research involving human participants and personal data.** Trinity, c2017. Disponível em: <https://www.research-integrity.admin.cam.ac.uk/files/policy_on_the_ethics_of_research_involving_human_participants_and_personal_data_oct_2016.pdf>. Acesso em: 10 dez. 2017

UNIVERSITY OF ILLINOIS URBANA CHAMPAIGN. **Illinois data bank policy framework and definitions.** Champaign, c2016. Disponível em: <https://databank.illinois.edu/policies>. Acesso em 10 dez. 2017.

UNIVERSITY OF MICHIGAN. **Prepare our data.** Ann Arbor, c2017. Disponível em: <https://deepblue.lib.umich.edu/data/prepare-your-data>. Acesso em: 18 dez. 2017.

UNIVERSITY OF OXFORD. **Ethical issues and data protection.** Oxford, c2013-2016. Disponível em: <http://researchdata.ox.ac.uk/home/managing-your-data-at-oxford/ethical-legal-commercial/>. Acesso em: 10 dez. 2017.

UNIVERSITY OF PENNSYLVANIA. Penn Libraries. **Data planning and management**. Philadelphia, c2017. Disponível em: <http://guides.library.upenn.edu/data-management>. Acesso em: 18 dez. 2017.

UNIVERSITY OF VIRGINIA. Library. **LibraData deposit cheklist**. Charlottesville, [201-]. Disponível em: <https://www.library.virginia.edu/libra/datasets/libra-data-deposit-checklist/>. Acesso em: 18 dez. 2017.

UNIVERSITY OF WISCONSIN MADISON. **Data security**. Madson, c2011-2017. Disponível em: <http://researchdata.wisc.edu/data-security/>. Acesso em: 10 dez. 2017.

UTRECHT UNIVERSITY. **Research data management**. Utrecht, c2017. Disponível em: <https://www.uu.nl/en/research/research-data-management/university-policy-framework>. Acesso em: 10 dez. 2017.

WARREN, S. D.; BRANDEIS, L. D. The right to privacy. **Harvard Law Review**, Cambridge, v. 4, n. 5, p. 193-220, 1890.

WESTIN, A. F. **Privacy and freedom**. New York: Atheneum, 1967.

YALE UNIVERSITY. **Terms of use.** New Haven, c2017. Disponível em: <https://isps.yale.edu/research/data/terms-of-use>. Acesso em: 10 dez. 2017.