REPOSITÓRIO DE DADOS CIENTÍFICOS: aspectos sobre privacidade de dados

Scientific Data Repositories: aspects about data Privacy

Elizabete Cristina de Souza de Aguiar Monteiro[1], Elaine Parra Affonso[2], Victor Ubiracy Borba[3]; Ricardo César Gonçalves Sant'Ana[4]

(1) UNESP, Av. Hygino Muzzi Filho,737, Mirante, Marília-SP, beteaguia@yahoo.com.br
(2) UNESP, Av. Hygino Muzzi Filho,737, Mirante, Marília-SP, elainepff@gmail.com
(3) UNESP, Av. Hygino Muzzi Filho,737, Mirante, Marília-SP, borba.victor.borba@gmail.com
(4) UNESP, Av. Hygino Muzzi Filho,737, Mirante, Marília-SP, 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 há menção sobre aspectos de privacidade nos PGDs, e evidenciada 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, mencionam em seus PGDs medidas para proporcionar privacidade de dados, 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.

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 subjects referenced in the research should be present in the Data Management Plan (PGD), both from researchers and available by from repositories. The purpose of this work was to analyze university data repositories to identify aspects of privacy. For this, it has been verified if there exists a mention of privacy aspects 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 need 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 ensure ethical aspects.

Keywords: Data repository; Data privacy; Data anonymization.

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).

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.

A ambiência de repositórios de dados viabiliza armazenar, representar, gerenciar, disseminar, disponibilizar, e preservar dados neles depositados. Reunir conjuntos de dados nesses repositórios propicia compartilhamento, acesso e reuso de dados entre pesquisadores. As atividades inerentes ao gerenciamento de dados são 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 (MONTEIRO, 2017). Tanto o pesquisador quanto o repositório dispõe de PGD. O pesquisador elabora seu PGD no início de sua pesquisa para gerenciar seus dados. Do mesmo modo, os repositórios de dados disponibilizam PGDs para orientar pesquisadores que vão depositar seus dados e profissionais atuantes no repositório.

Questões de privacidade são fatores preponderantes a serem registrados 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.

Assim, estratégias como anonimização de dados e técnicas de criptografia devem ser adotadas pelos profissionais que detém esses dados.

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 [...]".

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 as estratégias e verificação dos tipos de dados envolvidos no processo de proteção de dados pessoais.

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 da 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, 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.

Os atributos sensíveis são aqueles que representam os dados confidenciais (ex.: doenças, salário, exames médicos, lançamentos de cartão de crédito) que quando expostos podem colocar o indivíduo em situações constrangedoras e, quando não causam ameaças, são denominados atributos não sensíveis (DE CAPITANI DI VIMERCATI et al., 2012).

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

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, técnicas de anonimização são relevantes para que o conjunto de dados possa ser compartilhado sem que ocorram ameaças a privacidade dos participantes das pesquisas. Affonso, Oliveira e Sant'Ana (2017) ressaltam que, por meio de técnicas de anonimização, tais como, supressão, generalização, adição de ruídos ou troca de dados (swapping), é possível obter um conjunto de dados anonimizados, que quando disponibilizado, permite acesso aos dados do sujeito, mantendo protegida a sua identidade e minimizando ameaças a privacidade.

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. Desta forma, esse conjunto de dados, quando compartilhado 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 (Apêndice A).

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).

No contexto da coleta de dados científicos, participam os atores: sujeito alvo (participante da pesquisa), pesquisador, detentor de dados (profissional responsável pelos dados no repositório), comitê de ética, e sociedade (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, os dados precisam ser anonimizados para que não ocorra ameaça à privacidade dos sujeitos referenciados no conjunto de dados (Apêndice B).

Ressalta-se que no processo de coleta de dados (Apêndice B), 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.

2 Objetivos

O objetivo deste trabalho foi analisar os repositórios de dados das universidades para identificar aspectos de privacidade de dados na fase de coleta do repositório. Para tanto, buscou-se especificamente: Verificar se há menção sobre aspectos de privacidade nos PGDs; e evidenciar as medidas de privacidade propostas no PGD dos repositórios identificados.

3 Procedimentos Metodológicos

Utilizou-se a metodologia quantitativa e qualitativa. 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. 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.

4 Resultados

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.

As análises demonstraram que: 55 universidades dispõem de repositórios de dados. Dessas, 36 têm PGD. Das universidades que possuem PGD, 78% mencionam aspectos de privacidade nos seus PGDs.

Em relação aos aspectos de privacidade mencionados nos PGDs dos repositórios analisados, elencam-se as seguintes medidas para garantir a proteção de dados dos participantes de pesquisas científicas:

- Consentimento informado;
- Alinhamento da coleta de dados realizada pelo pesquisador de acordo com a política de dados da *Health Insurance Portability and Accountability Act* (HIPAA)[1];
- Aderência a *Family Education Rights and Privacy Act* (FERPA)[2];

[1] Regras para garantir a privacidade de dados pessoais de saúde e seu acesso por profissionais de saúde e outros.

[2] Leis que protegem os dados dos estudantes e seu acesso pelos pais, escolas e outros.

- Supressão de dados identificadores e dados sensíveis;
- 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;
- Uso de técnicas de perturbação para esconder/mascarar dados sensíveis;
- Criptografia de dados;
- Uso de *checklist* para o pesquisador verificar se realizou anonimização de dados e consentimento informado;
- Instrução para que o pesquisador não deposite dados sensíveis no repositório;
- Anonimização de dados seguindo o protocolo do *Institutional Review Board* (IRB)[3];
- Disponibilização de termos de uso e código de conduta sobre uso de dados pessoais e segurança de dados;
- Instruções para que pesquisadores identifiquem dados identificadores, semi-identificadores e sensíveis;
- Armazenamento separado para dados sensíveis;

Embora 78% 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 é realizada as medidas de proteção da privacidade, ou como o pesquisador deverá proceder para realizar anonimização dos dados. Ressalta-se ainda que, três repositórios apenas citam a necessidade de proteção de dados pessoais, no entanto, não apresentam políticas ou medidas para proteção de dados pessoais.

4 Considerações Finais

A precaução dos repositórios relacionada às questões da privacidade dos dados, principalmente 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

[3] IRB é um órgão administrativo estabelecido para proteger os direitos, o bem-estar e aspectos sobre privacidade dos sujeitos humanos participantes de pesquisas.

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.

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**, 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.

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

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.

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.

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

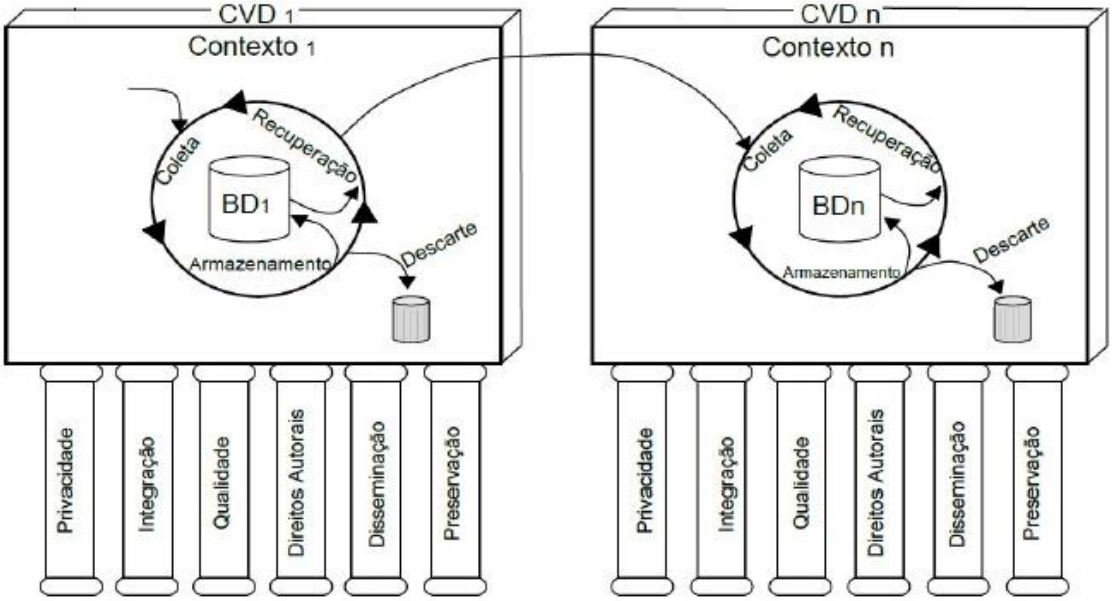
SAMARATI, P.; SWEENEY, L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. **Technical report, SRI International**, 1998. Disponível em: <https://epic.org/privacy/reidentification/Samarati_Sweeney_paper.pdf>. Acesso em: Maio de 2017.

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.

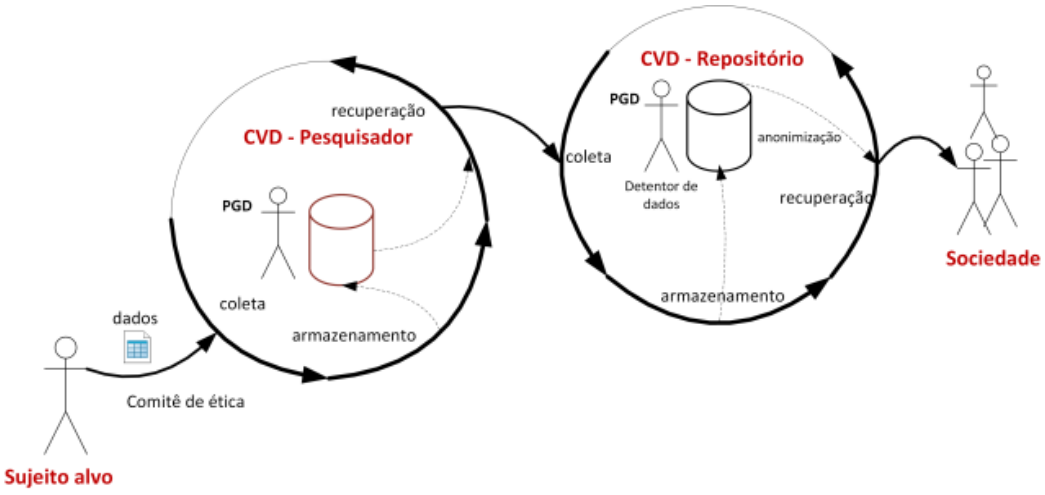
SWEENEY, L. k-anonymity: A model for protecting privacy. **International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems**, v. 10, n. 05,p. 557-570, 2002. Disponível em: <http://www.worldscientific.com/doi/abs/10.1142/S0218488502001648>. Acesso em: 14 jun. 2017.

Apêndice A – Ciclo de Vida dos Dados para a Ciência da Informação (CVD-CI)



Fonte: SANT'ANA, 2016

Apêndice B – Processo de Coleta de dados - Pesquisador e Repositório



Fonte: (Dados da Pesquisa, 2017).