Criação de Planos de Gestão de Dados em Projetos de Ciência de Dados para Detecção de *Fake News* Apoiados pelos princípios FAIR

Jorge Zavaleta[1], Annatércia Pinheiro[1], Renato Cerceau[2,3], Cabral Lima[1], Maria Luiza Machado Campos[1], Sérgio Manuel Serra da Cruz[1,4]

[1]Programa de Pós-graduação em Informática – Universidade Federal do Rio de Janeiro (UFRJ) Caixa Postal 68.530 – Rio de Janeiro – RJ – Brasil

[2]Departamento de Ensino e Pesquisa – Instituto Nacional de Cardiologia (INC) – Rio de Janeiro – RJ – Brasil

[3] Programa de Pós-Graduação em Telemedicina e Telessaúde – Universidade do Estado do Rio de Janeiro (UERJ) – Rio de Janeiro – RJ – Brasil

[4]Programa de Pós-graduação em Humanidades Digitais – Universidade Federal Rural do Rio de Janeiro (UFRRJ) – Seropédica – RJ – Brasil

{jorge.zavaleta,serra,mluiza}@ppgi.ufrj.br, annatercia@ufrj.br, cerceau@gmail.com,cabrallima@ufrj.br

Abstract. *Data Science researchers are experiencing an increasingly multifaceted reality concerning data governance. The paradigm shifts from disconnected data silos to data management plans (DMP) and standardized online repositories are adhering to the FAIR principles. This manuscript discusses, compares current DMP platforms, and describes the creation of a DMP in a Machine Learning project focused on Fake News detection. As a result, we describe a use case with the construction of the PGD on the DS-Wizard platform and offer an executable article that may be executed by the readers.*

Resumo. *Pesquisadores da área de Ciência de Dados vivenciam uma realidade cada vez mais multifacetada no que diz respeito à governança de dados. Mudança do paradigma de silos de dados desconectados para planos de gestão de dados (PGD) e repositórios padronizados online aderentes aos princípios FAIR ainda não é uma realidade. Este texto discute, compara plataformas e, descreve o percurso semi-automatizado de criação PGD, com aplicação em projetos de Machine Learning adotados na detecção de Fake News. Como resultados, oferecemos um trajeto para elaboração de PGDs na plataforma DS-Wizard e a oferta de um artigo do tipo executável sobre o projeto que pode ser executado pelos leitores.*

1. Introdução

Pesquisadores de diversas áreas do conhecimento vivenciam uma realidade cada vez mais multifacetada. Eles atuam concomitantemente como produtores de conhecimentos sob a forma de grandes *datasets* e consumidores de dados. Porém, muitos ainda subestimam a importância de gerenciar espaço-temporalmente seus próprios *datasets*. A

mudança do paradigma de silos de dados desconectados (em instituições que lidam direta ou indiretamente com geração de conhecimento) para uso de repositórios padronizados *online* aderentes aos Planos de Gestão de Dados (PGD) e ainda não são realidades [Koutkias, 2019; Wilkinson et al., 2017]. Por outro lado, paralelamente, tanto a indústria ligada à transformação digital, academia quanto os órgãos de fomento começam a estabelecer diretrizes mais rígidas relacionadas à governança de dados científicos, com vista a ampliar a competitividade, a transparência e a reprodutibilidade das pesquisas no sistema científico moderno [Henning, 2019; Wilkinson et al., 2017].

Frente a esta nova realidade, os pesquisadores, de forma particular ou em grupos, enfrentam desafios adicionais para organizar (sistematizar ou compartilhar) seus *datasets*. O enfrentamento do problema passa tanto pela evidente necessidade de reconhecimento da demanda (com sensibilização dos profissionais para sua execução) quanto pela necessidade de entendimento pelos pesquisadores de quais elementos são adequados para constarem nos seus planejamentos e, também, na formulação e compartilhamento de um conjunto de procedimentos para que terceiros posam lidar com seus *datasets* de pesquisa.

As atividades de governança de dados científicos têm ganhado maior importância no dia-a-dia dos pesquisadores, instituições e empresas; demandam esforços de alinhamento com as melhores práticas da Ciência Aberta [European Commission, 2016] e com as recomendações de diversos organismos internacionais tais como *Research Data Alliance* (*RDA*), iniciativa GO-FAIR (que visa implementar os princípios FAIR), *World Data System (*WDS) e *Commitee on Data (*CODATA) [Karimova et al., 2021; Pasquetto et al., 2017]. O aprimoramento da governança, acrescido de novas estratégias de motivação da comunidade científica, vêm sendo promovidos ativamente por organismos internacionais, editais e agências de fomento. Um dos elementos centrais é a recomendação da elaboração dos PGD [Karimova et al., 2021; Lefebvre et al., 2020].

Essencialmente, os PGD são documentos digitais formais que visam responder duas perguntas básicas: i) Quais dados serão gerados pelo projeto de pesquisa; ii) Como serão preservados e disponibilizados, considerando questões éticas, legais, de confidencialidade entre outras. Segundo Koers et al.[Koers et al., 2020], os PGD promovem não apenas o desenvolvimento de uma ampla gama de políticas, padrões, práticas e tecnologias de apoio para um melhor gerenciamento dos dados de pesquisa, mas também ampliam sua reprodutibilidade e o retorno do investimento das pesquisas.

Os PGD são fontes de apoio ao cientista e não um fim em si mesmo. Porém, sua importância e elaboração ainda são desconhecidas em grande parte da comunidade científica, em especial entre aqueles que desenvolvem pesquisa em Ciência de Dados (CD). Por exemplo, os PGD em projetos envolvendo experimentos de *Machine Learning* (ML) podem facilitar o acesso e reuso dos *datasets* por terceiros. Adicionalmente, podem tornar os experimentos mais localizáveis e reprodutíveis [Williams et al., 2017].

Henning (2019) destaca que para serem adequadamente tratadas, as plataformas geradoras de PGD e os *datasets*, idealmente, devem ser aderentes a certos requisitos, denominados de princípios FAIR (Localizável, Acessível. Intercambiável e Reutilizável). A autora afirma que: “Não basta ter um Plano de Gestão de Dados: é preciso ser FAIR [Wilkinson et al., 2017]”. Segundo esta abordagem, os PGDs e

datasets devem estar em um formato adequado que sejam mais facilmente localizáveis, acessíveis, interoperáveis e reutilizáveis tanto por máquinas quanto por usuários humanos.

A produção de PGD aderente aos princípios FAIR, em projetos de CD ainda é pouco frequente na literatura. Este trabalho tem como objetivo contribuir com a disseminação (na comunidade científica brasileira de CD e afins) das necessidades de adoção de plataformas digitais para construção de PGD e, dessa forma, contribuir com a governança de dados científicos, em especial as que se apoiam nas melhores práticas da Ciência Aberta.

Este artigo apresenta um estudo inicial que busca elucidar o processo de criação de PGDs e a geração de roteiro padronizado para o contexto de projetos de CD. Sua aplicação prática direcionada para ajudar uma ampla gama de pesquisadores ligados às áreas de Computação ou não*;* ilustra-se o processo de elaboração do PGD em caso de uso voltado para pesquisas de ML adotados na detecção de *Fake News* na área da Saúde trabalho também disponibiliza como um artigo do tipo executável [Lasser, 2020], disponível no Github, permitindo que os leitores possam avaliar a metodologia e verificar os resultados interativamente.

2. Plano de Gestão de Dados (PGD)

Tecnicamente, os PGD são modelos de conhecimento. Um PGD é materializado por um documento elaborado através de plataformas digitais em formato de texto de até duas páginas, onde consta uma descrição intencional dos dados e metadados produzidos por um projeto de pesquisa. Adicionalmente contém outros (meta)descritores complementares tais como: métodos de aquisição; restrições legais ou éticas; tecnologias demandadas para trabalhar com estes dados; políticas de preservação e compartilhamento; descrição de formatos e padrões; e os custos induzidos por recursos e serviços adicionais necessários para a governança de dados [Aguia, 2021; Lefebvre et al., 2020].

2.1 Motivação para preparação do PGD em projetos de Ciência de Dados

O PGD deve estruturar o percurso das atividades de pesquisa, integrando-se com outros sistemas e fluxos de trabalho, acompanhando todas as etapas, estratégias e resultados de um projeto [Karimova et al., 2021; Simms and Jones, 2017]. Ao ser produzido atendendo aos princípios FAIR, o documento auxiliará na legitimidade dos dados do projeto e em esclarecimentos junto aos órgãos de fomento, quando dos pedidos de subvenção financeira [Karimova et al., 2021].

Dentre as principais vantagens de elaborar PGD para projetos de CD destacamos: *i)* o planejamento prévio e organização antecipada de questões que irão influenciar a compreensibilidade e reuso dos *datasets*; *ii)* prevenção ou redução da probabilidade de contratempos, como perda ou uso inadequado dos *datasets*; *iii)* viabilizar a validação de resultados e a reprodutibilidade de resultados de pesquisa por terceiros, *iv)* favorecimento da atribuição de crédito aos seus autores que passam a ser devidamente citados e referenciados por pesquisadores que reutilizem os *datasets* no futuro; *v)* evitar a duplicação de esforços e investimentos, especialmente no tratamento de dados.

2.2 Organização dos PGD

Os PGD são concebidos a partir de meta-modelos (adotados pela instituição de pesquisa/fomento) e contém um conjunto de perguntas que devem ser respondidas pelo pesquisador responsável com um nível de detalhe adequado ao escopo do projeto [European Commission, 2016]. Por exemplo, o modelo empregado pela Comissão Europeia apresenta perguntas relacionadas com um conjunto de seções: 1) Resumo dos dados; 2) Aderência aos Princípios FAIR; 3) Alocação de recursos; 4) Segurança dos dados; 5) Aspectos éticos; e 6) outras informações relevantes ao projeto. Destacamos que na prática, não existe uma estrutura de organização pré-estabelecida e mandatória a ser utilizada nos modelos das plataformas digitais produtoras de PGD.

3. Plataformas de suporte para a criação dos PGD

Criar um PGD não é uma tarefa complexa, no entanto requer algum esforço adicional por parte do cientista além de conhecimentos sobre o ciclo de vida dos dados científicos e governança de dados em ferramentas de Ciência Aberta [Karimova et al., 2021; Sayogo and Pardo, 2013].

Sob essa perspectiva, nesta subseção, analisaremos as características das principais plataformas utilizadas na criação de PGD disponíveis na literatura, tais como DS-Wizard, DMPOnline, DMPTool, Argos e a ferramenta customizada da FIOCRUZ resumidas nas próximas seções, detalhadas, comparadas e aplicadas ao projeto em estudo encontram-se no repositório do mesmo. De modo geral elas coordenam de forma semi-automatizada o fluxo de trabalho na criação e gerenciamento dos PGD [Karimova et al., 2021].

3.1 DS Wizard, DMPOnline, DMPTool e Argos

Estas plataformas online são o resultado dos esforções por colaborar com pesquisadores e administradores de dados a criar PGDs inteligentes para projetos de Ciência Aberta alinhados com os princípios FAIR atendendo aos requisitos dos patrocinadores, possibilitando colaboração em tempo real, customização de questionários e listas de verificação para garantir consistência e compatibilidade entre planos de contextos diferentes, além de oferecer serviços para simplificar o gerenciamento, validação, monitoramento e manutenção dos PGDs auxiliando no gerenciamento de *datasets*, proveniência de dados e versionamento, ademais de apresentar de forma simplificada as informações do plano numa interface limpa [Argos, 2021; Dmptool, 2021; Ds-wizard, 2021; Pergl et al., 2019; Simms and Jones, 2017].

3.2 Ferramentas personalizadas

No cenário mundial existem poucas plataformas personalizadas geradoras de PGD. Em geral, as instituições e empresas utilizam-se das plataformas apresentadas no item anterior. No Brasil, a FIOCRUZ que vem adotando a perspectiva da Ciência Aberta em diversos projetos, incluindo a elaboração de uma política institucional de gestão, fomentando a abertura de dados para pesquisa, ali optou por elaborar um meta-modelo próprio para os seus PGD [Fundação Oswaldo Cruz, 2018] em razão de segurança e politicas institucionais sobre dados abertos.

A instituição elaborou uma plataforma personalizada para criar PGD alinhados aos princípios FAIR e acionável por máquina com base no documento *Practical Guide*

To The International Alignment Of Research Data Management [Fundação Oswaldo Cruz, 2018; Veiga et al., 2019]. A ferramenta será utilizada por programas de pós-graduação, pesquisadores e algumas instituições parceiras da área da saúde vinculadas à instituição.

4. Criação dos PGD para Projetos de Ciência de Dados

Nesta seção, discutimos os principais elementos presentes nos modelos de conhecimento presentes nas plataformas de suporte para a criação dos PGD. Adicionalmente, apresentamos os resultados de cinco experimentos de criação de PGD para um projeto de ML usado na detecção de *Fake News* na área da Saúde em Redes Sociais.

Devido a questões de espaço, os detalhes da metodologia, coleta de dados e algoritmos desenvolvidos estão descritos sob a forma de um artigo do tipo executável disponível no Github do projeto, nele podem ser encontrados maiores detalhes sobre os experimentos e discussões complementares sobre ele.

A preferência foi pelo DS-Wizard. Trata-se de uma plataforma abrangente e de fácil utilização, que se conecta com recursos externos, usa métricas preditivas automatizadas para avaliar as respostas do questionário em relação aos princípios FAIR e Ciência Aberta. Oferece documentos textuais e relatórios gráficos aos cientistas. Além disso, os PGD podem ser gerados em diversos formatos sendo legíveis por humanos e acionáveis por máquinas. Nas próximas subseções destacamos os principais elementos de um PGD para projetos de CD.

4.1 Elementos de Informações Administrativas

No PGD são mantidos metadados administrativos, tais como: Nome do projeto, autor(es), ORCIDs, instituição, e-mails, resumo do projeto, fontes de financiamento, datas de início e término e outras informações relevantes para uma melhor identificação do projeto.

4.2 Elementos de Resumo dos Dados

Há o registro dos dados coletados e produzidos pelo projeto. Um dado é um fato, um valor documentado ou um resultado de medição, a coleção de dados organizados e processados é chamado *dataset* que pode ser classificado em: estruturado e não-estruturado. Os dados podem ser adquiridos por processos manuais, mecânicos ou computacionais a partir de observações, processos laboratoriais, algoritmos, fenômenos físicos ou simulações computacionais.

Os *datasets* devem estar em formatos estandardizados com a finalidade de ser usados por outros pesquisadores, descrição dos métodos de obtenção dos dados, tecnologias de armazenamento, processamento e distribuição, tamanho do *dataset* e a forma em que serão usados os dados no projeto.

4.3 Elementos de Dados FAIR

Os princípios FAIR são considerados em todas as etapas no PGD na plataforma DS-Wizard. Essa condição permite que pesquisadores façam indicações precisas para cada um dos subprincípios FAIR. No caso deste projeto destacamos:

F1) Fornecer metadados descritivos minuciosos para garantir que os *datasets* possam ser mais facilmente encontrados por humanos e máquinas atribuindo um identificador global único e persistente.

F2) Registrar ou indexar os metadados em recursos de busca como repositórios que providenciem uso de cadernos eletrônicos (*Jupyter notebook, Jupterlab*) para garantir boa precedência da análise dos dados e documentação dos dados com padrões de metadados tais como *Dublin Core* [Dublin Core, 2021], *Data Documentation Initiative* (DDI) [DDI, 2021] e *Digital Object Identifier* (DOI) [DOI, 2021]. Entre os repositórios com estes serviços estão o *Zenodo* e o *ZivaHub*. O repositório deste projeto é https://doi.org/10.5281/zenodo.4697918.

F3) Padronizar as palavras chaves e nomes usados para capturar a proveniência retrospectiva usando o modelo W3C PROV e as bibliotecas Python PROV.

A1) Inserir metadados recuperáveis por identificador usando um protocolo de comunicação padronizado. Indicar as condições em que dados podem ser reutilizados ou não (dados de parceiros comerciais, confidenciais ou de população em risco) sob licença aberta (*Creative Commons Attribution-Only*) ou com restrições de usos específicos [DLS, 2021]. O repositório *DataFirst* tem uma série de condições de acesso para diferentes níveis de sensibilidade, enquanto o *ZivaHub* tem a opção de abrir os metadados enquanto mantém os arquivos de dados confidenciais [DataFirst, 2021].

A2) Usar protocolos abertos, gratuitos e universalmente implementáveis e que permitam procedimentos de autenticação e autorização [Wilkinson et al., 2016].

I1) Padronizar dados, utilizar vocabulários normalizados e formatos que permitam a descrição dos dados usando os princípios FAIR. Os metadados devem incluir referências qualificadas para outros metadados usando identificadores [DLS, 2021].

I2) Utilizar formato padronizado "*comma-separated values (csv)*" para facilitar a interoperabilidade entre sistemas.

R1) Inserir descritores semânticos de forma minuciosa com uma pluralidade de atributos específicos e relevantes, aumentado a capacidade de reutilização [DLS, 2021; Wilkinson et al., 2016], dados e metadados devem estar associados com sua proveniência e, serem liberados com licenças de uso claramente definidas [Veiga et al., 2019].

R2) Usar os subprincípios A1 e A2, todos os dados podem se tornar abertos imediatamente, verificando se o repositório escolhido tem licenças acionadas por máquina (está se tornando cada vez mais standard). Tanto o *ZivaHub* quanto o *DataFirst* fornecem licenças acionáveis por máquina [DLS, 2021].

5. Conclusões

As atividades de governança de dados científicos têm ganhado importância no dia-a-dia dos pesquisadores, instituições e empresas, mas a produção dos PGD ainda não estão consolidados junto à comunidade científica. Estima-se que grande parte dos *datasets* produzidos em projetos de CD possam se perder ou estar sob risco, em razão da fragilidade das nossas instituições caso não estimulem a institucionalização de

repositórios de dados e PGDs. Neste trabalho avaliou-se cinco plataformas geradoras de PGD, verificou-se que a DS-Wizard é a mais adequada para projetos de CD que utilizem experimento de ML voltados para a área de Saúde. A plataforma tem curva de aprendizado suave e pode ser utilizada por usuários inexperientes e produzir PGD compatíveis com os princípios FAIR. Neste trabalho também disponibilizamos o percurso metodológico, experimentos, *datasets* e os resultados em um artigo reprodutível. Como trabalhos futuros, implementar e disponibilizar as diferentes etapas do PGD no repositório.

Agradecimentos

Este estudo foi parcialmente financiado pela CAPES-TecnoDigital, CNPq - Bolsa DT - II (315399 / 2018-0), FAPERJ – Código E-26/210.192/2020.

6. Referências

Aguia. (2021). *Plano de Gestão de Dados*. Agência USP de Gestão Da Informação Acadêmica. https://www.aguia.usp.br/apoio-pesquisador/dados-pesquisa/plano-gestao-dados-2/

Argos. (2021). *Argos - Tool for Data Management Plan*. Argos. https://argos.openaire.eu/splash/

DataFirst. (2021). *Welcome DataFirst*. DataFirst. https://datafirst.uct.ac.za/

DDI. (2021). *Data Document Initiative*. Document, Discover and Interoperate. https://ddialliance.org/

DLS. (2021, April 5). *Stages of Research Data Management: Share & Publish*. Digital Library Services. http://www.digitalservices.lib.uct.ac.za/dls/services/rdm/share-publish

Dmptool. (2021). *dmptool: Build your Data Management Plan*. https://dmptool.org/about_us

DOI. (2021). *Digital Object Identifier System*. The DOI System. https://www.doi.org/index.html

Ds-wizard. (2021). *Data Stewardship Wizard*. DSW. https://ds-wizard.org/

Dublin Core. (2021). *Dublin Core*. Innovation in Metadata Design, Implementation & Best Practice: Dublin Core Metadata Initiative. https://dublincore.org/specifications/dublin-core/

European Commission. (2016). *H2020 Programme. Guidelines on FAIR Data Management in Horizon 2020*. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf?utm_content=bufferc22c5&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer

Fundação Oswaldo Cruz. (2018). *Grupo de Trabalho em Ciência Aberta. Termo de Referência: Gestão e Abertura de Dados para Pesquisa na Fiocruz*. https://www.arca.fiocruz.br/handle/icict/26803

Henning, P. C. (2019). Não basta um Plano de Gestão de Dados: é preciso ser FAIR. In Icict (Ed.), *Encontro da Rede Sudeste de Repositórios Institucionais* (Issue 1). Icict. https://www.arca.fiocruz.br/handle/icict/33372

Karimova, Y., Ribeiro, C., and David, G. (2021). Institutional Support for Data Management Plans: Five Case Studies. *Metadata and Semantic Research: 14th International Conference, MTSR 2020*, *1355*, 308–319. https://doi.org/10.1007/978-3-030-71903-6_29

Koers, H., Bangert, D., Hermans, E., van Horik, R., de Jong, M., and Mokrane, M. (2020). Recommendations for Services in a FAIR Data Ecosystem. *Patterns*, *1*(5), 100058. https://doi.org/10.1016/j.patter.2020.100058

Koutkias, V. (2019). From Data Silos to Standardized, Linked, and FAIR Data for Pharmacovigilance: Current Advances and Challenges with Observational Healthcare Data. *Drug Safety*, *42*(5), 583–586. https://doi.org/10.1007/s40264-018-00793-z

Lefebvre, A., Bakhtiari, B., and Spruit, M. (2020). Exploring research data management planning challenges in practice. *It - Information Technology*, *62*(1), 29–37. https://doi.org/10.1515/itit-2019-0029

Pasquetto, I. V., Randles, B. M., and Borgman, C. L. (2017). On the Reuse of Scientific Data. *Data Science Journal*, *16*(8), 1–9. https://doi.org/10.5334/dsj-2017-008

Pergl, R., Hooft, R., Suchánek, M., Knaisl, V., and Slifka, J. (2019). "Data Stewardship Wizard": A Tool Bringing Together Researchers, Data Stewards, and Data Experts around Data Management Planning. *Data Science Journal*, *18*(1). https://doi.org/10.5334/dsj-2019-059

Sayogo, D. S., and Pardo, T. A. (2013). Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. *Government Information Quarterly*, *30*(SUPPL. 1), S19–S31. https://doi.org/10.1016/j.giq.2012.06.011

Simms, S. R., and Jones, S. (2017). Next-Generation Data Management Plans: Global, Machine-Actionable, FAIR. *International Journal of Digital Curation*, *12*(1), 36–45. https://doi.org/10.2218/ijdc.v12i1.513

Veiga, V. S. de O., Henning, P., Dib, S., Penedo, E., Lima, J. D. C., Silva, L. O. B. da, and Pires, L. F. (2019). Plano de gestão de dados fair: uma proposta para a Fiocruz. *Liinc Em Revista*, *15*(2), 275–286. https://doi.org/10.18617/liinc.v15i2.5030

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Wilkinson, M. D., Verborgh, R., da Silva Santos, L. O. B., Clark, T., Swertz, M. A., Kelpin, F. D. L., Gray, A. J. G., Schultes, E. A., van Mulligen, E. M., Ciccarese, P., Kuzniar, A., Gavai, A., Thompson, M., Kaliyaperumal, R., Bolleman, J. T., and Dumontier, M. (2017). Interoperability and FAIRness through a novel combination of Web technologies. *PeerJ Computer Science*, *2017*(4). https://doi.org/10.7717/peerj-cs.110

Williams, M., Bagwell, J., and Nahm Zozus, M. (2017). Data management plans: the missing perspective. *Journal of Biomedical Informatics*, *71*, 130–142. https://doi.org/10.1016/j.jbi.2017.05.004