Revista Científica da Universidade Eduardo Mondlane, Série: Letras e Ciências Sociais ISSN 2307-3918

Pecha Kucha

REPOSITÓRIOS DE DADOS DE PESQUISA: critérios core *CoreTrustSeal*

Lyvia Rocha de Jesus Araujo[1,2], Eloísa Príncipe[1,2] e Maria Simone de Menezes Alencar[3]

[1]Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), Brasil
[2]Universidade Federal de Rio de Janeiro (UFRJ), Brasil
[3]Universidade Federal do Estado do Rio de Janeiro (UNIRIO), Brasil

RESUMO: A publicação de dados de pesquisa geralmente demanda que os conjuntos de dados tenham sido previamente compartilhados em um repositório e possuam um identificador persistente que garanta sua fácil recuperação e acesso. Em muitos casos, as revistas nas quais os dados serão publicados solicitam ou fornecem aos autores informações sobre os repositórios em que estes podem armazenar e compartilhar seus dados. Esta pesquisa tem o objetivo de avaliar as políticas dos repositórios de instituições brasileiras, indexados no diretório Re3data, de acordo com os critérios do certificado *CoreTrustSeal*, visando identificar se as informações publicadas por estes repositórios podem certificar a qualidade e a segurança dos mesmos. Conclui-se que, apesar de nenhum dos repositórios consultados possuir o selo de certificação *CoreTrustSeal*, em sua maioria, os repositórios têm seguido os critérios previstos nesta certificação, garantindo a segurança e a preservação dos dados. Sugere-se que cada vez mais a comunidade acadêmico-científica possa ter acesso a estas informações de modo a credibilizar os espaços em que seus dados são armazenados, ressaltando que a descrição mais esclarecida dos repositórios de dados pode aumentar a garantia da qualidade dos dados de pesquisa publicados em revistas híbridas ou especializadas na publicação de artigos de dados.

Palavras-chave: Critérios de confiabilidade, Repositórios de dados, Re3data, *CoreTrustSeal*.

RESEARCH DATA REPOSITORIES: CoreTrustSeal requirements

ABSTRACT: Publishing research data usually requires that the datasets have been previously shared in a repository and have a persistent identifier to ensure easy retrieval and access. In many cases, the journals in which the data will be published the request or provide authors with information about the repositories in which they can store and share their data. This research aims to evaluate the policies of Brazilian institutions' repositories, indexed in the Re3data directory, according to the CoreTrustSeal certificate requirements, in order to identify if the information published by these repositories can certify their quality and security. It is concluded that, although none of the repositories consulted has the CoreTrustSeal certification seal, most of them have followed the requirements of this certification, ensuring the security and preservation of the data. It is suggested that increasing the academic-scientific community can have access to this information in order to give credibility to spaces in which their data are stored, pointing out that a more enlightened description of data repositories can enhance the quality assurance of research data published in hybrid journals or journals specialized in the publication of data papers.

Keywords: Trustworthiness requirements, Data repositories, Re3data, CoreTrustSeal.

Correspondência para: (correspondence to:) lyviaaraujo@aluno.ibict.br

INTRODUÇÃO

A Ciência Aberta preconiza uma série de novas ações para seus atores, buscando práticas abertas, transparentes, colaborativas e inclusivas em todo processo do ciclo da pesquisa científica. Dentre essas dinâmicas, destacam-se o compartilhamento e o reuso de dados de pesquisa abertos, que se caracterizam como "[...] todo e qualquer tipo de registro coletado, observado, gerado ou usado pela pesquisa científica, tratado e aceito como necessário para validar os resultados da pesquisa pela comunidade científica." (SALES e SAYÃO, 2019, p. 36).

Os dados de pesquisa podem ser publicados como suplementos em revistas científicas tradicionais; em revistas de dados (*data journals*) dedicadas, exclusivamente, a publicação de dados de pesquisa; compartilhados através de troca pessoal; publicados em *sites* de pesquisadores ou laboratórios; ou depositados em repositórios temáticos ou institucionais (KIM, 2020).

A publicação de dados de pesquisa em periódicos científicos e sob a forma de artigos de dados demanda que os conjuntos de dados tenham sido previamente compartilhados em um repositório e possuam um identificador persistente que garanta sua fácil recuperação e acesso. Contudo, em estudo sobre as características de *data papers* (artigos de dados), Jihyun Kim (2020) aponta que, raramente, estes artigos e as revistas nas quais os *data papers* são publicados solicitam ou fornecem aos autores informações sobre os repositórios em que os dados podem ser compartilhados. Para o autor, o fornecimento transparente de informações sobre a reputação de repositórios, suas práticas de curadoria e plano de preservação podem contribuir com a avaliação da qualidade dos dados e da garantia de reúso dos dados publicados (KIM, 2020).

Em meio à preocupação de organizações científicas e instituições de fomento com a preservação dos dados, o *World Data System of the International Science Council* (WDS), e a *Research Data Alliance* (RDA) têm gerado mecanismos para certificar a qualidade dos repositórios e firmar o compromisso de preservação a longo prazo dos dados abertos compartilhados. Nesse cenário, em 2017 foi lançado o *CoreTrustSeal*[1], um certificado geral para comprovar a segurança e a confiabilidade de repositórios digitais.

Considerando o que fora apontado por Kim (2020) e a abrangência dos critérios da certificação *CoreTrustSeal* para repositórios digitais, esta pesquisa teve como objetivo dimensão de compatibilidade dos critérios definidos pela atualização 2020-2022 do certificado com a descrição e as políticas dos repositórios de dados brasileiros, indexados no diretório *Re3data*[2].

METODOLOGIA

A pesquisa caracteriza-se como descritiva e exploratória. O universo da pesquisa é composto pelos repositórios de dados de instituições brasileiras, indexados no diretório *Re3data*. A busca foi realizada selecionando os repositórios administrados por instituições brasileiras, já que na disposição dos filtros do *Re3data*, as informações de localização se referem às instituições que administram os repositórios (*Re3data Coref*, 2022)[3]. Os repositórios incluídos nesta pesquisa foram avaliados de acordo com os 16 critérios do *CoreTrustSeal* 2020-2022, que são divididos em três blocos (Tabela 1).

Para a análise dos dados foram criados dois quadros no aplicativo *Google Sheets* com as principais características dos repositórios, e os critérios de avaliação de confiabilidade definidos pelo *CoreTrustSeal* 2020-2022, sendo realizada uma síntese comparativa dos repositórios listados. Foram excluídos desta pesquisa: Programas ou Grupos de colaboração, que não se classificam exclusivamente como repositórios de dados (2); repositórios com *links* inválidos[4](1); e repositórios ainda em construção[5](1), sendo avaliados um total de nove repositórios.

TABELA 1: Critérios *CoreTrustSeal* 2020-2022

Infraestrutura Organizacional	
1) Missão e escopo	O repositório tem a missão explícita de fornecer acesso e preservar os dados no seu domínio?
2) Licenças	O repositório mantém todas as licenças aplicáveis que cobrem o acesso e utilização de dados e controla o seu cumprimento?
3) Continuidade do acesso	O repositório possui um plano de continuidade para assegurar o acesso contínuo e a preservação de sua disponibilidade?
4) Ética e confiabilidade	O repositório assegura, na medida do possível, que os dados são criados, tratados, compartilhados e utilizados em conformidade com as normas disciplinares e éticas?
5) Infraestrutura	O repositório dispõe de financiamento adequado e de pessoal qualificado administrado por um sistema claro de gestão para levar a cabo sua a missão de forma eficaz?
6) Orientação de especialistas	O repositório assegura regularmente o aconselhamento e o feedback de especialistas para garantir a sua contínua relevância e melhoria?
Gestão de objetos digitais	
7) Integridade e confidencialidade dos dados	O repositório fornece provas de que opera um sistema de gestão de dados e metadados adequado para assegurar sua integridade e autenticidade durante os processos?
8) Avaliações / Feedbacks	O repositório recebe retorno de seus usuários?
9) Procedimentos documentados de armazenamento	O repositório define como os registros dos dados podem ser comprovados, comprovando e avaliando sua proveniência?
10) Plano de preservação	O repositório possui um plano explícito de preservação que garanta o acesso aos dados compartilhados por um longo período?
11) Qualidade dos dados	São exigidas pelo repositório informações suficientes sobre os dados que garantam sua qualidade?
12) Fluxo de trabalho	São definidos e documentados fluxos de trabalho de acordo com as atividades do repositório?
13) Descoberta e identificação dos dados	Os conjuntos de dados compartilhados possuem metadados e identificadores persistentes que facilitem sua recuperação?
14) Reúso dos dados	O repositório toma medidas para assegurar que os dados sejam reusáveis?
Tecnologia	
15) Infraestrutura técnica	O repositório foi desenvolvido sobre infraestrutura estável que possa ser mantido a longo prazo e garanta a disponibilidade dos dados armazenados?
16) Segurança	O repositório prevê a análise de potenciais ameaças, a avaliação dos riscos e a criação de sistemas de proteção para os dados?

Fonte*: CORETRUSTSEAL*, 2019.

RESULTADOS

A pesquisa realizada em 29 de março de 2022 retornou um total de 13 repositórios que possuíam vínculo institucional com uma ou mais instituições brasileiras, dos quais foram avaliados nove repositórios. O processo de avaliação considerou a documentação disponibilizada nos *sites* dos repositórios, quantificando os resultados em cinco níveis de conformidade, adaptados do *CoreTrustSeal* (2019): 0) critério não citado na descrição/documentação do repositório; 1) critério citado genericamente na documentação, porém não é observado no repositório; 2) repositório aponta políticas e definições relacionadas, mas não declara o procedimento específico adotado; 3) repositório declara que irá implementar o critério; 4) repositório especifica e aplica o critério na descrição do *site*/documentação fornecida.

Os repositórios avaliados nesta pesquisa e os resultados de sua avaliação podem ser observados no Tabela 2.

Foi verificado que todos os nove repositórios consultados possuem escopo e missão bem definidos, estão sobre o financiamento estável de instituições de ensino e pesquisa confiáveis, detém sistema de gestão de dados que assegura a integridade dos mesmos, têm padrão de submissão aberta ou fechada de dados bem delimitado e registrado, exigem informações suficientes para comprovar a qualidade dos dados armazenados, estão desenvolvidos sobre uma infraestrutura estável que garante sua permanência online a longo prazo**,** e estão protegidos pelas políticas de segurança dos sistemas em que foram desenvolvidos (*Dspace, Dataverse* e *Ckan*).

TABELA 2: Resultados da avaliação dos repositórios

Repositórios	CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022 (Critérios)															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
PPBio Data Repository	4	2	2	4	4	4	4	0	4	0	4	0	4	2	4	4
FAPESP COVID-19 Data Sharing/BR	4	4	4	4	4	4	4	4	4	2	4	0	4	0	4	4
Base de Dados Científicos da UFPR	4	0	0	0	4	4	4	0	4	0	4	0	3	0	4	4
IBICT Cariniana Dataverse Network	4	0	2	0	4	4	4	4	4	0	4	0	4	0	4	4
SciELO Data	4	4	4	4	4	4	4	4	4	4	4	0	4	4	4	4
CEDAP Research Data Repository	4	0	0	0	4	0	4	0	4	0	4	0	4	2	4	4
Open Research Data @PUC-Rio	4	4	2	0	4	4	4	0	4	4	4	0	4	4	4	4
UNESP Institutional Repository	4	0	2	2	4	4	4	4	4	0	4	0	4	0	0	4
Dados Abertos De Pesquisas	4	0	0	0	4	0	4	0	4	0	4	0	4	0	4	4

Oito dos nove repositórios estão em conformidade com o critério 13, possuindo metadados ricos e disponibilizam identificador persistente para todos os conjuntos de dados publicados, com exceção da Base de Dados Científicos da Universidade Federal do Paraná, cujo fornecimento de *Digital Object Identifier* (DOI) está temporariamente suspenso. Seis e sete dos repositórios cumprem com os critérios 3 e 6, respectivamente, expondo em sua descrição ou em documentação complementar o plano de continuidade para salvaguarda de médio a longo prazo do domínio em que os dados são armazenados e assegurando a participação de especialista na decisão dos dados que neles são armazenados. Menos da metade dos repositórios estão em conformidade com os critérios 2, 4, 8, 10 e 12.

CONSIDERAÇÕES FINAIS

O estudo se limitou à avaliação das características dos repositórios a partir dos materiais bibliográficos e descritivos oferecidos pelos mesmos e à exploração de seus domínios. Apesar de nenhum dos repositórios consultados possuir o selo de certificação *CoreTrustSeal,* em sua maioria os repositórios têm seguido os critérios nela previstos, garantindo a segurança e a preservação dos dados. Desde 2018, vem sendo cobrada uma taxa administrativa de € 1.000,00, equivalente à R$ 5.017,17[6], pelo processo de revisão dos repositórios que buscam a certificação *CoreTrustSeal*. O alto custo do processo da certificação pode dificultar a certificação dos repositórios de instituições públicas, que constantemente encontram dificuldades para a obtenção de financiamentos que cubram despesas de alto valor além das necessidades básicas de manutenção dos repositórios.

Por ter se limitado aos registros publicados e não utilizar o contato direto com as equipes de gestão dos repositórios, é reconhecido que alguns critérios possam ter sido avaliados de forma indevida. Diante disso, sugere-se o aprofundamento desta pesquisa, consultando diretamente as instituições responsáveis pelos repositórios para se obter um quadro de avaliação mais próximo de sua realidade. Mesmo diante de suas limitações, cabe destacar os esforços das instituições na descrição transparente e aberta das características de seus repositórios e das políticas que os regem. Nesse cenário, é sugerido que cada vez mais a comunidade acadêmico-científica possa ter acesso a estas informações de modo a credibilizar os espaços em que seus dados são armazenados.

Por fim, ressalta-se a ideia de que a descrição mais esclarecida dos repositórios de dados, defendida por Kim (2020), pode aumentar a garantia da qualidade dos dados de pesquisa publicados em revistas híbridas ou *data journals*. Mesmo os repositórios que não possuem a certificação podem se beneficiar da explicitação dos critérios de qualidade e segurança que adotam. Para que as práticas de compartilhamento e reuso de dados sejam bem quistas pela comunidade acadêmica, é preciso fortalecer a confiança dos pesquisadores nas ferramentas utilizadas para sua execução.

REFERÊNCIAS

CORETRUSTSEAL TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS 2020–2022. Zenodo, 2019. Disponível em: https://zenodo.org/record/3638211#.YltNEXXMK00. Acesso em: 20 set. 2022.

CORETRUSTSEAL. **About**. S/d. Disponível em: https://www.coretrustseal.org/about/. Acesso em: 20 set. 2022.

CORETRUSTSEAL. **Review of Requirements**. S/d. Disponível em: https://www.coretrustseal.org/why-certification/review-of-requirements/. Acesso em: 20 set. 2022.

CORETRUSTSEAL. **Administrative fee**. S/d. Disponível em: https://www.coretrustseal.org/apply/administrative-fee/. Acesso em: 20 set. 2022.

KIM, Jihyun. An Analysis of Data Paper Templates and Guidelines: Types of Contextual Information Described by Data Journals. **Science Editing**, vol. 7, n. 1, p. 16–23, 2020. Disponível em: https://doi.org/10.6087/kcse.185. Acesso em: 20 set. 2022.

LIN, D. *et al*. The TRUST Principles for digital repositories. **Sci Data**, vol. 7, n. 144, 2020. Disponível em: https://doi.org/10.1038/s41597-020-0486-7. Acesso em: 20 set. 2022.

MAPPING THE GLOBAL REPOSITORY LANDSCAPE. 2022. Disponível em: https://coref.project.re3data.org/blog/mapping-the-global-repository-landscape. Acesso em: 20 set. 2022.

SALES, L. F.; SAYÃO, L. F. Uma proposta de taxonomia para dados de pesquisa. **Conhecimento em Ação**, v. 4, n. 1, 2019. Disponível em: https://revistas.ufrj.br/index.php/rca/article/view/26337. Acesso em: 20 set. 2022.

NOTAS

[1] A certificação *CoreTrustSeal foi* lançada pelo WDS e o *Data Seal of Approval* (DSA), com o objetivo de garantir a qualidade de repositórios digitais a partir do estabelecimento de 16 critérios formais embasados nos padrões *nestor-Seal DIN* 31644 e ISO 16363, que abrangem aspectos sobre infraestrutura organizacional, gestão de objetos digitais e Tecnologia (CORETRUSTSEAL, *s.d*).

[2] O Re3data é um diretório global, criado em 2012 com o financiamento pela Fundação Alemã de Pesquisa, que abrange repositórios de dados de pesquisa de diversas áreas do conhecimento científico.

[3] A localização da instituição que administra um repositório não necessariamente é mesma localização onde são armazenados os seus servidores. Por exemplo, o servidor do Dataverse de uma instituição brasileira pode estar localizado no país de origem do sistema, enquanto esta o administra remotamente do Brasil.

[4] *Links* inválidos correspondem à *URLs* corrompidas. O repositório não abre a partir do *link* fornecido.

[5] Repositórios que ainda não foram completamente desenvolvidos, mas se encontram disponíveis para consulta em versão beta.

[6] De acordo com a cotação do Banco Central do dia 18 de Abril de 2022.