Pecha Kucha

DOI: 10.21680/2447-7842.2023v9n2ID33825

Projeto Laguna: infraestrutura de um lago de dados científicos em acesso aberto

Laguna project: infrastructure of na open access scientific data lake

Washington Luís Ribeiro de Carvalho Segundo [1]
Adilson Luiz Pinto [2]
Fabio Lorensi do Canto [3]
Patricia da Silva Neubert [4]

Submetido em: 17/04/2023	Aprovado na ConfOA: 14/06/2023	Publicado em: 04/12/2023

Resumo: Existe um grande volume de dados sobre o ecossistema de Ciência, Tecnologia e Inovação brasileiro espalhados em diferentes bases de dados e repositórios. No entanto, nem todas as fontes permitem a recuperação e análise de dados de forma centralizada. O projeto Laguna foi desenvolvido para criar uma infraestrutura contendo dados de fontes abertas em diferentes formatos e níveis de tratamento, baseada no conceito de lago de dados, um repositório amplo e centralizado que permite armazenar dados estruturados, semi-estruturados e não estruturados em qualquer escala e extraídos de fontes baseadas nos princípios FAIR. O objetivo é aumentar a visibilidade e o uso dos resultados de pesquisa, incluindo patentes, programas de computador e outros produtos, serviços e processos com potencial para aplicação prática e exploração econômica, podendo ainda auxiliar processos de avaliação por agências de fomento, construindo indicadores adequados ao cenário brasileiro.

[1] Doutorado em Informática.
[2] Doutorado em Documentação.
[3] Doutorado em Ciência da Informação.
[4] Doutorado em Ciência da Informação.

Palavras-chave: ciência aberta; repositórios abertos; dados científicos; lago de dados; informação científica.

Abstract: There is a large volume of data on the Brazilian Science, Technology and Innovation ecosystem spread across different databases and repositories. However, not all sources allow for centralized data retrieval and analysis. The Laguna project was developed to create an infrastructure containing data from open sources in different formats and levels of treatment, based on the concept of a data lake, a broad and centralized repository that allows storing structured, semi-structured and unstructured data at any scale and drawn from sources based on FAIR principles. The objective is to increase the visibility and use of research results, including patents, computer programs and other products, services and processes with potential for practical application and economic exploitation, and may also help evaluation processes by development agencies, building adequate indicators to the Brazilian scene.

Keywords: open science; open repositories; scientific data; data lake; scientific information.

1 INTRODUÇÃO

O projeto Laguna, coordenado pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) em colaboração com pesquisadores de quatro instituições federais de ensino superior brasileiras, tem por objetivo a criação de uma infraestrutura informacional aberta para dados do ecossistema de informações brasileiro em Ciência, Tecnologia e Inovação (CT&I). O projeto organiza sua infraestrutura computacional com base em um lago de dados (Fang, 2015), um modelo de repositório centralizado que permite armazenar dados estruturados, semiestruturados e não estruturados em qualquer escala e extraídos de diferentes fontes (Giebler *et al.*, 2019). O projeto está em implementação, com conclusão estimada em 36 meses. Uma infraestrutura baseada em um *cluster* de servidores

on-premises foi montada e os primeiros testes de coleta e de processamento de dados estão sendo realizados.

2 METODOLOGIA

O lago de dados será composto por dados coletados em repositórios e bases de dados de relevância reconhecida em CT&I que atendam total ou parcialmente aos princípios FAIR (Wilkinson *et al.*, 2016). As fontes internacionais de dados previamente selecionadas foram o OpenAlex[5], o Wikidata[6], o CrossRef[7], o OpenCitations[8], o OpenAIRE Research Graph[9], o ISSN Portal[10], o Sistema Regional de Información em línea para Revistas Científicas de América Latina, el Caribe, España y Portugal (Latindex)[11], o Directory of Open Access Journals (DOAJ)[12] e o Google Scholar Metrics[13]. As fontes brasileiras pré-selecionadas foram as plataformas Lattes[14] e Sucupira[15], Oasisbr[16] e a Biblioteca Digital de Teses e Dissertações (BDTD)[17].

Os conjuntos de dados coletados nessas fontes serão submetidos a diferentes tipos de análise, incluindo técnicas avançadas de inteligência artificial, análise em tempo real, aprendizado de máquina, painéis e visualizações. Esse conjunto de técnicas potencializa o valor dos dados (Fang, 2015; Giebler *et al.*, 2019). A metodologia de tratamento e preparação dos dados compreenderá as seguintes etapas:

a) Coleta: Os dados serão coletados por meio de APIs públicas ou ferramentas de busca e extração disponíveis em repositórios e bancos de dados que

[5] Disponível em: https://openalex.org/.
[6] Disponível em: https://www.wikidata.org/?uselang=pt.
[7] Disponível em: https://www.crossref.org/.
[8] Disponível em: https://opencitations.net/.
[9] Disponível em: https://graph.openaire.eu/.
[10] Disponível em: https://portal.issn.org/.
[11] Disponível em: https://latindex.org/latindex/.
[12] Disponível em: https://doaj.org/.
[13] Disponível em: https://scholar.google.com/citations?view_op=metrics_intro&hl=en
[14] Disponível em: https://www.lattes.cnpq.br/.
[15] Disponível em: https://sucupira.capes.gov.br/sucupira/.
[16] Disponível em: https://oasisbr.ibict.br/vufind/.
[17] Disponível em: https://bdtd.ibict.br/vufind/.

atendem total ou parcialmente aos princípios FAIR. Se necessário, serão utilizadas ou desenvolvidas ferramentas de extração de dados (web scraping) de fontes complexas. Será utilizado um protocolo para envio e recebimento de mensagens, que opera via solicitações em interfaces REST (Representational State Transfer), realizando chamadas HTTP (HyperText Transfer Protocol) com respostas de documentos em formato JSON (JavaScript Object Notation). Também será utilizado o protocolo OAI-PMH (Open Archives Initiative - Protocol Metadata Harvesting), com chamadas HTTP, mas com respostas XML (eXtensible Markup Language) em diferentes padrões: desde o OAI-DC (Open Archives Initiative - Dublin Core) até modelos mais genéricos, como o RDF (Resource Description Framework), com alto poder de expressividade.

b) Seleção e separação: A seleção e separação dos dados serão realizadas por meio de filtragem e categorização. As informações auxiliares para a coleta (as informações adicionais) serão eliminadas, e os arquivos coletados serão desmembrados entre os diferentes tipos de entidades descritas em seu conteúdo (também chamado de carga útil).

c) Transformação e conexão: Os dados já fragmentados, classificados e categorizados podem ser ainda mais adaptados e validados, formando relacionamentos com registros de outras fontes.

d) Organização, classificação e indexação: Os dados serão organizados, classificados e indexados. As classificações servirão como base para a construção de interfaces de busca, serviços da web e painéis de visualização.

e) Recuperação e visualização: Indicadores serão construídos para serem exibidos em painéis de visualização. Ferramentas de exibição serão usadas para redes de colaboração, dados geoespaciais, séries temporais e esquemas de tabulação dinâmica, entre outros. Um modelo semântico será usado de acordo com padrões internacionais, compatível com representações de sistemas usados em outros países, visando obter níveis avançados de interoperabilidade.

3 RESULTADOS ESPERADOS

O principal resultado esperado é o desenvolvimento de um amplo repositório de dados abertos relativos à CT&I brasileira, com infraestrutura de lago de dados e armazenamento e processamento em nuvem, que possibilitará o alcance de outros resultados, decorrentes do tratamento e da análise dos conjuntos de dados, como a criação de sistemas de recomendação de entidades científicas.

4 CONSIDERAÇÕES FINAIS

O projeto tem como objetivo aumentar a visibilidade, a recuperação e o uso dos resultados de pesquisas, incluindo informações sobre patentes, programas de computador e outros produtos, serviços e processos com potencial para aplicação prática e exploração econômica. Também com potencial de auxiliar os processos de avaliação, com base na análise de indicadores apropriados para o cenário brasileiro, minimizando a dependência de plataformas comerciais para avaliação científica.

REFERÊNCIAS

Fang, H.L. (2015). Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. In: *Proceedings of the 2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)* (pp. 820–824).

Giebler, C; Göger, C. Hoos, E., Schwarz, H. & Mitschang, B. (2019). Leveraging the Data Lake: Current State and Challenges. In: Ordonez, C., Song, IY., Anderst-Kotsis, G., Tjoa, A., Khalil, I. (Eds.), *Big Data Analytics and*

Knowledge Discovery. DaWaK 2019. Lecture Notes in Computer Science, v. 11708. Springer, Cham. https://doi.org/10.1007/978-3-030-27520-4_13

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton G., Axton, M., *et al.* (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3 (1), 1–9. https://doi.org/10.1038/sdata.2016.18