15ª Conferência Lusófona de Ciência Aberta (ConfOA)
Acesso Aberto e Dados de Investigação Abertos: sistemas, políticas e práticas

Modalidade: Pecha Kucha

A PUBLICAÇÃO DE DADOS EM HUMANIDADES DIGITAIS COM FAIR: *datasets* de pesquisa do projeto MPO compartilhados no CKAN da UNIRIO

Cláudio José Silva Ribeiro
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)/Programa de Pós-graduação em Biblioteconomia (PPGB).
Rio de Janeiro, RJ, Brasil
https://orcid.org/0000-0002-9571-1707

Martha Tupinambá de Ulhôa
Universidade Federal do Estado do Rio de Janeiro (UNIRIO)/Programa de Pós-graduação em Música (PPGM).
Rio de Janeiro, RJ, Brasil
http://orcid.org/0000-0002-6886-1267.

RESUMO:
A iniciativa *Open Science* trouxe um incremento para a disponibilização de dados de pesquisa e a possibilidade de reutilizá-los no contexto de novas investigações. Este relato apresenta o protótipo com a iniciativa de compartilhamento de dados do projeto Música em Periódicos Oitocentistas seguindo as recomendações apresentadas pelos princípios FAIR. A metodologia é um estudo de caso que combina métodos, reunindo os aspectos teóricos e práticos para a publicação dos conjuntos de dados gerados pelo processo de FAIRificação. Os resultados apresentam a estrutura de um conjunto de dados que contém registros com descrições textuais de notícias sobre música no século XIX. Os conjuntos de dados estão publicados no CKAN sob licenciamento *Creative Commons Attribution License*.

Palavras-chave: periódicos musicais; CKAN; reúso de dados de pesquisa; século XIX.

INTRODUÇÃO

O movimento do acesso aberto trouxe uma expectativa crescente para pesquisadores à medida que apresenta muitos benefícios, em especial com a redução de tempo dispendido nas fases iniciais da pesquisa. É possível afirmar que pesquisadores utilizam boa parte do tempo de suas investigações fazendo a coleta, organização e provendo a garantia de qualidade dos dados, ao invés de utilizar o tempo na própria pesquisa (Chen; Jagerhorn, 2022).

Reduzir o tempo utilizado no tratamento de dados de pesquisa pode ser benéfico para novas pesquisas. Sinaci *et al*. (2020, p. 30) convalidam nossa afirmação quando observam que o fluxo de trabalho relacionado ao processo de *FAIRification* impulsiona a descoberta do conhecimento em áreas específicas e incrementa o reúso de dados.

Ademais, sabe-se que esse reúso para dados que foram gerados com o uso de recursos públicos é elemento importante e contribui para o desenvolvimento de uma sociedade mais justa e igualitária (Silva, 2021). Em pesquisa recente e convalidando nossa intenção de investigação, Martin-Melon, Hernández-Pérez e Martínez-Cardama (2023) registram que a *The EUA Open Science Agenda* 2025 determina como área prioritária o tratamento dos dados de pesquisa e destacam, ainda, que essa tarefa é elemento fundamental para possibilitar o sucesso da investigação e garantir sua transparência científica com acesso aberto e reúso.

Esta iniciativa implementa a etapa final do projeto que se iniciou como registrado em trabalhos anteriores e sintetizados a seguir. Em Ribeiro *et al*. (2023) e Ribeiro e Ulhôa (2024, no prelo) foram apresentados: a definição do escopo da investigação e o planejamento das atividades para o reúso de possíveis interligações com projetos e verbetes Wikipedia e Wikidata. Em Ribeiro e Ulhôa (2023) foram identificadas as questões de competência, metadados candidatos e foi dado início ao fluxo de FAIRificação. Já em Ribeiro e Ulhôa (2024a) foi enunciado e explorado um caso de reúso de dados textuais presentes no *dataset* que serviu de campo empírico por este protótipo, mas antes de disponibilizá-lo no CKAN. Esta trajetória reflete o estudo de caso completo que passou pelos ciclos de implementação de projeto (planejamento, execução, entregas e monitoramento) (PMBOK, 2021) em conjunto com o cumprimento das etapas apresentadas nos subfluxos de pré-FAIRificação, de FAIRificação e pós-FAIRificação.

Diante dessa contextualização, este relato tem como objetivo demonstrar o processo de publicação na fase de pós-FAIRificação, em infraestrutura *Comprehensive Knowledge Archive Network* (CKAN) já presente na UNIRIO, para disponibilização, reúso e monitoramento de um conjunto de dados extraídos da base de Música em Periódicos Oitocentistas (MPO).

O contexto do protótipo

Pode-se identificar diferentes iniciativas promovidas por vários entes ligados ao ensino e à pesquisa em prover infraestrutura para dados de pesquisa (Gabriel Junior *et al*., 2022). A UNIRIO vem trabalhando em dotar o seu ambiente de divulgação científica de soluções tanto para a gestão e preservação de dados de pesquisa (*DataHórus*), quanto para publicações (*Hórus*) (UNIRIO, 2018).

A solução *DataHórus* ainda está em estágio de planejamento, não permitindo seu uso como ambiente para compartilhamento de dados. A proposta de gerar um protótipo com o CKAN foi motivada pela existência de infraestrutura dessa tecnologia na UNIRIO e, portanto, reduzindo o esforço para implantação de outra tecnologia em um contexto de exiguidade de recursos (pessoas, processos e ferramentas). O CKAN traz, na visão de Costa *et al*. (2017), facilidades, pois "[...] os dados são depositados pelos usuários por meio da interface provida pela ferramenta ou pela *application programming interface* (API). Os dados, ou conjunto de dados, são descritos conforme suas descrições (metadados) [...]" (p. 15, grifo nosso). Zastrow e Fabas (2023) convalidam a proposta de uso do CKAN e registram que a solução tem a vantagem de possibilitar o armazenamento de metadados e objetos de forma independente. Beer *et al*. (2023) complementam e asseveram que a possibilidade de integrar múltiplos formatos é um item de destaque para o CKAN.

No contexto do protótipo, outro ponto favorável à adoção do CKAN para a descrição dos conjuntos de dados é que sua estrutura é compatível com os padrões DCAT e *Dublin Core*. Isso torna viáveis os processos de harmonização e interoperabilidade com outras infraestruturas para publicação e disseminação de conjuntos de dados (Albertoni *et al*., 2023; Datacite ..., 2024).

A versão do CKAN que está disponível na UNIRIO é 2.6.2 e a estrutura de *datasets* e recursos planejada foi organizada segundo essa versão. A estrutura segue o preconizado em Costa *et al*. (2017, p.16) com a definição de Organização, múltiplos Conjuntos de Dados, Recursos (é o conteúdo do *dataset*) e possibilidade de visualização. Os Conjuntos de Dados podem ser descritos com: título; descrição; etiquetas; licença; organização; visibilidade; fonte e versão. Além desses, foram acrescidos os metadados: nome do autor e nome do mantenedor.

Descrição dos recursos

Como mencionado anteriormente, no CKAN os *datasets* são carregados como recursos. Para cada conjunto de dados foram implementados diferentes recursos que objetivam tornar o seu uso facilitado. Os campos para detalhamento de cada recurso também seguiram a proposta de Costa *et al.* (2017) contendo os seguintes elementos: descrição, formato, fonte (atributo de proveniência que descreve a origem do *dataset*), autor, mantenedor, versão, data de criação, data de atualização e licenciamento (definido na criação do recurso: *cc-by*).

Os recursos são a matéria-prima do reúso. Quanto mais opções de recursos forem publicadas, maior a possibilidade de utilização e reutilização. Nesse sentido, o uso de formatos que são passíveis de interpretação semântica por meio de mecanismos automatizados é essencial para a adoção de FAIR. Formatos como JSON, XML, RDF e OWL devem ser incorporados como recursos para viabilizar a compreensão semântica de conjuntos de dados ligados. Segundo Soiland-Reyes *et al*. (2022), isso ocorre tanto no contexto de *linked data,* quanto para uso de *FAIR Data Objects* (FDO). Em ambas as abordagens o enriquecimento é útil para uso programático via API[1] no CKAN.

PROCEDIMENTOS METODOLÓGICOS

A presente pesquisa é um estudo de caso que se caracteriza como de natureza aplicada e exploratória em relação ao seu objetivo. A complementação da revisão de literatura se deu sobre BRAPCI, Google Acadêmico e Portal de Periódicos CAPES.

A extração de dados foi desenvolvida por meio de *queries* SQL sobre o banco de dados PostgreSQL. Os resultados foram carregados no *OpenRefine* para tratamento conforme apresentado em Ribeiro e Ulhôa (2023), onde foram promovidas as complementações para ajustes na semântica do conjunto de dados.

O cadastramento dos *datasets* seguiu a interface do CKAN com o preenchimento dos metadados e posteriormente a criação de um recurso que contém o conjunto de dados em formato csv. Os dados foram carregados nesse recurso por meio de *upload*.

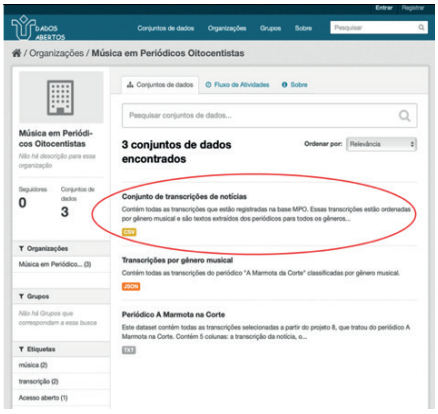
RESULTADOS

A descrição e o formato foram definidos na concepção do *dataset*. No protótipo, optou-se por inserir os dados do projeto MPO como fonte, pois esse atributo descreve a proveniência, garantindo a referência ao projeto de origem (Ribeiro e Ulhôa, 2023). No campo de mantenedor optou-se por inserir os dados dos responsáveis pelo ambiente de cadastramento. Não houve personalização de metadados, pois na definição deste protótipo não foi identificada a necessidade de incorporar novos metadados ou campos personalizados.

Foi gerado um registro de organização no CKAN denominado Música em Periódicos Oitocentistas. Em seguida, foram criados os grupos apresentados na figura 1: conjunto de transcrições de notícias (etiqueta formato CSV); transcrições por gênero musical (etiqueta formato JSON) e periódico Marmota na Corte (etiqueta formato txt).

1 Disponível em: https://docs.ckan.org/en/2.9/api/.

FIGURA 1 – Estrutura de *datasets* utilizada no protótipo



Fonte: dados.unirio.br.

FIGURA 2 – *Dataset* de transcrições de notícias

Fonte: dados.unirio.br.

Os dados foram carregados como recursos. Para cada conjunto de dados foram propostos diferentes recursos que objetivam tornar o reúso facilitado. A figura 2 apresenta a interface do CKAN com um recurso publicado em formato CSV.

CONSIDERAÇÕES FINAIS

Estes resultados apontam para a viabilidade da proposta e sinalizam para uma ação multiplicadora deste piloto, pois contribuem para a adoção de pressupostos da *Open Science* e compartilhamento de dados de pesquisa com o propósito de reúso em investigações no campo das humanidades digitais. As etapas do *FAIRification workflow* sistematizam o processo de transformação e contribuem para agilizar o trabalho. Como continuidade, pretende-se que os conjuntos de dados que estão sendo investigados e coletados estejam disponíveis e publicados fazendo uso do *FairDataPoint* (FDP). O nível de *FAIRness* deste protótipo poderá ser avaliado na etapa de alinhamento com o *FairDataPoint*, conforme Santos *et al*. (2023).

Cabe registrar que a iniciativa de compartilhamento aqui apresentada foi utilizada como fonte para o reúso de dados de pesquisa. O *dataset* destacado na figura 2 possui o total aproximado de 3.700 registros de conteúdo em formato texto, que representam as notícias publicadas em jornais do século XIX. Esse *corpus textual* foi explorado em pesquisa sobre *text as data* (Ribeiro; Ulhôa, 2024a).

Espera-se que a apresentação deste relato, que registra uma experiência prática investigativa no campo das humanidades digitais, possa contribuir com outros esforços semelhantes que buscam a adoção de princípios FAIR em repositórios de dados de pesquisa.

REFERÊNCIAS

ALBERTONI, Riccardo *et al*. The W3C Data Catalog Vocabulary, version 2: Rationale, design principles, and uptake. **Data Intelligence**, [*s.l.*], v. 6, n. 2, p. 457-487, 2024.

BEER, Anna *et al*. **Leibniz data manager – an adaptive research data management system**. *In:E-SCIENCE-TAGE 2023: EMPOWER YOUR RESEARCH – PRESERVE YOUR DATA,* 2023, Heidelberg. **Anais** […] Heidelberg, 2023. p 1. DOI: https://doi.org/10.11588/heidok.00033144. Disponível em: https://archiv.ub.uni-heidelberg.de/volltextserver/33144/7/leibnitz_data_manager_E-science-Tage_2023.pdf. Acesso em: 12 fev. 2024.

CHEN, Xiaoli; JAGERHORN, Martin. Implementing FAIR workflows along the research lifecycle. **Procedia Computer Science**, [*s.l.*], v. 211, p. 83-92, 2022. DOI: https://doi.org/10.1016/j.procs.2022.10.179. Disponível em: https://www.sciencedirect.com/science/article/pii/S1877050922016441. Acesso em: 12 fev. 2024.

COSTA, Lucas Rodrigues *et al*. **Guia do usuário CKAN**. Brasília: Ibict, 2017. 80p. ISBN 978-85-7013-126-3. 2017. DOI: 10.18225/978-85-7013-126-31. Disponível em: http://ridi.ibict.br/handle/123456789/1113. Acesso em: 14 fev. 2024.

DATACITE METADATA WORKING GROUP *et al*. **DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs Note**. DataCite eV, 2024.

GABRIEL JUNIOR, Rene Faustino *et al*. **Iniciativa para o desenvolvimento e implementação de repositórios de dados de pesquisa**. RNP/Ibict/CNPq. 2022.

MARTIN-MELON, Roberto; HERNÁNDEZ-PÉREZ, Tony; MARTÍNEZ-CARDAMA, Sara. Research data services (RDS) in Spanish academic libraries. **The Journal of Academic Librarianship**, v. 49, n. 4, p. 102732, 2023.

PINTO, Adilson Luiz *et al*. Brazil Developing Current Research Information Systems (BrCRIS) as data sources for studies of research. **Iberoamerican Journal of Science Measurement and Communication**, [*s.l.*], v. 2, n. 1, 2022.

PROJECT MANAGEMENT INSTITUTE. **A guide to the project management body of knowledge (PMBOK Guide)**. 7 ed. Newtown Square, PA: Project Management Institute, 2021.

RIBEIRO *et al*. Knowledge Organization no Processo de FAIRificação de Datasets: Estruturando a Semântica e Interligando as Notícias do Banco de Dados de Periódicos Musicais Oitocentistas. *In*: ISKO-Brasil, Londrina. Organização e do conhecimento em diferentes contextos: desafios e perspectivas na era da datificação, 2023. **Anais [...]**. Londrina: PPGCI-UEL, 1, p. 363-372.

RIBEIRO, Claudio José Silva; ULHÔA, Martha Tupinambá de. El uso compartido de conjuntos de datos de investigación del proyecto Música en Periódicos Ochocentistas: un prototipo con el uso de la solución CKAN. **Revista EDICIC**, [*s.l.*], v. 3, n. 3, p. 1-18, 2023. Disponível em: http://ojs.edicic.org/index.php/revistaedicic/article/view/214. Acesso em: 28 fev. 2024.

RIBEIRO, Claudio José Silva; ULHÔA, Martha Tupinambá de. Reúso de dados de pesquisa em humanidades digitais: investigando textos em notícias do projeto Música em Periódicos Oitocentistas. *In*: XIV EDICIC – Diálogos na Ciência da Informação, 2024a. **Anais [...].** Lisboa.

RIBEIRO, Claudio José Silva; ULHÔA, Martha Tupinambá de. Linked data com Wikipédia e Wikidata: reduzindo os silos de informação na Web com notícias sobre Música em Periódicos Oitocentistas (MPO). *In*: PESCHANSKI, João Alexandre; JURNO, Amanda Chevtchouk. (orgs.). **A Wikimedia no Brasil:** o poder e os desafios do conhecimento livre. 2024b. EDUFBA, no prelo.

SANTOS, Luiz Olavo Bonino da Silva *et al*. FAIR data point: a FAIR-oriented approach for metadata publication. **Data Intelligence**, [*s.l.*], v. 5, n. 1, p. 163-183, 2023.

SILVA, Fabiano Couto Corrêa. **Gestão de dados científicos**. Interciência, 2021.

SINACI, A. Anil *et al*. From raw data to FAIR data: the FAIRification workflow for health research. **Methods of Information in Medicine**, [*s.l.*], v. 59, n. S 01, p. e21-e32, 2020.

SOILAND-REYES, Stian *et al*. Updating linked data practices for FAIR digital object principles. **Research Ideas and Outcomes**, [*s.l.*], v. 8, p. e94501, 2022.

UNIRIO. **Política de acesso aberto à informação técnico-científica e aos dados de pesquisa da Universidade Federal do Estado do Rio de Janeiro**. Resolução 5.055 de 18 de outubro de 2018. Disponível em: https://www.unirio.br/bibliotecacentral/arquivos/Resolucaono5.055de10deOutubrode2018.PDF.

ZASTROW, Thomas; FABAS, Nicolas. Research data publication at large scale. **Proceedings of the Conference on Research Data Infrastructure**, [*s.l.*], v. 1, 2023. DOI: 10.52825/cordi.v1i.289. Disponível em: https://www.tib-op.org/ojs/index.php/CoRDI/article/view/289. Acesso em: 7 ago. 2024.