INTEROPERANDO DADOS DE PESQUISA DO DATAVERSE PARA O FAIR DATA POINT

Interoperating Dataverse Lookup Data to the FAIR Data Point
Interoperabilidad de datos de investigación de Dataverse en el FAIR Data Point

Henrique Fernandes Rodrigues
Bacharelando em Ciência da Computação, Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro – RJ, Brasil.
Auxiliar de Pesquisa, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília - DF, Brasil.
Lattes: http://lattes.cnpq.br/9081660982164633
ORCID: https://orcid.org/0000-0002-6777-3935

Tatyane Guedes Martins da Silva
Bacharela em Biblioteconomia, Universidade de Brasília (UnB), Brasília - DF, Brasil.
Bibliotecária - Bolsista, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília - DF, Brasil.
Lattes: http://lattes.cnpq.br/7310861285054095
ORCID: https://orcid.org/0000-0002-1743-0467

Cássio Teixeira de Morais
Bacharel em Biblioteconomia e Arquivologia, Universidade de Brasília (UnB), Brasília – DF, Brasil. Bibliotecário – Pesquisador, Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasília – DF, Brasil.
Lattes: http://lattes.cnpq.br/3368268946691719
ORCID: https://orcid.org/0000-0003-2840-4624

Rene Faustino Gabriel Junior
Doutor em Ciência da Informação. Universidade Federal do Rio Grande do Sul (UFRGS) Porto Alegre - RS, Brasil.
Professor Adjunto, Universidade Federal do Rio Grande do Sul (UFRGS) Porto Alegre - RS, Brasil.
Lattes: http://lattes.cnpq.br/5900345665779424
ORCID: http://orcid.org/0000-0003-1021-3360

Resumo

Introdução: A expansão das iniciativas em ciência aberta gera a necessidade de implementar repositórios de dados de pesquisa para disponibilização e preservação desses insumos. Os princípios FAIR apresentam diretrizes para expor os dados de forma apropriada e possibilitar seu reuso. Nesse contexto, os metadados

que descrevem esses conteúdos são de suma importância, viabilizando o reconhecimento e transparência sobre os dados existentes. Assim, portais de dados em Dataverse, como o LattesData, possibilitam a descrição rica de metadados de seus conjuntos de dados. Aproveitar essas descrições com o uso de ferramentas semânticas como o FAIR Data Point, possibilitam a ampliação da consulta e transmissão dos metadados. Dessa forma, é fundamental que se tenha interoperabilidade entre portais dessa natureza. **Metodologia**: Esse artigo tem o objetivo relatar a experiência da exploração da viabilidade e estudos das possibilidades da integração automatizada desses portais por meio de seus metadados. **Resultados**: Como resultado é apresentada uma análise sobre os metadados e a construção do código *crosswalk* responsável pela compatibilidade e integração. **Conclusão**: Assim, os metadados são disponibilizados no catálogo FAIR Data Point para consultas em triplas como o SPARQL, demonstrando a importância e viabilidade do processo em estudo.

Palavras-chave: dados de pesquisa; FAIR Data Point; dataverse; crosswalk.

Abstract

***Introduction**: The expansion of open science initiatives generates the necessity of implementing research data repositories to make these inputs available and to be preserved. The FAIR principles provide guidelines for properly exposing data and enabling reuse. In this context, the metadata that describe these contents are of paramount importance, enabling acknowledgement and transparency over existing data. Thus, Dataverse data portals, such as LattesData, enable the rich metadata description of its datasets. Taking advantage of these descriptions with the use of semantic tools such as the FAIR Data Point, facilitate the query and transmission of metadata. Thus, it is important to have interoperability between portals of this nature. **Methodology**: This article aims to report the experience of exploring the feasibility and studies of the possibilities of automated integration of these portals through their metadata. **Results**: As a result, an analysis of the metadata and the construction of the crosswalk code responsible for compatibility and integration is presented. **Conclusion**: Thus, the metadata are made available in the FAIR Data Point catalog for queries in triples such as Sparql, demonstrating the importance and feasibility of the process under study.*

***Keywords**: research data; FAIR Data point; dataverse; crosswalk.*

Resumen

***Introducción**: La expansión de la ciencia abierta está creando la necesidad de implementación de depósitos de datos de investigación para que estas entradas estén disponibles y se conserven. Los principios FAIR proporcionan pautas para exponer correctamente los datos y permitir su reutilización. En este contexto, los metadatos que describen estos contenidos son de suma importancia, permitiendo una fácil exploración y transparencia sobre los datos existentes. Por lo tanto, los portales de datos de Dataverse, como LattesData, permiten la rica descripción de metadatos de sus conjuntos de datos. Aprovechando estas descripciones con el uso de herramientas semánticas como el FAIR Data Point, facilitar la consulta y transmisión de metadatos. Por lo tanto, es importante tener interoperabilidad entre portales de esta naturaleza. **Methodología**: Este artículo tiene como objetivo relatar la experiencia de explorar la factibilidad y estudios de las posibilidades de integración automatizada de estos portales a través de sus metadatos. **Resultados**: Como resultado, se presenta un análisis de los metadatos y la construcción del código crosswalk responsable de la*

compatibilidad e integración. ***Conclusión****: Así, los metadatos quedan disponibles en el catálogo FAIR Data Point para consultas en triples como Sparql, demostrando la importancia y viabilidad del proceso en estudio.*

Palabras clave*: datos de investigación; FAIR Data Point; dataverse; crosswalk.*

DOI 10.22477/vi.widat.54

1. Introdução

As pesquisas geram uma grande quantidade de dados, que são agrupados em conjuntos de dados, que são disponibilizados em repositórios de dados de pesquisa com o objetivo de possibilitar seu reuso, que por consequência pode beneficiar outras pesquisas. Contudo, todo esse trabalho pode estar comprometido se não forem resolvidas algumas dificuldades existentes pela falta de padronização e suporte adequado. *Softwares* como o Dataverse, permite que dados de pesquisa sejam armazenados e compartilhados apropriadamente. Além disso, os portais devem ser coesos com o processo de abertura científica, dialogar com os princípios FAIR, ou seja, sendo encontráveis, acessíveis, interoperáveis e viabilizando o correto reuso com uma eficiente descrição dos (meta)dados (UCSF, 2023).

Com a premissa FAIR, para que os princípios estabelecidos sejam alcançados, o uso de um conjunto mínimo de metadados se torna importante, visto que esses enriquecem semanticamente os dados possibilitando seu reuso, bem como a correta interpretação por máquinas (BONINO, 2023).

Neste contexto, o estudo tem como objetivo explorar a viabilidade de integração automatizada da plataforma Dataverse com o FAIR Data Point, com uma integração automatizada por meio de seus metadados. Para este fim, utilizou-se a instância do LattesData como fonte dos metadados.

2. Referencial teórico

Este artigo está organizado de forma a apresentar uma breve revisão bibliográfica dos conceitos intrínseco à pesquisa, e dos resultados obtidos a partir das reflexões e experimentos em um ambiente de testes.

2.1 Dataverse e o LattesData

O Dataverse está sendo desenvolvido, em código aberto, desde 2006, no *Harvard's Institute for Quantitative Social Science* (IQSS), junto de muitos colaboradores em todo o mundo para atender a demandas de repositórios para dados de pesquisa. O Dataverse é uma aplicação web visando compartilhar, preservar, citar, explorar e analisar dados de pesquisa. Com um forte embasamento arquivístico, este traz formas de preservar em longo prazo os dados. Ainda mais, este o faz de forma próxima ao ideal para trabalhar com dados, permitindo que os conteúdos publicados possuam melhores descrições, referências e acesso.

Neste sentido, essas características o aproximam dos princípios FAIR, importantes no contexto de trabalho e manutenção dos dados. Dessa forma, o Dataverse vem se destacando como referência na preservação de dados científicos, apresentando desenvolvimento e novas necessidades de integração (GDCC, 2022).

O LattesData[1], repositório em que se baseia o estudo, foi criado com o intuito de fomentar a ciência aberta por meio da disponibilização de dados de pesquisa, desenvolvido pelo CNPq, incentivando os pesquisadores a depositar e preservar seus dados de forma apropriada. Tais fatores, tornam o LattesData uma referência em repositório da produção científica nacional. Dessa forma, aprimorar sua estrutura e criar formas de expandir sua interoperabilidade são de grande importância para o desenvolvimento da pesquisa no Brasil (CNPQ, 2023).

2.2 Princípios FAIR

As primeiras manifestações referentes aos princípios FAIR surgiram no início de 2014 na conferência da *Jointly designing a data* FAIRPORT. Nesta conferência foram discutidos os obstáculos relativos à utilização e reutilização de dados e as propostas para a solução destes problemas, resultando no consenso da necessidade de criar uma infraestrutura global que suporte a abertura de dados de pesquisa voltados para a armazenamento, compartilhamento e a reutilização dos dados. As discussões ainda suscitaram a criação de princípios e práticas que pudessem orientar a descoberta, o acesso, a integração e a reutilização da vasta quantidade de dados e informação gerada pela ciência. Esses princípios foram denominando de FAIR, "acrônimo de *Findable, Accessible, Interoperable, Reusable*, visando implementar um conjunto de metadados definidos tanto para uso por mecanismos computacionais automatizados, quanto para uso por pessoas" (HENNING *et al.*, 2019).

1 Acessível em: https://lattesdata.cnpq.br/

2.3 FAIR Data Point

Um "FAIR Data Point" é uma implementação específica dos princípios FAIR, trata-se de um componente ou serviço que facilita a descoberta e o acesso a dados científicos de forma compatível com os princípios FAIR, o objetivo da implementação é fornecer uma interface padronizada para localizar e recuperar dados de forma mais eficiente (BONINO, 2023).

O FAIR Data Point é um catálogo que tem como objetivo preservar e difundir seu conteúdo. Alinhado aos princípios FAIR, esse os fomenta através de um elemento crucial, os metadados e, possibilitam melhor descrever os recursos existentes. Com isso, os recursos são facilmente encontrados e reutilizados visto sua boa descrição e difusão promovida (BONINO, 2023).

3. Metodologia

Para os processos deste estudo foi implementada uma instância nos servidores de trabalho em um servidor virtual composto por 4 núcleos de processador, 8GB de memória e 50GB de disco. Essa implementação utilizou o Docker, que possibilita maior facilidade na construção do serviço, bem como sua manutenção.

O processo consistiu em algumas etapas a fim de alcançar o objetivo. O desejado é realizar a exposição dos metadados presentes no repositório Dataverse LattesData em um catálogo de metadados FAIR Data Point. Para tal foi necessário: (i) realizar um estudo sobre os metadados presentes no Dataverse; (ii) identificar mecanismos para exportar os metadados de forma automática; (iii) distinguir como os metadados são armazenados no FAIR Data Point; (iv) determinar acesso e armazenamento automatizados ao FAIR Data Point; (v) mapear os metadados de acordo com o FAIR Data Point; (vi) ordenar a geração dos metadados triplificados e sua carga.

Uma vez levantados os metadados existentes, o fator principal é o alinhamento desses. Ou seja, fazer um mapeamento para que os metadados sejam bem representados em mais de uma estrutura. O processo de alinhamento se baseia nas fontes dos dados e seu valor semântico, por exemplo, ontologia empregada. Assim, um processo denominado *crosswalk*.

A primeira parte do processo requisitou a análise dos e sobre os metadados presentes no Dataverse. Para fins dos estudos, elegeu-se uma série de elementos descritivos, com um grupo mínimo inicial, mas podendo ser expandido de acordo com a necessidade com novos tipos de metadados.

Dos campos de metadados mínimos foram selecionados o *keyword; description; createTime*; e *title*. Em seguida, é importante verificar como tais metadados são obtidos. A plataforma em questão, o Harvard Dataverse, disponibiliza uma API que permite o acesso de dados, tanto de forma pública como com credenciais, permitindo ações mais restritas. De forma mais específica, foram selecionados os meios de acesso público, ou seja, sem restrições ou necessidade de permissões, verificando a possibilidade com qualquer base aberta. Esses utilizam requisições HTTP REST, para obter os dados.

Os metadados obtidos se apresentam com diferentes formatos: DDI, Dublin Core, JSON. Os conjuntos por esses apresentados, no entanto, são reduzidos para saídas em DDI e *dublin core*. Por tal fato, foi selecionada a saída por *json*, na qual, embora não sejam utilizados termos controlados, esse formato disponibiliza todos os metadados do conjunto de dados, não só os obrigatórios ou básicos, e ainda associa as chaves internas (Manual Dataverse API).

Para realizar a interoperabilidade com a estrutura do *FAIR Data Point* foi necessário identificar como os dados são representados. Identificou-se que o sistema apresenta um conjunto de dados ligados descritos com ontologias (*namespace*). Para fins do estudo, restringiu-se a estrutura mínima comporta por *dcat:keyword; dct:description; dct:issued*; e *dct:title*. É possível expandir e qualificar os metadados do FAIR Data Point, da mesma forma que o

Dataverse. Porém para fins deste estudo inicial, optou-se em primeiro testar a integração e futuramente expandir e qualificar os metadados para interoperabilidade.

De forma análoga ao processo realizado com o Dataverse é necessário analisar como pode ser realizada a ingestão de metadados de forma automatizada. O *FAIR Data Point* disponibiliza uma API baseada em HTTP REST, permitindo que requisições sejam feitas, passando os metadados. O repositório valida o conteúdo submetido, que deve ser composto por triplas. Em seguida, são criados os elementos: *catalog, dataset* e *distribution* dependendo da forma que foi feita a requisição e os dados fornecidos.

Uma vez verificados os metadados de origem e como são descritos, bem como os de destino, é importante verificar a relação entre esses. Assim, é realizado o processo de *crosswalk* (ZENG, 2006), verificando como as ontologias se alinham, definindo uma relação de equivalência dos valores na ontologia, esquema de destino. Desse modo, foi gerado um quadro (QUADRO 1) organizando essas informações.

Quadro 1 – Mapeamento dos Metadados do LattesData (Dataverse) e do FAIR Data Point (CrossWalk)

API Dataverse	API FAIR Data Point	Descrição
keyword	*dcat:keyword*	Palavras-chave
description	*dct:description*	Descrição dos Dados
createTime	*dct:issued*	Data
title	*dct:title*	Título dos dados

Fonte: Elaboração própria, com base nos dados da pesquisa (2023).

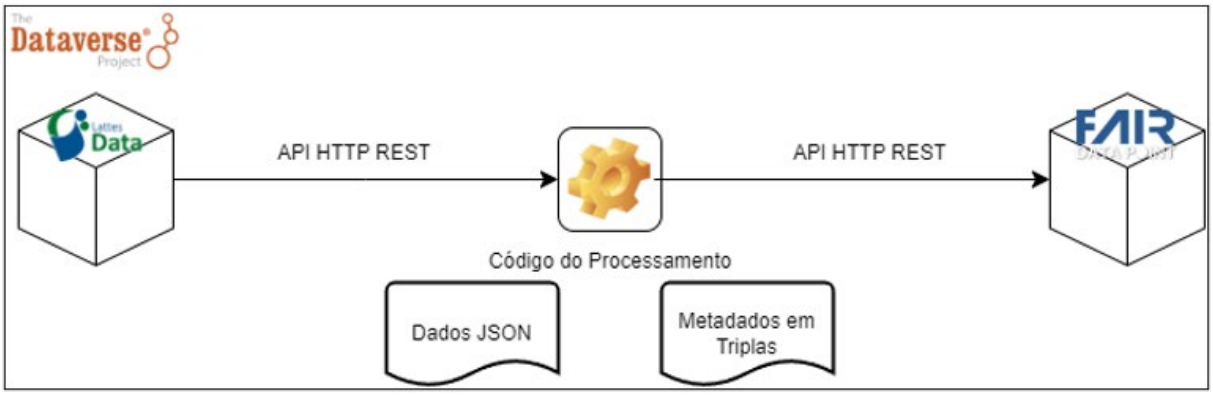
Uma vez definidas as relações dos metadados (*crosswalk*) de forma conceitual, foi necessário aplicar o modelado ao conjunto existente, bem como disponibilizar no destino, o *FAIR Data Point*. Assim, um script em *Python* foi desenvolvido para realizar os processos de automação. Dessa forma, o código recupera os dados do Dataverse por uma API, e converte no novo modelo através das triplas e, depois, carrega esses com novas requisições ao *FAIR Data Point,* monitorando a mensagem de retorno, que em caso de sucesso retorna 201. Todos os elementos após a ingestão dos metadados podem ser consultados pela interface gráfica ou recuperados pelas APIs.

4. Resultados

A fim de concretizar os procedimentos propostos, o script em código em Python foi aplicado para realizar a interoperabilidade entre o LattesData e o *FAIR Data Point*. Para tal, são explorados possíveis meios de conexão que em ambos, envolveu comunicação utilizando as APIs HTTP Rest. Através desse meio foram feitas as requisições dos dados, bem como a submissão para carga, apontando as configurações devidas.

O código em questão realiza a obtenção dos metadados existentes no LattesData e organizá-la-ás em nas triplas no modelo proposto no estudo. Por fim, com as triplas formadas, essas são carregadas no *FAIR Data Point*, que armazenará e difundirá os metadados conforme estruturado. A seguir, pode ser visto um diagrama que representa o processo descrito na Figura 1.

Figura 1 – Diagrama da coleta, processamento e ingestão dos metadados



Fonte: Elaboração própria, com base nos dados da pesquisa (2023).

Para os testes, o script foi inserido no agendador de tarefas do *Linux* para realizar o acionamento do *script* de forma sistematizada em horários pré-definido (Cron). O resultado da integração fica disponível para consulta do usuário visitante ou para serviços de descoberta.

Como foram utilizados os elementos mínimos de descrição de metadados, o resultado do Dataset apresentado na Figura 2, demonstra apenas os dados identificados no modelo proposto, para a incorporação de novos campos é necessário ampliar o modelo e adaptar o *script* em *Python* de forma a atender todos os metadados. Ressalta-se que este estudo foi experimental, para verificar e explorar a viabilidade de realizar a interoperabilidade entre essas duas plataformas.

Figura 2 – Tela inicial do FAIR Data Point

Fonte: Elaboração própria, com base nos dados da pesquisa (2023).

Uma das grandes vantagens do uso da *FAIR Data Point* é a possibilidade de realizar consultas em *SPARQL* diretamente nas triplas armazenadas, e com isso a integrações com outros sistemas que utilizem dados ligados (LinkedData).

5. Considerações finais

Uma vez explorados os metadados e possibilidades por meio da execução do planejado, é possível entender melhor o processo. Considera-se que o processo de integração das mencionadas plataformas é viável e pode trazer benefícios para o uso dos dados, como a visibilidade, bem como disponibilizar triplas para gerar consultas inteligentes e inferências pelos sistemas computadorizados.

De forma geral, considera-se que o trabalho atual realizou uma investigação e levantamento da interoperabilidade entre os sistemas, atendendo de forma satisfatória o objetivo proposto, contudo, o estudo não se encerra, sendo necessário melhorar e qualificar os metadados trocados entre os sistemas, principalmente com a incorporação da descrição dos dados da pesquisa, como o formato DDI disponibilizado pelo Dataverse, sendo essas problemáticas propostas de estudos futuros, ampliando os processos ontológicos aplicados na abordagem.

Referências

CONZETT, P. Dataverse Community Survey 2022: report. **Septentrio Reports**, v. 1, p. 1-75, 2022. Disponível em: https://septentrio.uit.no/index.php/SapReps/article/view/6872. Acesso em: 19 maio 2023.

HENNING, P. C. *et al.* GO FAIR e os Princípios FAIR: o que representam para a expansão dos dados de pesquisa no âmbito da Ciência Aberta. **Em Questão**, v. 25, n. 2, p. 389-412, 2019. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/84753/52667. Acesso em: 19 maio 2023.

KING, G. An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. **Sociological Methods & Research**, v. 36, n. 2, p. 173–199, nov. 2007. Disponível em: https://gking.harvard.edu/files/dvn.pdf. Acesso em: 19 maio 2023.

LATTESDATA. **Histórico**. CNPq, Brasília, 2022. Disponível em: https://lattesdata.cnpq.br/dvn/about/. Acesso em: 19 maio 2023.

SANTOS, L. O. B. S. *et al.* FAIR Data Point: a fair-oriented approach for metadata publication. **Data Intelligence**, v. 5, n.1, p. 163–183, 2023. Disponível em: https://direct.mit.edu/dint/article/5/1/163/112599/FAIR-Data-Point--A-FAIR-Oriented-Approach-for. Acesso em: 19 maio 2023.

UNIVERSITY OF CALIFORNIA SAN FRANCISCO (UCFS). Data Sharing & Data Management - Why Share Data?. UCSF, [202?]. Disponível em: https://www.library.ucsf.edu/research-and-data-services/data-sharing-data-management/why-share-data/. Acesso em: 13 maio 2023.

ZENG, M. L.; CHAN, L. M. Metadata interoperability and standardization: a study of methodology part I: achieving interoperability at the schema level. **D-Lib Magazine**, v. 12, n. 6, p. 1, 2006. Disponível em: https://www.dlib.org/dlib/june06/chan/06chan.html. Acesso em: 19 maio 2023.