Criação de um *pipeline* de dados abertos conectados para indicadores educacionais do PNE.

Abílio Nogueira Barros [1], Andreza Alencar[1],Aldéryck Félix de Albuquerque , Ibsen Mateus Bittencourt[2], Rafael Ferreira Mello[1]

[1]Departamento de Computação - Universidade Federal Rural de Pernambuco (UFRPE)

[2]Universidade Federal de Alagoas (UFAL)

{abilionbarros,derycck}@gmail.com, {andreza.leite,rafael.mello}@ufrpe.br, ibsen@feac.ufal.br

Resumo. *Este estudo apresenta o processo de pesquisa e design de um pipeline para a publicação de Dados Abertos Conectados (DAC), abordando desde a análise de ferramentas até a definição da estrutura para modelagem ontológica e disponibilização dos dados. A proposta foi validada por meio de um estudo de caso aplicado aos indicadores do Plano Nacional de Educação (PNE). Além disso, analisamos ferramentas que, embora não tenham sido incorporadas diretamente, podem ser utilizadas em variações desta solução conforme as necessidades do projeto e do domínio dos dados. Por fim, o estudo detalha a estrutura e a implementação do pipeline em operação, demonstrando sua aplicabilidade no serviço disponibilizado ao público.*

1. Introdução

O uso de dados abertos conectados (DAC) é fundamental para o desenvolvimento de soluções inovadoras. Essa abordagem permite a integração de diversos conjuntos de dados, ampliando as possibilidades de análise e descoberta de informações [Bandeira et al. 2015]. A conexão desses dados possibilita um panorama mais completo e abrangente, facilitando a identificação de padrões, tendências e relações que poderiam ser negligenciados em conjuntos de dados isolados.

O processo de disponibilização do DAC é complexo e desafiador. Este requer uma análise cuidadosa desde a coleta das informações até a forma de disponibilização e o incremento subsequente após sua exposição na rede. A dificuldade em integrar informações de diversas fontes e consolidá-las em um formato e local unificados constitui um obstáculo significativo na disponibilização desse tipo de dados.

As etapas para a publicação de DACs estão diretamente relacionadas às escolhas tecnológicas empregadas em seu desenvolvimento e, principalmente, em sua manutenção. Diante disso, é essencial estabelecer uma integração entre o conhecimento produzido por pesquisas ou sistemas previamente desenvolvidos no contexto de publicação de DACs. Muitas vezes, as ferramentas utilizadas são descontinuadas, seja em alguma etapa do processo ou em sua completude, resultando na finalização dos serviços de suporte e manutenção dessas ferramentas. Essa descontinuidade é um dos maiores obstáculos para a implementação prática de soluções de DAC.

A escolha das ferramentas de software para esse processo é particularmente desafiadora, pois a maioria delas, incluindo bibliotecas e *plugins*, são desenvolvidos de forma

open-source. Isso acarreta o risco de descontinuidade ou desatualização, o que pode afetar a compatibilidade entre essas ferramentas e aquelas já consolidadas no mercado. Por exemplo, atualizações e incompatibilidades entre estas ferramentas e sistemas de gerenciamento de bancos de dados amplamente utilizados no meio acadêmico e comercial podem dificultar a integração e a manutenção de um sistema de DAC eficiente.

2. Contextualização do problema

A construção da ontologia nesta pesquisa teve como ponto de partida a avaliação de ontologias educacionais existentes, conforme descrito em [Carneiro and Brito 2005]. Contudo, a estrutura de metas e indicadores analisada exigiu uma abordagem ampliada para um contexto analítico. Esse mapeamento ontológico de indicadores guarda semelhanças com abordagens adotadas na área da saúde, como discutido em [Ferronato et al. 2016]. Nesse cenário, a ferramenta Protégé destacou-se como a mais utilizada para modelagem e visualização de ontologias, sendo assim escolhida para essa etapa do estudo.

A montagem da pipeline também envolveu a busca por ferramentas adequadas para publicar Dados Abertos Conectados (DAC). Foram feitos testes e análises para escolher as melhores opções para converter dados relacionais em triplas RDF.

O levantamento das ferramentas foi guiado por estudos recentes que abordam a temática e detalham as soluções utilizadas. Com base nessas referências, foram selecionados componentes que possibilitam a construção de uma pipeline robusta e replicável. Além disso, tutoriais e materiais complementares disponíveis na web foram consultados para aprofundar o conhecimento sobre as ferramentas identificadas.

A escolha do PDI (*Pentaho Data Integration*) e ETL4LOD+ se deu com base no estudo de [Rodrigues and Maciel 2022] que descreve o processo de abertura de dados da UFMT, incluindo extração, transformação e conexão. A escolha desse trabalho se deu por sua proximidade com a pipeline idealizada. Os autores utilizaram o PDI com o plugin ETL4LOD+, que executa a transformação e disponibilização dos dados abertos, posteriormente publicados no CKAN. No entanto, a abordagem não contemplava a publicação de DAC, requisito essencial para este projeto.

A tentativa de replicar o processo enfrentou dificuldades, como a complexidade do ambiente e a falta de atualizações do plugin, descontinuado há mais de quatro anos. Como o CKAN não seria usado, sua configuração também foi descartada. Essas limitações levaram à busca por outro plugin compatível com o PDI.

3. Desenvolvimento da *pipeline*

Para a definição deste processo de publicação de DAC foi realizada uma pesquisa e levantamento de trabalhos que serviram como base para o entendimento das ferramentas que já foram utilizadas anteriormente por outros grupos de trabalho na mesma área. Com isso, realizamos um esboço dessa *pipeline* unindo o conhecimento adquirido tanto do entendimento do cenário básico que ela ira atender quanto as lições aprendidas descritas em outros projetos.

A Figura 1 apresenta o processo idealizado para o projeto. Antes de iniciar a busca pelas ferramentas a serem utilizadas, foram definidos três pilares fundamentais do processo de DAC, com base em [Rautenberg et al. 2019] e adaptados às necessidades

específicas do projeto. Esses pilares são: (i) criação da ontologia e processamento e exportação das triplas RDF; e (ii) disponibilização do endpoint SPARQL.

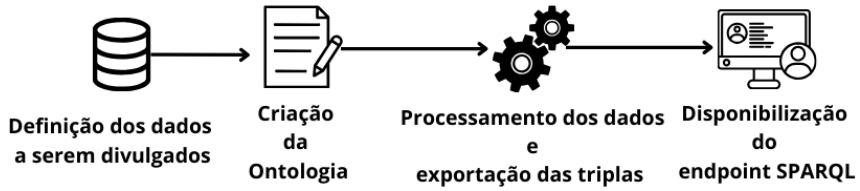

Figura 1. *Pipeline* proposta

Após os pilares definidos passamos as buscas de ferramentas que atuassem em uma ou mais etapas do processo de publicação de DAC.A seguir será apresentado o pipeline proposto e as atividades que foram necessárias para a definição de cada etapa do processo.

Foi realizada a etapa de definição dos dados a serem divulgados, a qual é independente de ferramentas e centrada em decisões de projeto. Nessa fase, são estabelecidas a origem e a governança dos dados, determinando quais serão abertos e conectados, considerando seu uso por usuários externos. Além disso, define-se o formato de acesso pela pipeline (CSV, Parquet ou banco de dados), em alinhamento com a equipe de infraestrutura, levando em conta a atualização periódica e as regras de governança.

3.0.1. Criação da ontologia

Em seguida é necessário a criação da ontologia, foi utilizada a ferramenta gratuita *Protégé*. O objetivo dessa ferramenta é criação, edição ou a visualização da ontologia criada, bem como sua definição de entidades e atributos, similar ao processo de elaboração de modelagem de tabelas relacionais para bancos de dados. Vale ressaltar que o foco deste estudo não é a criação formal de ontologias, mas sim a pipeline de publicação de DAC.

Concluída a ontologia, o próximo passo é a geração das triplas RDF (Resource Description Framework). O RDF, especificado pelo W3C, é um padrão para modelar e descrever recursos na web de maneira legível por humanos e interpretável por máquinas [Pan 2009].

No RDF, os dados são estruturados em triplas: sujeito, predicado e objeto. O sujeito representa o recurso principal, o predicado define a relação ou propriedade, e o objeto pode ser um valor literal ou outro recurso vinculado ao sujeito.

No contexto de DAC, o RDF é utilizado para integrar dados de diversas fontes, garantindo interoperabilidade e reutilização. Essencial para a Web Semântica, permite que máquinas interpretem e utilizem os dados de forma significativa. Definido esse formato para o *endpoint* SPARQL, o próximo passo é escolher uma ferramenta para executar a publicação dos dados.

Com esse objetivo, essa etapa da *pipeline* foi implementada utilizando a ferramenta *Pentaho Data Integration* junto ao *plugin Kettle Jena*.

O *Pentaho Data Integration* (PDI), também conhecido como *Kettle*, é uma ferra-

menta de ETL (*Extract, Transform, Load*) de código aberto que faz parte da suíte *Pentaho*, uma plataforma completa de *Business Intelligence* (BI). O PDI utiliza uma interface gráfica de usuário (*GUI*) chamada *Spoon*, que permite aos usuários projetar, testar e implantar processos ETL de maneira intuitiva e visual. Isso reduz a necessidade de codificação manual e facilita a criação e manutenção de fluxos de trabalho de dados.

Sendo uma ferramenta de código aberto, o PDI oferece flexibilidade e personalização, com uma comunidade ativa contribuindo com sua evolução e conexão com outras ferramentas. Um exemplo de ferramenta é o plugin do Jena, uma ferramenta desenvolvida em Java e já bem consolidada na temática do *DAC*. Esse plugin permite que usuários do *Kettle* trabalhem com dados RDF nativamente dentro de seus fluxos de trabalho de ETL, não sendo necessário o uso da ferramenta que tinha como requisito o domínio da linguagem Java. Uma vez que já se possua o PDI configurado e instalado na máquina basta que seja adicionado o *plugin* como descrito na página dos desenvolvedores na etapa *Build Steps* aqui.

As etapas implementadas na ferramenta PDI são:

1. *Data input*: Etapa é responsável pela inserção dos dados, fazendo com que estes sejam inseridos no fluxo da ferramenta para serem transformados em triplas RDF .
2. *Set URI Prefix* : Definição de qual será a URI. A definição de URIs na criação de uma ontologia é crucial para assegurar a identificação única, interoperabilidade e desambiguação dos recursos.
3. *Create Resources URI*: Seleção dos campos de dados inseridos são mapeados para os atributos definidos na ontologia para que assim possa refletir corretamento os tipos de cada dado presente nas triplas formadas.
4. *Create Jena Model*: Criação do modelo Jena que será responsável por unir todas as informações anteriores e realizar o mapeamento entre os dados inseridos e as ontologias utilizadas para realizar a estruturação das triplas RDF.
5. Serialize Jena Model: Por fim, a última etapa é responsável por compilar as informações anteriores, gerando assim todas as triplas RDF's mapeadas pelo modelo e gerando o arquivo com as triplas prontas para serem carregadas.

3.0.2. Disponibilização do *endpoint SPARQL*

Endpoint SPARQL é um serviço que permite a execução de consultas SPARQL (*SPARQL Protocol and RDF Query Language*) sobre um conjunto de dados RDF (*Resource Description Framework*) acessível via HTTP. No contexto de DAC, um endpoint SPARQL é uma interface pública que possibilita a consulta de dados abertos de forma padronizada e interligada, promovendo a transparência, reutilização e interoperabilidade dos dados.

A ferramenta escolhida para esse serviço foi o *OpenLink Virtuoso*, uma plataforma de servidor de banco de dados robusta para armazenamento e consulta de dados RDF e SPARQL. Amplamente utilizado para implementar endpoints SPARQL devido a suas características avançadas e flexibilidade, o Virtuoso é uma solução consolidada adotada por diversos entes públicos como o governo da Espanha aqui, a Biblioteca Nacional do Chile aqui, e a *DBPEDIA*, um dos principais portais mundiais de DAC.

4. Estudo de caso: Metas do PNE

Definir o estudo de caso em um contexto real permite capturar desafios que um ambiente controlado não abarcaria. Segundo [Kitchenham et al. 1995], essa abordagem possibilita uma avaliação mais precisa do método ou ferramenta, considerando interações com outros sistemas, fatores organizacionais e restrições do ambiente.

Neste projeto, aplicamos a *pipeline* na construção da plataforma ConectaPNE, que está sendo desenvolvida em colaboração com o Ministério da Educação (MEC) para apoio ao monitoramento do Plano Nacional de Educação (PNE). O PNE brasileiro tem a missão de guiar o desenvolvimento educacional do país ao longo de 10 anos [Macena et al. 2018]. Ele define metas, estratégias e ações para aprimorar a qualidade da educação em todas as suas etapas e modalidades, incluindo a formação de professores e a gestão democrática das escolas. O monitoramento contínuo do PNE é essencial para garantir que as metas sejam cumpridas e possibilitar ajustes estratégicos conforme necessário [Vinente and Duarte 2015].

Diante disso, o objetivo da abertura dos dados utilizados na plataforma ConectaPNE é disponibilizar as informações sobre indicadores do PNE ao nível municipal para que estes possam ser acessados diretamente no formato de DAC.

Vale ressaltar que a *pipeline* proposta neste trabalho pode se adequar a qualquer domínio de dados onde exista a viabilidade da abertura dos dados e não somente ao contexto deste estudo de caso. A seguir, passaremos por todas as etapas da *pipeline* anterior para demonstrar o funcionamento dela no fluxo na plataforma.

4.1. Estrutura prévia dos dados

A figura 2 fornece um visão geral de como está estruturada a etapa de ETL (extração, transformação e carga) do projeto, citando as ferramentas utilizadas neste processo. Os dados em sua maioria são oriundos de fontes públicas com atualização anual. O detalhamento completo do processo de ETL e da arquitetura de dados pode ser encontrado nos seguintes trabalhos [Barros et al. 2023],[Barros et al. 2022].

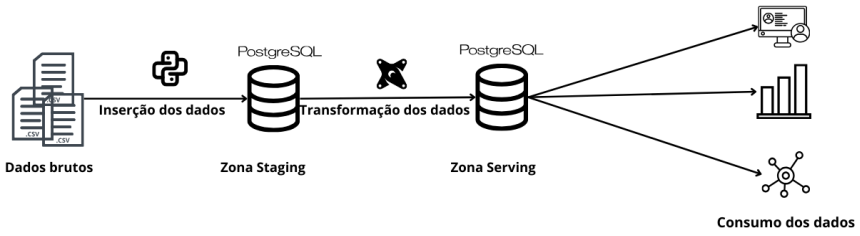

Figura 2. *Endpoint SPARQL.*

4.1.1. Seleção dos dados para publicação e criação da ontologia

Na primeira etapa da *pipeline* realizamos a definição dos dados a serem disponibilizados. Neste estudo de caso escolhemos os indicadores das três primeiras metas do PNE, abordando assim os dados relacionados a 5 indicadores que serão publicados no formato de DAC.

Por decisão de projeto, o formato de disponibilização desses dados foi feito via arquivos CSV, devido à periodicidade de atualização ser baixa, apenas anualmente. Assim, não se fazia necessário a conexão da *pipeline* diretamente ao banco de dados, o que poderia onerar ainda mais o processo, visto que um código em *Python* pode ser ativado manualmente a cada atualização anual desses resultados e gerar os arquivos necessários.

Para o desenvolvimento da ontologia, foram criadas as entidades e posteriormente adicionados seus atributos e relacionamentos. Para esta etapa foi utilizada a ferramenta gratuita *Protégé*.

4.2. Processamento dos dados e exportação das triplas RDF

Para esta etapa a abordagem escolhida foi carregar os dados direto da nossa base de dados e modelando a estrutura de dados do *python* como dicionários e *dataframes*. Esse carregamento deve ser feito por funções *python* para que assim possa ser atualizado sempre que necessário. A criação dessa função irá permitir a adição de novas metas e indicadores.

4.2.1. *Input* dos dados

A primeira etapa é realizar a inserção dos dados, a depender da fonte escolhida pelo projeto, que neste estudo de caso é *Comma-Separated Values*(CSV). Nessa etapa é possível fazer a seleção dos campos que devem ser lidos, visto que a base pode ter outros elementos que não devem ser carregados.

4.2.2. Definição da URI e Concatenação dos campos e Criação do Modelo *Jena*

Nessa etapa é adicionada a URI que se tornarão a base dos elementos, criando o elemento *resource_uri* para realizar a geração de triplas RDF com base nas URIs concatenadas. Após isso é possível realizar uma última avaliação do tipo das colunas e adicionar a *resource_uri* como *target* nesse processamento. Essa penúltima etapa envolve a inicialização do modelo responsável pela criação das triplas. Nesta fase, os campos são preenchidos com as ontologias necessárias e, para cada campo, são atribuídas as propriedades de cada coluna, bem como seu tipo conforme as ontologias informadas. Por fim, é possível identificar erros ou exceções dependendo da completude dos dados e dos campos principais que devem ser criados, como ilustrado na Figura 3.

4.2.3. Serialização do modelo Jena

Para a última etapa, a serialização do modelo *Jena*. Serialização é o processo de transformar objetos ou estruturas de dados em um formato que pode ser facilmente armazenado ou transmitido e posteriormente reconstruído. Nesta fase, os objetos *Jena* criados na etapa anterior são transformados em triplas RDF e exportadas em arquivos *TURTLE*.

4.3. Disponibilização dos dados

Assim, nesta etapa as triplas RDF's, geradas em arquivos da extensão.ttl, são carregadas na ferramenta *Virtuoso* via interface web denominada Conductor, como podemos ver na Figura 4.

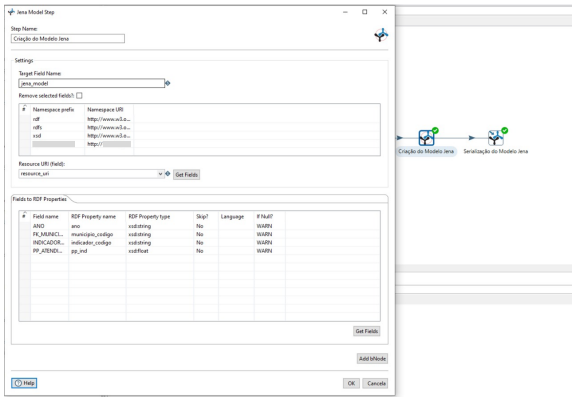

Figura 3. Criação do modelo Jena

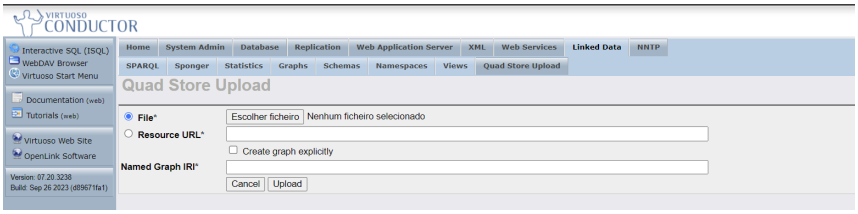

Figura 4. Visão do Conductor para inserção das triplas RDF's.

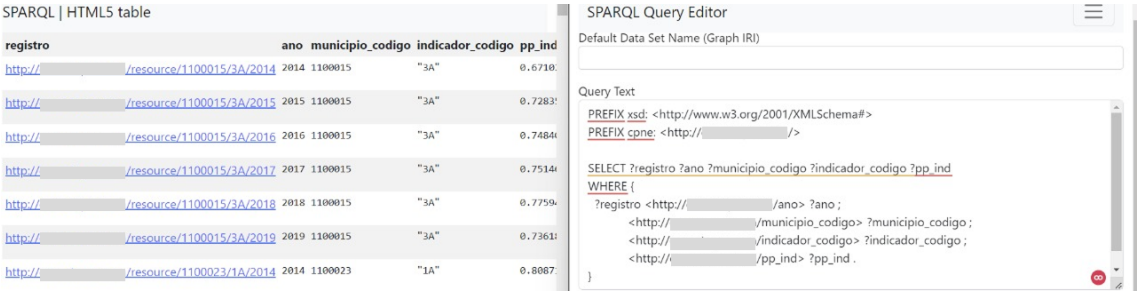

Figura 5. Resultado da consulta SPARQL e a consulta realizada ao endpoint SPARQL.

Uma vez que as triplas são carregadas, é disponibilizado a realização das consultas no ambiente *SPARQL*, como ilustrado na figura 5. Atualmente o serviço já se encontra ativo e pode ser acessado em https://conectapne.nees.ufal.br/sparql/.

5. Conclusão e trabalhos futuros

Este estudo de caso propõe um pipeline para a publicação de Dados Abertos Conectados (DAC) aplicada aos indicadores do PNE fornecendo acesso a seus indicadores em formato de consulta SPARQL. A principal contribuição foi a conversão de dados relacionais para triplas RDF de forma semi-automática. Além de atender ao escopo do projeto, o trabalho amplia o conhecimento sobre DAC na área de educação pública.

A divulgação de uma pipeline estável e escalável é fundamental para garantir a manutenção contínua da disponibilização dos dados, especialmente no contexto de dados abertos e dados abertos conectados, onde a sustentabilidade e a confiabilidade dos processos são aspectos cruciais.

Para trabalhos futuros prevemos a realização de melhorias na *pipeline* como a concatenação de modelos, a otimização da disponibilização das triplas e a melhoria da ontologia. Também aspiramos submeter as ontologias a portais específicos de âmbito mundial, como o LOV Linked Dataset, que reúne diversas ontologias globalmente.

Referências

[Bandeira et al. 2015] Bandeira, J. M., Alcantara, W., Barbosa Sobrinho, A., Ávila, T. J. T., Bittencourt, I., and Isotani, S. (2015). Dados abertos conectados. *III Simpósio Brasileiro de Tecnologia da Informação*.

[Barros et al. 2023] Barros, A., Albuquerque, A., Alencar, A., Mello, R., Alves, G., and Bittencourt, I. (2023). Arquitetura de dados educacionais como plataforma para governo inteligente - utilizando dados abertos para apoio à gestão educacional baseada em evidências. In *Anais do XI Workshop de Computação Aplicada em Governo Eletrônico*, pages 130–140, Porto Alegre, RS, Brasil. SBC.

[Barros et al. 2022] Barros, A., Alencar, A., Nascimento, A., Albuquerque, A., and Mello, R. (2022). Elaboração do conjunto de dados agregados do censo da educação básica. In *Anais do IV Dataset Showcase Workshop*, pages 35–45, Porto Alegre, RS, Brasil. SBC.

[Carneiro and Brito 2005] Carneiro, R. E. and Brito, P. d. (2005). Definição de uma ontologia em owl para representação de conteúdos educacionais. *VII ENCONTRO DE ESTUDANTES DE INFORMÁTICA DO ESTADO DO TOCANTINS. Centro Universitário Luterano de Palmas (CEULP/ULBRA)*.

[Ferronato et al. 2016] Ferronato, A. C. C., Pires, F. R., and Bernardini, F. C. (2016). Um modelo para integração e disponibilização de dados na área de saúde governamental. In *Anais do XII Simpósio Brasileiro de Sistemas de Informaçao*, pages 124–127. SBC.

[Kitchenham et al. 1995] Kitchenham, B., Pickard, L., and Pfleeger, S. L. (1995). Case studies for method and tool evaluation. *IEEE software*, 12(4):52–62.

[Macena et al. 2018] Macena, J. d. O., Justino, L. R. P., and Capellini, V. L. M. F. (2018). O plano nacional de educação 2014–2024 e os desafios para a educação especial na perspectiva de uma cultura inclusiva. *Ensaio: Avaliação e Políticas Públicas em Educação*, 26:1283–1302.

[Pan 2009] Pan, J. Z. (2009). Resource description framework. In *Handbook on ontologies*, pages 71–90. Springer.

[Rautenberg et al. 2019] Rautenberg, S., de Souza, L., Dall'Agnol, J. M. H., and Michelon, G. A. (2019). *Guia prático para publicação de dados abertos conectados na web*. Appris Editora e Livraria Eireli-ME.

[Rodrigues and Maciel 2022] Rodrigues, F. A. and Maciel, C. (2022). Um método para captura e compartilhamento de dados abertos educacionais via um processo etl. In *Anais do X Workshop de Computação Aplicada em Governo Eletrônico*, pages 133–144. SBC.

[Vinente and Duarte 2015] Vinente, S. and Duarte, M. (2015). O plano nacional de educação (2014-2024) e a garantia de um sistema educacional inclusivo: possibilidade ou utopia. *Olhares: Revista do Departamento de Educação da Unifesp*, 3(2):133–151.