Um Modelo para Integração de Dados Baseado em Ontologias com Foco em Análises Exploratórias: Uma Aplicação Prática em Dados de Manutenção de Locomotivas Diesel-elétricas

Pedro Paulo Rezende Silva Domingos and José Maria Parente de Oliveira

[1]*Instituto Tecnológico de Aeronáutica (ITA), São José dos Campos, SP, Brasil*

Abstract

The use of data for decision making has been increasingly essential in organizations. However, for these applications to occur efficiently, it is of paramount importance that descriptive and exploratory analyzes be carried out on the available data. The use of ontologies for this initial phase may be relevant considering the whole semantic concept involved. However, a solution that accesses data sources of different formats is necessary for this integration to occur in a simplified way. This article proposes a model where a *framework* in *Python* language directly accesses data sources from different *file systems* and makes the data available in triple format within a previously elaborated ontology, with practical applications in the railway sector, in the field of maintenance data for diesel-electric locomotives.

Resumo

A utilização de dados para tomadas de decisão tem sido cada vez mais primordial nas organizações. Porém, para que essas aplicações ocorram de forma eficiente, é de suma importância que análises descritivas e exploratórias sejam feitas nos dados disponíveis. O uso de ontologias para essa fase inicial pode ser relevante considerando-se todo o conceito semântico envolvido. Contudo, torna-se necessária uma solução que acesse fontes de dados de diferentes formatos para que essa integração ocorra de uma forma simplificada. Este artigo propõe um modelo onde um *framework* em linguagem *Python* acessa diretamente fontes de dados de *file systems* diferentes e disponibiliza os dados em formato de triplas dentro de uma ontologia previamente elaborada, com aplicações práticas no setor ferroviário, no domínido de dados de manutenção de locomotivas diesel-elétricas.

Keywords

framework, ontology, python, data access, file systems

1. Introdução

O uso eficiente dos dados disponíveis tem sido essencial para o sucesso das organizações. O foco está mudando de obter informações para encontrar as informações corretas [1]. As organizações contemporâneas estão cada vez mais reconhecendo a importância de analisar como seus processos de negócio são conduzidos, visando garantia de qualidade, otimização e

Proceedings of the XVI Seminar on Ontology Research in Brazil (ONTOBRAS 2023) and VII Doctoral and Masters Consortium on Ontologies (WTDO 2023), Brasilia, Brazil, August 28 - September 01, 2023.

domingos@ita.br (P. P. R. S. Domingos); parente@ita.br (Prof. Dr. J. M. P. d. Oliveira)

melhoria contínua [2]. Para tais situações, torna-se importante representar fielmente a realidade do negócio aos dados que estão sendo analisados para se obterem *insights* concretos.

Nesse contexto, o uso de ontologias se demonstra relevante, com a finalidade de descrever o domínio de atuação, implicando um contexto semântico aos dados disponíveis. Uma ontologia é uma representação semântica de conceitos relacionados e seus relacionamentos, revelando a essência desses conceitos e as ligações entre eles [3].

Contudo, as ferramentas disponíveis para acesso aos dados por meio de uma ontologia dependem que os dados estejam estruturados em bancos de dados ou já no formato de triplas, ou que as fontes sejam convertidas para esses formatos por meio de bases de dados federadas. Existem atualmente soluções que atuam também em formatos não-estruturados como *NoSQL* ou dados tabulares como arquivos CSV, porém não contemplam outros formatos de arquivo como *Parquet*, por exemplo, ou formatos específicos a alguns dispositivos industriais. Logo, há a necessidade inicial de uma estruturação ou adequação das fontes de dados para que esta integração seja realizada.

Em cenários onde as fontes de dados de uma organização ainda se encontram em fase de estruturação, torna-se relevante a aplicação de análises exploratórias para identificar o real valor dos dados disponíveis. Entende-se que as ontologias podem contribuir nessa etapa por acrescentar valor semântico ao conjunto de dados, auxiliando na identificação de atributos que podem trazer valor ao negócio. Porém adequar os dados para a realização dessas análises pode resultar em um esforço considerável considerando-se ser esta uma fase de muitas incertezas quanto à eficácia destes.

Logo, tem-se a necessidade de se desenvolver uma solução de integração de dados no formato de ontologia, acessando diretamente estes diferentes *file systems* por meio de um mapeamento e possibilitando a realização de consultas nos dados disponíveis. De uma certa forma, uma implementação como essa pode se traduzir em uma forma simplificada de acessar dados por meio de ontologias, principalmente em casos de análises exploratórias.

O presente artigo propõe a elaboração de um *framework* desenvolvido em *Python* que acesse diretamente fontes de dados de diferentes formatos e as integre por meio de uma ontologia e de um arquivo de mapeamento. As experimentações serão realizadas no cenário industrial, em uma empresa do setor ferroviário, com foco em dados de manutenção de locomotivas diesel-elétricas. Atualmente essa empresa se encontra no estágio inicial de alavancar seus processos com o uso de dados. Nesta organização, utiliza-se parte dos dados disponíveis há alguns anos para tomadas de decisão, porém por meio de muito trabalho repetitivo e com o uso constante de planilhas. Com isso, muitas dessas fontes de dados não são integradas, por conta do pouco conhecimento do domínio com um todo. Há também um desconhecimento quanto à real viabilidade que essa integração traria, já que não se sabem os reais ganhos dessa ação.

Tem-se então a necessidade de se realizarem análises descritivas e exploratórias em um grande volume de dados em fontes distintas com *file systems* diferentes para, a partir daí, ter-se um processo de engenharia de dados para a estruturação dos dados e um mapeamento inicial de possíveis implementações. O uso de ontologias torna-se relevante para guiar esse fluxo de estruturação dos dados, já que toda a carga semântica que pode ser aplicada pode contribuir significativamente para o melhor entendimento dos dados.

2. Objetivo

A pesquisa tem como objetivo principal desenvolver e validar um modelo de integração de dados, utilizando-se ontologias para prover o conceito semântico do domínio e um *framework* para acesso direto à fontes de dados de diferentes *file systems*, com aplicação prática à dados de manutenção de locomotivas diesel-elétricas. Têm-se como objetivos específicos:

- Levantar os dados disponíveis na empresa em estudo por meio de um mapa conceitual, para o posterior desenvolvimento de uma ontologia para a integração dos dados;
- Elaborar ontologias que representem o domínio de manutenção de locomotivas diesel-elétricas, envolvendo dados de utilização dos ativos e de seus componentes, e parâmetros de funcionamento do motor diesel e dos itens elétricos;
- Desenvolver um *framework* que possibilite o uso de uma ontologia definida em OWL e, por meio de um mapeamento em formato JSON, acesse as fontes de dados, realize consultas e permita o uso de raciocínio para inferências nos dados disponíveis;
- Validar o modelo de integração de dados desenvolvido, identificar possíveis melhorias e descrever os próximos passos do estudo.

3. Domínio de dados de manutenção de locomotivas diesel-elétricas

A competitividade e a capacidade de reação das indústrias dependem consideravelmente da eficácia com que mantêm seus ativos e garantem a disponibilidade de seus recursos [4]. Nesse sentido, processos de manutenção preditiva e de manutenção baseada na condição por meio do uso dos dados ampliam o conhecimento dos ativos e a capacidade de atuação destes.

Além de ser vantajosa em relação aos demais tipos quanto ao custo, a manutenção preditiva direciona a empresa para o conceito de indústria 4.0 [5]. O conhecimento e os dados coletados de *inputs* humanos, máquinas, meio ambiente e de processos também podem auxiliar nas tomadas de decisão a fim de reduzir o custo operacional e melhorar a qualidade de produtos ou serviços [6].

O ambiente ferroviário apresenta uma ampla diversidade em relação aos dados, sendo esses provenientes de sistemas gerenciais com dados de manutenção, ciclo de vida de componentes, informações técnicas e registradores de microprocessadores controladores. Conta também com diferentes tipos de ativos e grupos destes, além da complexidade de muitos não possuírem uma localização fixa. Locomotivas e vagões, por exemplo, ficam em constante operação ao longo da via férrea, sendo afetados inclusive por diferentes condições que o ambiente se encontram.

No caso de uma locomotiva, a complexidade é ainda maior, devido à sua abrangência de especialidades técnicas e à quantidade de painéis eletrônicos e sensores instalados. Trata-se de um ativo multivariável, com um sistema fortemente interligado e não linear[7].

Na empresa em questão, a diversidade de fontes de dados e sua descentralização implicam em um cenário de maior complexidade. Tratam-se de arquivos de extensão *Parquet*, CSV, XLSX e *file systems* específicos de alguns painéis controladores de locomotivas, disponíveis em diversos locais da organização. Por consequência, para uma definitiva integração de dados tem-se a

necessidade de avaliar o real valor que essas fontes trazem, evitando-se assim retrabalhos e custos desnecessários com armazenamento de dados em serviços como, por exemplo, um *data lake*.

4. Integração de dados por meio de ontologias e acesso direto às fontes de dados

Os possíveis usos das representações ontológicas do domínio da manufatura e industrial não se limitam às aplicações em arquiteturas de controle, mas também podem apoiar projeto, simulação, planejamento e programação, avaliação de desempenho e integração de dados em campo [8]. O uso de ontologias pode facilitar a integração de dados de várias maneiras e ir além das abordagens tradicionais de uso de elementos e modelos de dados comuns.

Existem algumas formas de acessar os dados, sendo um deles o sistema *Ontology-Based Data Access* (OBDA). No OBDA, uma camada conceitual é fornecida na forma de uma ontologia que define um vocabulário compartilhado, modela o domínio, oculta a estrutura das fontes de dados e enriquece dados incompletos com conhecimento prévio. Em seguida, as consultas são inseridas sobre essa visão conceitual de alto nível, retirando dos usuários a necessidade de entendimento das fontes de dados, a relação entre elas ou a codificação dos dados. As consultas são traduzidas pelo sistema OBDA em consultas sobre fontes de dados potencialmente muito grandes (geralmente relacionais e federadas) [1].

Uma das ferramentas OBDA utilizadas é o *Ontop*. Este é um sistema OBDA de código aberto lançado sob a licença *Apache*, desenvolvido na Universidade Livre de Bozen-Bolzano. O sistema *Ontop* expõe bancos de dados relacionais como grafos RDF virtuais, vinculando os termos (classes e propriedades) da ontologia às fontes de dados por meio de mapeamentos. Esse gráfico RDF virtual pode então ser consultado usando SPARQL traduzindo as consultas SPARQL em consultas SQL nos bancos de dados relacionais. Este processo de tradução é transparente para o usuário [1].

Existem outras soluções como o R2RML [9] e o xR2RML[10] que atuam com o mapeamento de alguns tipos de arquivos como CSV, XML e até mesmo bancos *NoSQL*, porém não contemplando tipos de arquivos que têm sido utilizados atualmente como o *Parquet* por exemplo.

Na linguagem *Python* existe a biblioteca *Owlready2* que permite o carregamento de ontologias salvas em OWL, além da execução de *queries* em *SPARQL* e raciocínio por meio do *HermiT*. Os dados também devem estar em formato de triplas ou em bancos de dados para a execução das consultas.

Ambas as soluções precisam que os dados estejam em bancos de dados ou formato de triplas para serem acessados. No caso do *Ontop* e do seu uso no *Protegé*, as fontes de dados podem ser acessadas por meio de banco de dados federados, utilizando ferramentas e *plugins* de terceiros para realizar essa conversão. No *Owlready* é possível popular a ontologia carregada com triplas RDF para acessar os dados, porém não existe uma estrutura de mapeamento que realize esse processo. Contudo, o *Owlready* é uma biblioteca da linguagem *Python*, o que facilita o desenvolvimento de códigos que permitam o acesso direto às fontes de dados e as utilize para popular a ontologia.

Dessa forma, vê-se a oportunidade de se desenvolver um *framework* que realize esse acesso

direto à diferentes *file systems*, realize um mapeamento dos dados à ontologia utilizada, e os trate em um formato único, para que, em memória de execução, os dados possam ser acessados.

5. Modelo proposto

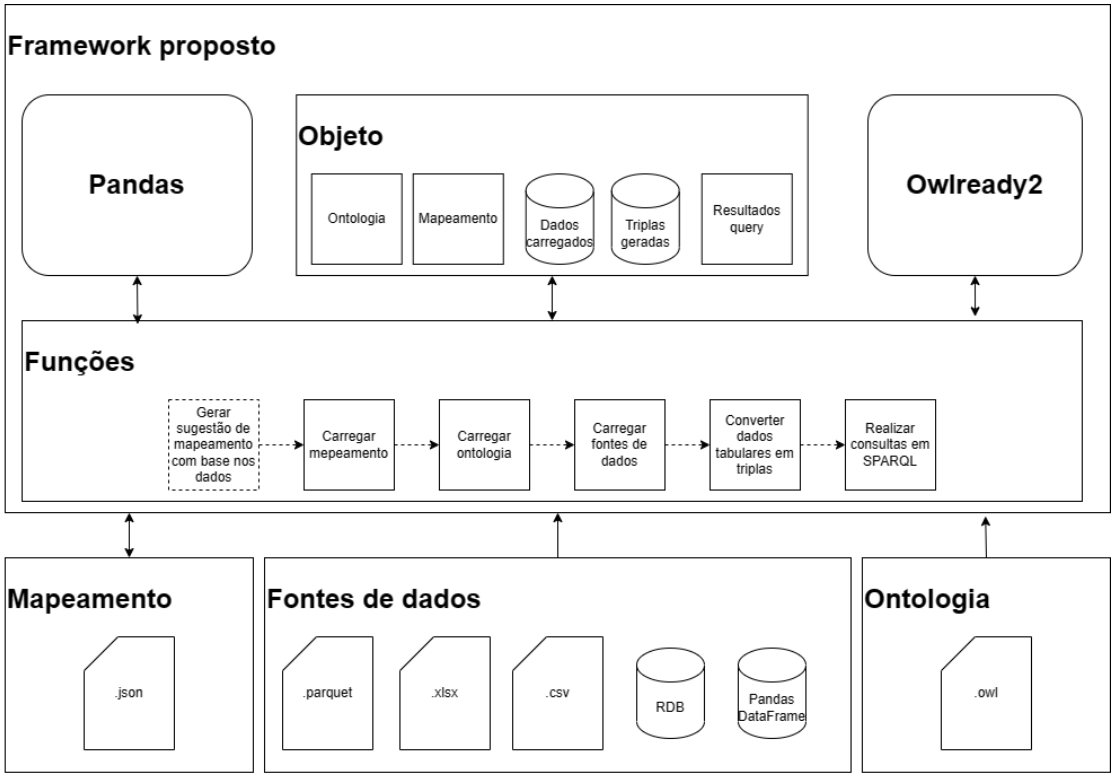

Figure 1: Arquitetura do *framework* proposto

O *framework* proposto (Figura 1) se baseia na criação de um objeto que permita o carregamento de uma ontologia no formato OWL e um arquivo de mapeamento de extensão JSON. Por meio destes, o código deve identificar os *file systems* correspondentes às fontes de dados e abri-las no formato de um *dataframe Pandas* [11], e converter os dados para triplas referentes aos mapeamentos elaborados. Dessa forma os dados necessários para realizar as consultas e as inferências ficam disponíveis em memória para realizar as análises.

Listing 1: Exemplo de arquivo de mapeamento em JSON

```
{
    "base_ocorrencias" :
    {
        "data_source_path":
        "C:\\Users\\Teste\\ontologia_locomotivas\\avarias.csv",
```

```
        "triples":
        [
            {
                "subject":
                {
                    "data_source_attribute_name": "N mero de S rie",
                    "ontology_subject_name": "Locomotiva"
                },
                "predicates_and_objects":
                [
                    {
                        "data_source_attribute_name": "Modelo",
                        "ontology_predicate_name": "temModelo",
                        "ontology_object_name": "Modelo_Locomotiva"
                    },
                    {
                        "data_source_attribute_name": "Grupo",
                        "ontology_predicate_name": "pertenceAoGrupo",
                        "ontology_object_name": "Grupo_Locomotiva"
                    }
                ]
            },
        ]
    }

    "base_logs" :
    {
        "data_source_path":
        "C:\\Users\\Teste\\ontologia_locomotivas\\logs_locomotivas.parquet
            ",

        "triples":
        [
            {
                "subject":
                {
                    "data_source_attribute_name": "Data",
                    "ontology_subject_name": "Timestamp_log_falha"
                },
                "predicates_and_objects":
                [
                    {
                        "data_source_attribute_name": "Log",
```

```
                        "ontology_predicate_name": "registrouLog",
                        "ontology_object_name": "Identificador_log"
                    },
                    {
                        "data_source_attribute_name": "Locomotiva",
                        "ontology_predicate_name": "ocorridoNaLocomotiva",
                        "ontology_object_name": "Locomotiva"
                    }
                ]
            },
        ]
    }
}
```

O arquivo de mapeamento direciona a execução do *framework* quanto ao caminho dos arquivos e as relações entre os nomes dos atributos das fontes de dados com as classes e propriedades presentes na ontologia criada (Listing 1). A solução também contará com uma função que gera uma sugestão de mapeamento em JSON com base nos dados carregados, gerando a estrutura do arquivo com os atributos identificados, bastando ao usuário inserir os identificadores das classes e propriedades.

Para a aplicação prática do *framework* será necessária também a elaboração de uma ontologia no domínio de manutenção de locomotivas diesel-elétricas. O foco das consultas a serem realizadas estará voltado à identificação da vida útil de componentes instalados nos ativos, tanto com base na utilização quanto no tempo em operação, tendo como referência as datas de instalação de componentes e os parâmetros coletados.

Pretende-se também realizar comparativos entre o *framework* proposto e demais soluções disponíveis, a fim de validar os ganhos com a pesquisa, tanto em tempo de execução e esforços necessários.

6. Conclusão

O *framework* proposto e a ontologia referente aos dados de locomotivas estão em fase de desenvolvimento inicial. Contudo, as aplicações iniciais têm se mostrado eficientes, possibilitando conversões assertivas e uma boa performance em consultas *SPARQL* às fontes de dados. Além disso, o *framework* e a estrutura inicial dos arquivos de mapeamento têm demonstrado um relevante dinamismo, permitindo modificações sem a necessidade de alterações nas fontes de dados.

A expectativa é que os desenvolvimentos sejam concluídos no próprio ano de 2023, e que todo o modelo proposto seja disponibilizado na plataforma *Github* para utilização em futuras pesquisas, aplicações e melhorias.

Referências bibliográficas

[1] D. Calvanese, B. Cogrel, S. Komla-Ebri, R. Kontchakov, D. Lanti, M. Rezk, M. Rodriguez-Muro, G. Xiao, Ontop: Answering sparql queries over relational databases, Semantic Web 8 (2017) 471–487.

[2] D. Calvanese, T. E. Kalayci, M. Montali, S. Tinella, Ontology-based data access for extracting event logs from legacy data: the onprom tool and methodology, in: International Conference on Business Information Systems, Springer, 2017, pp. 220–236.

[3] J. Wan, B. Chen, M. Imran, F. Tao, D. Li, C. Liu, S. Ahmad, Toward dynamic resources management for iot-based manufacturing, IEEE Communications Magazine 56 (2018) 52–59.

[4] M. H. Karray, F. Ameri, M. Hodkiewicz, T. Louge, Romain: Towards a bfo compliant reference ontology for industrial maintenance, Applied Ontology 14 (2019) 155–177.

[5] L. Baldissarelli, E. Fabro, Manutenção preditiva na indústria 4.0, Scientia cum industria 7 (2019) 12–22.

[6] J. Yan, Y. Meng, L. Lu, L. Li, Industrial big data in an industry 4.0 environment: Challenges, schemes, and applications for predictive maintenance, IEEE Access 5 (2017) 23484–23491.

[7] X. Yu, Z. Zhao, X. Zhang, C. Sun, B. Gong, R. Yan, X. Chen, Conditional adversarial domain adaptation with discrimination embedding for locomotive fault diagnosis, IEEE Transactions on Instrumentation and Measurement 70 (2020) 1–12.

[8] E. Negri, L. Fumagalli, M. Garetti, L. Tanca, Requirements and languages for the semantic representation of manufacturing systems, Computers in Industry 81 (2016) 55–66.

[9] C. Debruyne, D. O'Sullivan, R2rml-f: Towards sharing and executing domain logic in r2rml mappings., LDOW@ WWW 1593 (2016).

[10] F. Michel, L. Djimenou, C. F. Zucker, J. Montagnat, Translation of relational and non-relational databases into rdf with xr2rml, in: 11th International Confenrence on Web Information Systems and Technologies (WEBIST'15), 2015, pp. 443–454.

[11] T. pandas development team, pandas-dev/pandas: Pandas, 2023. URL: https://doi.org/10.5281/zenodo.8092754. doi:10.5281/zenodo.8092754, if you use this software, please cite it as below.