Utilização de Dicionário Semânticos de Dados para cobertura vacinal da COVID-19

Leonardo Mageste da Cruz Herédia, Evaldo de Oliveira da Silva

Curso de Bacharelado em Engenharia de Software

UniAcademia – Campus Academia

36016-000 – Juiz de Fora, MG – Brasil

lheredia1195@gmail.com, evaldo.oliveira@gmail.com

Resumo. A conceituação sobre dados da área da saúde é um desafio tendo em vista a grande necessidade de consumo humano destes dados para diferentes objetivos, entre eles, estudos científicos na área médica ou a recuperação da informação por meio de grafos de conhecimento. Atualmente as estruturas de dados utilizadas não são livres de *schemas*, ou seja, dependem do formato dos repositórios de dados que as armazenam, e não permitem que seus objetos e propriedades sejam anotados semanticamente. A anotação semântica utiliza modelos conceituais que elevam o entendimento sobre os dados com base em ontologias. Com melhor entendimento sobre os dados, pesquisadores tem em mãos a possibilidade de conduzir estudos mais detalhados e precisos. Neste trabalho, é proposto que anotar semanticamente dados pode auxiliar nesse sentido. Para tal, é apresentado um estudo de caso com dados sobre a cobertura vacinal da COVID-19. Nele são gerados grafos de conhecimento com dados modelados através de um processo sistemático com base na técnica SDD no domínio aqui contextualizado.

Abstract. The conceptualization of data in the health area is a challenge in consideration of the great need for human consumption of these data for different purposes, including scientific studies in the medical field or the retrieval of information through knowledge graphs. Currently, the data structures used are not schema-free, that is, they depend on the format of the data repositories that store them, and do not allow their objects and properties to be semantically annotated. Semantic annotation uses conceptual models that enhance the understanding of data based on ontologies. With a better understanding of the data, researchers have at hand the possibility of conducting more detailed and accurate studies. In this work, it is proposed that semantically annotating data can help in this regard. To this end, a case study is presented with data on the vaccination coverage of COVID-19. It generates knowledge graphs with modeled data through a systematic process based on the SDD technique in the domain contextualized here.

1. INTRODUÇÃO

Na era da informação dados passaram a ter grande valor, para diferentes áreas de atuação, tanto para área empresarial, acadêmica e científica. Cada vez mais torna-se necessário disponibilizar grandes volumes de dados de forma organizada.

A gestão da informação é uma área que contribui para a organização dos dados e informações. Propõe o planejamento das necessidades de informação, técnicas de descrição de dados com metadados que favorecem a qualidade, a preservação e facilitam a descoberta de novas informações para o reúso de dados e conhecimento (MEDEIROS, 2018; MARCHIORI, 2002). A área de *eScience* propõe técnicas para manipular grandes volumes de dados, ou métodos computacionais sofisticados e computação de alto desempenho. A pesquisa em *eScience* abrange diferentes etapas de um processo de pesquisa, tais como, a criação de ferramentas computacionais, coleta e análise de dados, modelagem, e o reúso dos resultados da pesquisa. Além disso, pressupõe trabalho conjunto e multidisciplinar, em que cientistas da Informação e de Computação auxiliem pesquisadores de outras áreas a desenvolver pesquisas de forma mais rápida e eficiente (FAPESP, 2021).

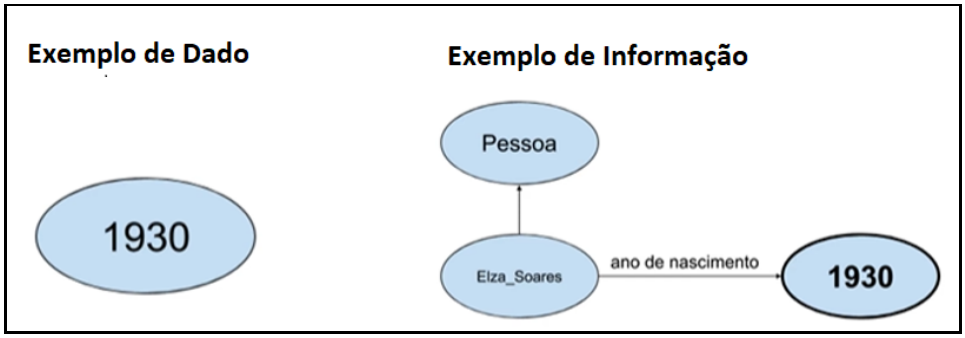
Este trabalho está contextualizado na área de *eScience* e utiliza como estudo de caso, dados sobre cobertura vacinal da COVID-19. Descreve a necessidade dos dados estarem preparados e organizados. A preparação de dados é uma das fases iniciais e propõe que os dados estejam modelados de forma a evitar interpretações equivocadas dos diferentes termos sobre um determinado domínio de aplicação. Os dados sobre a cobertura vacinal foram extraídos do site do Ministério da Saúde. No entanto, observou-se que as estruturas dos dados utilizados não são livres de *schemas*, e dependem do formato dos repositórios de dados que as armazenam, ou ainda não permitem que seus objetos e propriedade estejam documentados ou anotados com conceitos sobre o domínio.

A anotação semântica utiliza modelos conceituais que elevam o entendimento sobre os dados com base em ontologias. A conceituação sobre dados da área da saúde é um desafio tendo em vista a grande necessidade de consumo humano destes dados para diferentes objetivos, entre eles, estudos científicos na área médica ou a recuperação da informação por meio de grafos de conhecimento (GC). A conceituação propõe o uso de ontologias como mecanismo de representação do conhecimento sobre os dados utilizados.

No entanto, somente o uso de ontologias não é o suficiente para preparar e organizar os dados. Ontologias possuem uma estrutura taxonômica que classifica os conceitos sobre o domínio, mas que depende de técnicas de anotação semântica para que os dados estejam associados aos conceitos.

Ao analisar os diversos conhecimentos disponibilizados na web, percebe-se diferentes formatos de dados tais como arquivos .xml (*eXtensible Markup Language*), visualizadores .pptx (*Power Point*), planilhas .xlsx (*Excel*), documentos .docx (*Word*), arquivos de texto separados por vírgulas (CSV), diagramas, bases de dados. Não há uma forma simples de integrar esses dados de diferentes formatos para extrair informações conforme Figura 1 e gerar conhecimento, desta forma, dificuldades são encontradas ao preparar e organizar dados. Através das ontologias é possível criar um formato único para associar estes dados de forma semântica independente de formato, além de armazenar os conceitos e os seus relacionamentos. Ontologias têm sido utilizadas para representar e organizar o conhecimento a fim de anotar semanticamente dados e documentos (GONÇALVES, 2020).

Figura 1: Exemplo de dado e informação



Fonte: elaborada pelo autor.

Diante dos dados utilizados neste trabalho, observou-se a necessidade de manipular diferentes formatos e armazenados em diferentes repositórios. Com isso, também verificou-se a necessidade de integrar estes dados de forma que os mesmos se tornem mais acessíveis e interoperáveis para serem consumidos não somente por máquinas, mas também por diferentes profissionais da área de saúde. A técnica SDD (*Semantic Data Dictionary*) permite preparar e organizar dados de forma a serem anotados semanticamente com base em modelos ontológicos (RASHID et. al, 2017). Desta forma, com base nesta contextualização, a seguinte questão surge para

realizar esta pesquisa: “como realizar a anotação semântica de dados sobre cobertura vacinal da COVID-19?”. Para responder a esta questão necessitou-se aplicar um processo sistemático de integração de dados com base na modelagem ontológica da cobertura vacinal da COVID-19 e anotação semântica dos dados utilizando a técnica SDD.

A aplicação da técnica SDD utiliza dicionários semânticos de dados, para gerar GCs a partir de ontologias e de templates de metadados. Os grafos por sua vez podem ser explorados para filtrar dados necessários, verificar hipóteses em diferentes estudos científicos sobre a cobertura vacinal. Seguido da modelagem de ontologias aplicada através da técnica SDD e do GCs gerado, os dados da cobertura vacinal podem ser acessados por partes interessadas e profissionais de domínio para estudo e entendimento. Os grafos podem ser úteis para servir como ferramenta de planejamento de órgãos públicos na administração dos recursos da saúde para contribuir no controle da pandemia.

Este trabalho está dividido nas seguintes seções, além da introdução exposta: a Seção 2 descreve os conceitos de Ontologias, Grafos de Conhecimento e consultas SPARQL para extração de informações dos grafos. A Seção 3 aborda a utilização da técnica SDD [1]bem como trabalhos correlatos desenvolvidos. A Seção 4 apresenta um estudo de caso aplicando a técnica SDD, execução da ferramenta sdd2rdf e extração de conhecimento com consultas SPARQL. Por fim, a seção 5 faz as considerações finais e lista os trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

2.1. ONTOLOGIAS DE DOMÍNIO

A palavra "ontologia” tem origem na Filosofia a partir de um ramo que estuda a que estuda a ciência e a reflexão do “ser”, visando explicar as coisas do mundo e estabelecendo conceitos. Na área de computação o conceito é um pouco diferente, uma ontologia pode ser definida como um conjunto de conceitos fundamentais e suas relações, que capta como as pessoas entendem e interpretam o domínio (ramo de atuação) em questão e permite a representação de tal entendimento de maneira formal ou gráfica, compreensível por humanos e computadores (MIZOGUCHI, 2004).

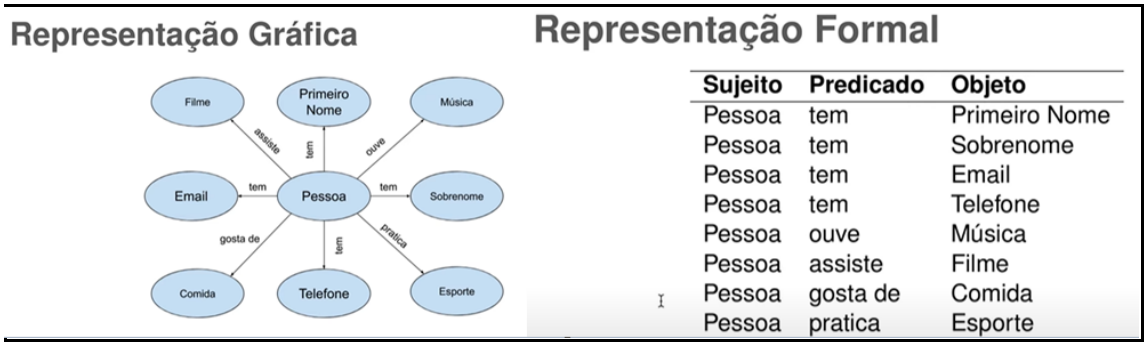
A conceituação pode ser elaborada contendo terminologias e vocabulários, estabelecendo características, propriedades e permitindo que o conhecimento seja

[1] *Semantic Data Dictionary*

reutilizado, evitando o retrabalho ou a redescoberta de terminologias equivalentes (GUARINO, 1998). O termo conceituação diz respeito a uma coleção de objetos, conceitos e entidades existentes em um determinado domínio e os relacionamentos entre eles. Uma conceituação é uma visão abstrata e simplificada do mundo que se deseja representar (ALMEIDA e BAX, 2003).

Conforme apresentado na Figura 2, a representação gráfica dos conceitos é definida por figuras ovais e seus relacionamentos definidos através de flechas, enquanto na representação formal o conceito é feito por meio de triplas. Essas triplas são representadas por um sujeito e um objeto (conceitos) e um predicado (relacionamento).

Figura 2: Representação gráfica e formal do conhecimento.



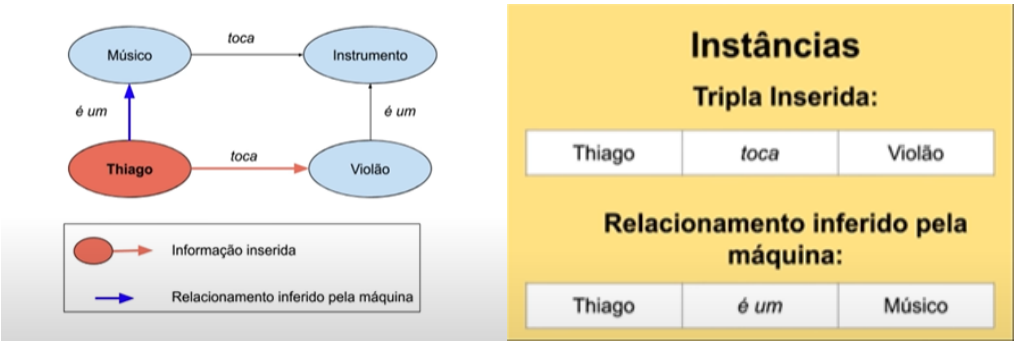
Fonte: Elaborada pelo autor.

Algumas vantagens podem ser vistas com o uso de ontologias:

- Gestão da Informação: conforme citado na introdução deste artigo as Ontologias permitem que diferentes formatos de dados (.XLSX, .CSV, .PPTX, JSON, XML e etc) sejam integrados em um único formato através de conceitos e relacionamentos. Muitos softwares armazenam dados e informações em milhares de linhas de código o que faz com que sejam entendidas por um número pequeno de pessoas, geralmente o pessoal de TI, através da ontologia poderia ser criado todo um emaranhado de conhecimento, uma teia de conhecimento, na qual todos pudessem acessar e entender a informação/conhecimento, gerando um maior engajamento na gestão de dados.
- Facilmente extensíveis: através das ontologias é possível incrementar cada vez mais com informações sem que se tenha muito processamento de dados ou limitações.

- Eficiência na Recuperação de Conhecimento: fácil recuperação da informação por meio do formato das triplas.
- Interoperabilidade entre homem e máquina: a partir das informações inseridas pelos humanos, através dos relacionamentos a máquina consegue inferir conhecimento. Nas imagens abaixo temos as ontologias “Músico toca instrumento” e “Violão é um instrumento”, ao inserir uma nova ontologia “Thiago toca Violão” a máquina por meio das triplas conseguirá entender os relacionamentos e informar que “Thiago é um músico”.

Figura 3: Exemplo de ontologias no formato gráfico e formal.



Fonte: Elaborada pelo autor.

2.2. GRAFOS DE CONHECIMENTO

De acordo com Terra (2002), o conhecimento tornou-se um recurso econômico proeminente, sendo mais importante que a matéria prima em determinadas situações. Ramos de atuação como industrial, empresarial, saúde dependem, é conhecimento gerado a partir dos dados para tomar decisões. É importante organizar o conhecimento do domínio para permitir que diferentes dados sejam integrados, independentemente de sua estrutura.

Um GC ou *Knowledge Graph* é uma estrutura composta de entidades como nós (ou vértices) e relações como tipos diferentes de arestas. São frequentemente constituídos a partir de várias fontes de dados (*datasets*). Uma instância representada por uma aresta é uma tripla (entidade principal, relação, entidade final ou sujeito, propriedade, objeto denotada como s, p, o), um exemplo de triplas pode ser visto na figura 3. Os tipos de entidades e relacionamentos são definidos por ontologias. Além disso, um GC apresenta entidades e conceitos fragmentados que podem ser conectados para formar uma solução

completa e estruturada para uma base de conhecimento, facilitando o gerenciamento, recuperação, uso e compreensão das informações (WANG et al., 2014).

A manipulação da diversidade de dados utiliza as representações de esquema que contém as identificações e o contexto das entidades presentes no GC. Esse esquema define uma estrutura de alto nível, onde a identidade dos objetos denota quais nós no grafo se referem à mesma entidade do mundo real, enquanto o contexto especifica os objetos considerados verdadeiros para o domínio específico.

Na relação formada por uma tripla um objeto pode também ser um literal, definindo uma propriedade para um recurso. Cada recurso relacionado a sujeito, predicado ou objeto, pode ser representado por um URI (*Uniform Resource Identifier*). Um URI identifica unicamente um recurso definido como um objeto de uma ontologia. Este recurso pode ser um URL (*Uniform Resource Locator*) que representa uma forma de acessar o recurso na Web.

2.3. – REPRESENTAÇÃO DO CONHECIMENTO

Hogan et al. (2020) destaca a importância das ontologias como modelos conceituais utilizados para anotar conjuntos de dados. Esta anotação permite gerar fragmentos de conhecimento do domínio que podem ser representados por GCs, como pode ser visto na Figura 4. A forma de representação gráfica das ontologias é usada para o entendimento e compreensão humano enquanto a representação formal é utilizada para entendimento de máquinas e computadores. Para a representação formal, consumida por computadores, utiliza-se algumas linguagens para a representação de ontologias, dentre as mais populares, RDF[2], RDF-S[3] e OWL[4].

O padrão RDF é um modelo para troca de dados na rede mundial através do padrão W3C e representa ontologias através de triplas (entidade principal, relação, entidade final ou sujeito, propriedade, objeto denotada como s, p, o). Através da anotação RDF:TYPE define o tipo de um novo conceito tal como uma classe, propriedade ou instância.

[2] *Resource Description Framework*
[2] *Resource Description Framework Schema*
[4] *Ontology Web Language*

Para representar a semântica através das triplas é necessária a definição de algumas tags para apresentar as propriedades do domínio. RDF-S é utilizado nesta especificação e adota anotações:

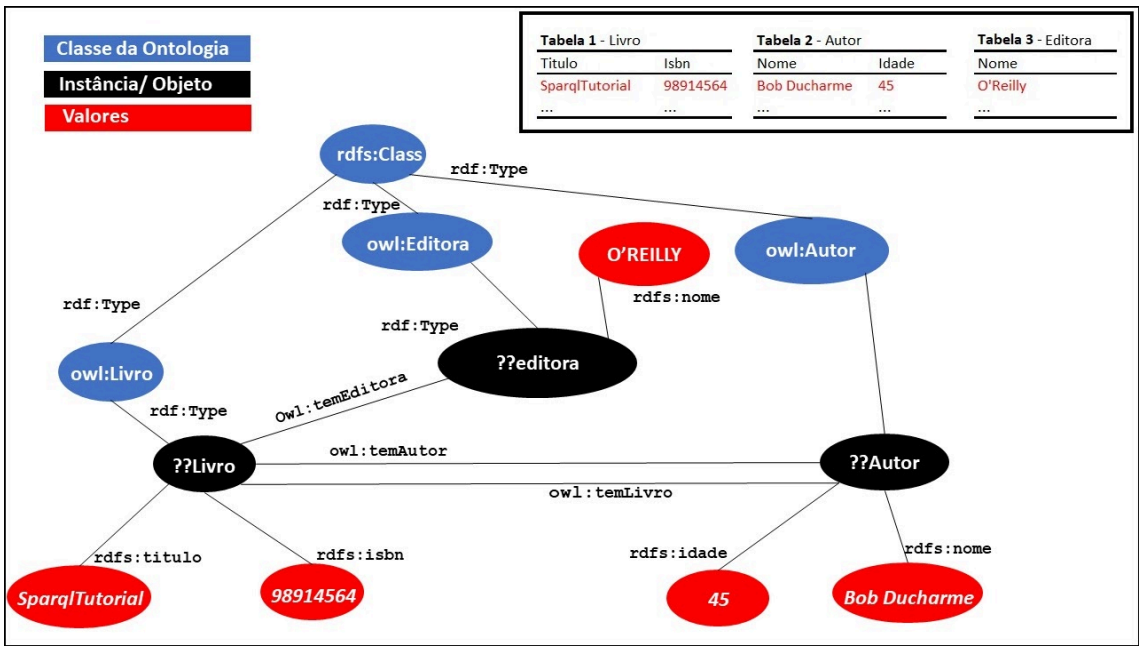
- Rdfs:label – define o termo de um novo conceito.
- Rdfs:comment – define um comentário sobre um novo conceito.
- Rdfs:subClassOf – define a superclasse ou classe “pai” de um novo conceito.
- Rdfs:domain – define o domínio de um novo relacionamento.
- Rdfs:range – define o range de um novo relacionamento.

Embora RDF-S possa ser usada para explicitar propriedades dos conceitos, ela apresenta limitações, principalmente ao apoiar o raciocínio computacional. Desta forma, uma linguagem mais expressiva foi desenvolvida e é conhecida como *Web Ontology Language* (OWL).

Horrocks et al. (2003) descreve a OWL como uma forma de satisfazer o formalismo exigido pela comunidade de Web Semântica para que programas possam compreender e responder a consultas de agentes (pessoas ou outros programas) por meio do uso de descrições ontológicas. No presente é a mais utilizada por possuir variantes da linguagem e suportar, de uma melhor forma, a extensão das ontologias com a inferência de novos conhecimentos. Aqui estão listadas algumas das tags utilizadas na linguagem OWL:

- Owl:Ontology - utilizada para definir uma ontologia.
- Owl:Class - utilizada para definir uma classe.
- Owl:ObjectProperty - utilizado para definir relacionamento de objetos ou instâncias.
- Owl:DataProperty - utilizado para definir relacionamentos de dados.
- Owl:NamedIndividual - utilizado para definir uma instância de uma classe.

Figura 4: Representação de um GC.



Fonte: elaborado pelo autor

2.4. EXTRAINDO INFORMAÇÕES DO CONHECIMENTO

Dados são representados no GC utilizando o modelo conceitual de RDF, em conjunto com as extensões de RDF-S e OWL. Esses dados podem estar armazenados em um banco de dados de triplas (*triple store*) e as informações podem ser recuperadas através da linguagem SPARQL.

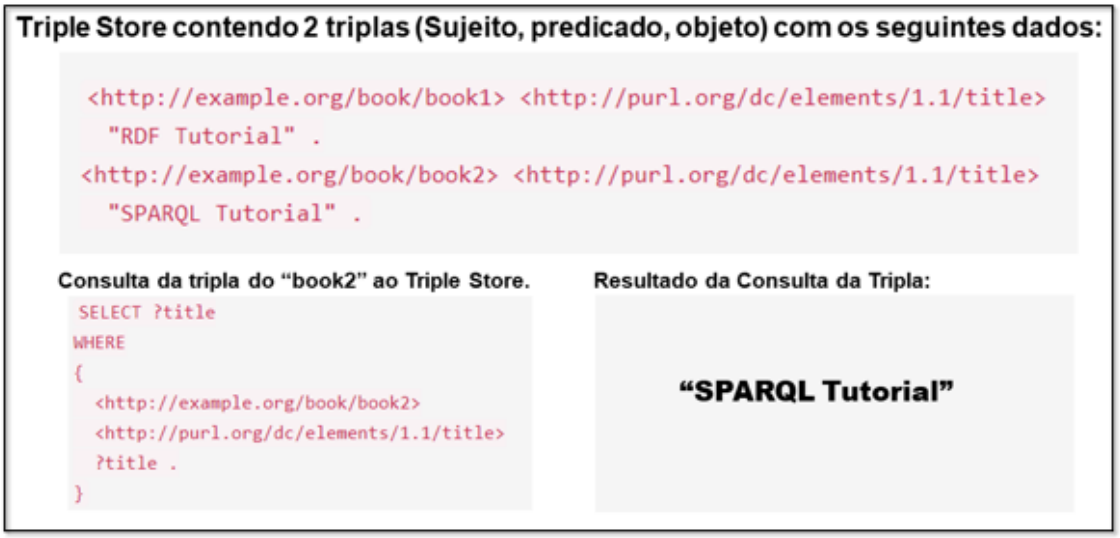
O SPARQL pode ser usado para expressar consultas em diversas fontes de dados, sejam os dados armazenados nativamente como RDF ou visualizadores de RDF por meio de softwares que fornecem os recursos necessários.

Pode-se estabelecer uma analogia entre SPARQL e a linguagem SQL [5]de consulta a bancos de dados relacionais, porém deve-se considerar que SPARQL tem uma sintaxe adequada a consultas a dados representados como um conjunto de triplas RDF. A consulta consiste em duas partes: a cláusula *SELECT* identifica as variáveis que aparecem nos resultados da consulta e a cláusula *WHERE* fornece o padrão gráfico básico para comparar com o gráfico de dados. As variáveis começam com um caractere "?" e podem ser definidas em qualquer uma das três posições de uma tripla (sujeito, predicado, objeto) no conjunto de dados RDF. Operadores e filtros podem ser acrescentados para criar consultas

[5] Structured Query Language

complexas ao conjunto de triplas. Conforme figura 5, uma recuperação de informação é feita no *triplestore*.

Figura 5: Representação recuperação de conhecimento com SPARQL



Fonte: Elaborada pelo autor

3. ANOTAÇÃO SEMÂNTICA COM DICIONÁRIOS SEMÂNTICOS DE DADOS

O vocábulo "anotação" implica, de uma forma geral, realizar um apontamento ou associação de alguma coisa. Uma anotação realiza apontamentos de alguns dados a outros dados: ela estabelece, dentro de algum contexto, uma relação entre dados anotados. Segundo BELLOZE et al.(2012) uma anotação semântica é uma associação entre as expressões ou termos relevantes de um documento e os conceitos descritos em uma ontologia.

Três tipos de anotações podem ser implementados: informais, formais e ontológicas. As anotações informais não são legíveis por máquina porque não usam uma linguagem formal. As anotações formais são compreensíveis por máquina, mas não usam termos ontológicos. Nas anotações ontológicas, a terminologia tem um significado comumente compreendido que corresponde a uma conceituação compartilhada por uma ontologia.

Anotações semânticas são utilizadas em diferentes domínios de aplicação para os mais diversos fins. Ferramentas de anotação manual permitem que os usuários adicionem anotações a recursos na web ou outros recursos e as compartilhem. Uma anotação de exemplo pode relacionar o texto "Azul" a uma ontologia, identificando-a

como companhia de transporte aéreo e não a cor como característica de alguém ou alguma coisa.

3.1. DICIONÁRIOS SEMÂNTICOS DE DADOS (SDD)

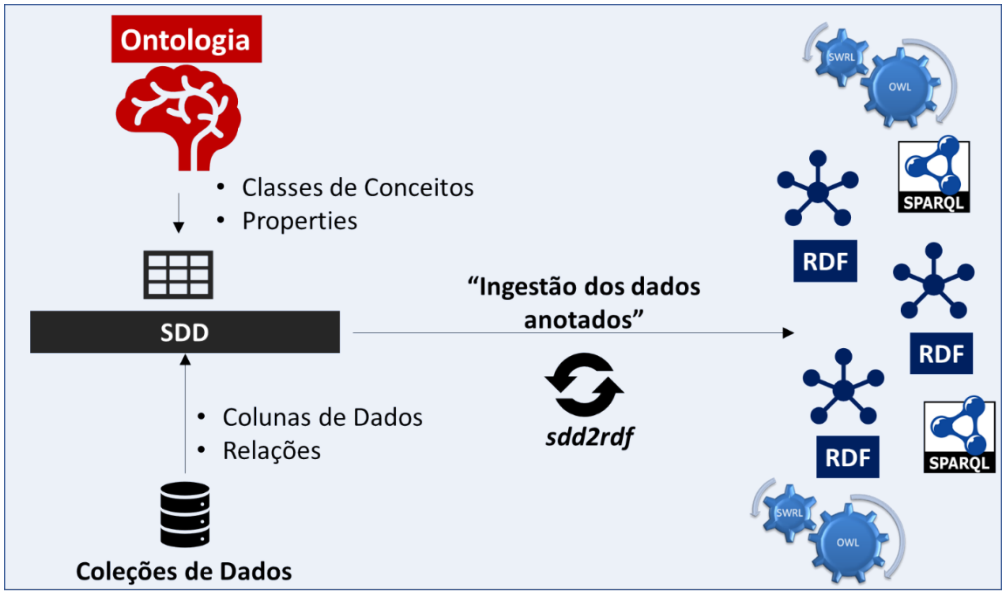
O SDD é um conjunto de padrões de metadados fundamentado em ontologias que descrevem objetos (representados por dados) em classes e relacionamentos (RASHID et. al, 2017). A anotação por SDD é feita de forma manual e associa os dados de um conjunto de dados(*dataset*) a conceitos (ou classes) nas ontologias, formalizando a semântica dos dados. A anotação dos dados pode ser realizada por profissionais relacionados com o domínio e os *datasets*, com apoio de ontologistas. A formalização do vocabulário abre caminho para a interoperabilidade dos dados que podem ser integrados de fontes diversas. Abaixo estão relacionados os artefatos utilizados na anotação por SDD:

- Ontologia de domínio. Formaliza os conceitos do domínio da pesquisa. Deve-se buscar reutilizar ontologias consolidadas.
- Dictionary Mapping (DM). Anota as colunas do *dataset*. Cada linha do DM mapeia uma coluna do *dataset*, formalizando-a conceitualmente, explicitando suas relações com os outros dados, bem como a sua proveniência.
- CodeBook. Estrutura os dados categóricos de um *dataset*, mapeando-os para conceitos correspondentes na ontologia.
- Infosheet. Organiza os metadados de descrição do SDD.
- GC (RDF). É gerado com interpretação da dupla: "SDD (templates de metadados) + Dados" pelo script de anotação *sdd2rdf*. Caso seja necessário persistir os dados, o usuário pode armazenar o grafo em um *triplestore* para consulta posterior dos dados.

Inicialmente, os dados mapeados para as ontologias pelo SDD são as colunas do próprio *dataset*. Os objetos caracterizados nos *datasets* podem estar implicitamente representados. Os objetos implícitos serão explicitados no SDD e formalizados no grafo final gerado. A explicitação dos objetos implícitos favorece a integração semântica dos dados nos níveis conceituais mais abstratos do domínio do projeto, permitindo normalizar e harmonizar interpretações de conceitos que descrevem dados que se deseja integrar.

Uma vez que a estrutura de anotação esteja pronta, a ferramenta sdd2rdf é utilizada para integrar os dados do *dataset* descrito pelo SDD formando um GC persistido em *triplestores*. A sdd2rdf (*SEMANTIC DATA DICTIONARY*, 2019) é um script/software, desenvolvido na linguagem python, que interpreta o SDD e converte os dados do *dataset* descrito pelo SDD gerando o GC expresso no padrão RDF (RASHID et. al, 2017). O GC possibilita a inferência de novos fatos enriquecendo o compartilhamento do conhecimento. O grafo RDF gerado se fundamenta em ontologias, e possibilita a integração semântica dos dados bem como a sua interoperabilidade. Para exemplificar o acesso aos dados anotados no grafo, o sdd2rdf cria alguns exemplos de consultas SPARQL. A Figura 6 representa de forma simplificada o processo de anotação feita com o uso de SDDs.

Figura 6: Representação da anotação de dados feita pelo SDD.



Fonte: elaborada pelo autor

3.2. TRABALHOS RELACIONADOS

Dentro da área de saúde foram encontrados trabalhos a respeito de prontuários com informações clínicas de registros médicos. GOODWIN e HARABAGIU (2013) enfatizam que existe uma quantidade extraordinária de informações clínicas disponíveis nos Registros Médicos Eletrônicos (EMR, *Electronic Medical Records*). No entanto, interpretar esse conhecimento normalmente exige um nível significativo de compreensão clínica. Isso pode ser facilitado pelo acesso a bases de conhecimento estruturadas. Para os autores, mesmo que as bases de conhecimento sejam vastas, na área biomédica as

informações disponíveis são muito limitadas. Em contraste, o texto clínico expressa muitas relações entre conceitos usando uma quantidade extraordinária de variação, ou seja, se um conceito médico está presente, incerto ou ausente. Os autores propõem um método para construir automaticamente um GC com conceitos clínicos relacionados. Para esse propósito, os dados foram classificados partindo do estado de crença de certos conceitos médicos. No segundo momento, projetou-se a técnica para construir o GC com conceitos qualificados pelo valor de crença do médico. Os autores demonstraram várias técnicas para inferir a similaridade entre conceitos médicos qualificados. Finalmente, apresentaram que a incorporação do conhecimento codificado a partir do grafo produz resultados competitivos quando aplicado à expansão da consulta para a recuperação de grupos de pacientes hospitalares.

Koopman et al. (2016) apresentam um modelo de recuperação a partir da inferência do conhecimento usando GCs. Os autores utilizam um modelo enriquecido semanticamente aplicado ao domínio médico. Este domínio apresenta recursos de conhecimento estruturados e textos não estruturados, sendo um desafio para recuperação da informação. Documentos na área médica foram utilizados para aplicar o modelo proposto. Estes documentos foram agrupados por métodos de pesquisa baseados em palavras-chave, que permitiu avaliar a relevância para recuperar novos documentos. O mecanismo de inferência do modelo promoveu a recuperação recuperando novos documentos relevantes não encontrados por abordagens anteriores baseadas em palavras-chave.

4.0. APLICAÇÃO DA TÉCNICA SDD E REPRESENTAÇÃO DO CONHECIMENTO

Para execução do processo de anotação semântica demonstrado na seção 3, foi utilizado um processo de anotação manual utilizando o conjunto de documentos: *Dictionary Mapping, Codebook, Infosheet* apresentados nas respectivas tabelas 1, 2 e 3.

4.1. COLETA DE DADOS DATASET COBERTURA VACINAL

Os dados utilizados neste trabalho sobre a cobertura vacinal foram extraídos do site do Ministério da Saúde[6] através de um documento disponibilizado no formato .CSV

[6] https://dados.gov.br/dataset/covid-19-vacinacao

e apresentam um *schema*, arquivo dicionário de dados, próprio definido na API de acesso do portal DATASUS.

4.2. ANOTAÇÃO SEMÂNTICA DOS DADOS

4.2.1. DICTIONARY MAPPING

A anotação semântica é criada com base no *schema*, colunas já pré-definidas, do dataset para dados explícitos e implícitos. Através da coluna "attributeOf" é colocada uma anotação do conjunto de conceitos do domínio referente ao atributo do *dataset* fornecido pelo DataSUS. As classes ontológicas são mapeadas na coluna "Entity" assim como suas relações na coluna "Relation", conforme tabela 1.

Tabela 1: Fragmento do *dictionary mapping* da cobertura vacinal.

Column	Attribute	attributeOf	Entity	Relation	inRelationTo
paciente_id	vacinacao:id	??paciente			
estabelecimenta_razaoSocia1	vacinacao:LocalVacinacao	??razaosocialvacinacao			
estabelecimento_noFantasia	vacinacao:LocalVacinacao	??nomefantasiavacinacao			
estabelecimento_municipio_nome	vacinacao:Municipio	??municipiolocalvacinacao			
estabelecimento_uf	vacinacao:UnidadeFederativa	??uflocalvacinacao			
vacina_grupo_atendimento_nome	vacinacao:GrupoDeAtendimento	??grupodeatendimentodopaciente			
vacina_dataAplicacao	vacinacao:dataDaAplicacaoDaVacina	??datadaaplicacaodavacinanopaciente			
vacina_nome	vacinacao:NomeDaVacina	??vacina			
??vacinacao			vacinacao:Vacinacao	vacinacao:vacinacaoECompostaPor	vacinacao:Paciente
??vacinacao			vacinacao:Vacinacao	vacinacao:vacinacaoECompostaPor	vacinacao:LocalDeVacinacao
??paciente			vacinacao:Paciente	vacinacao:pacienteRecebe	vacinacao:Vacina
??vacina			vacinacao:Vacina	vacinacao:vacinaPossui	vacinacao:FabricanteDaVacina
??vacina			vacinacao:Vacina	vacinacao:vacinaPossui	vacinacao:CategoriaDaVacina

Fonte: Elaborado pelo autor.

4.2.2. PREENCHIMENTO CODEBOOK

O codebook estrutura os dados categóricos de um dataset, mapeando-os para conceitos correspondentes na ontologia. Dado o schema do dataset de cobertura vacinal os conceitos de "paciente_racaCor_valor" e "paciente_enumSexobiologico" foram categorizados.

TABELA 2: Codebook dataset cobertura vacinal.

Column	Code	Class
paciente_enumSexobiologico	M	:Masculino
paciente_enumSexobiologico	F	:Feminino
paciente_racaCor_valor	1	:Branca
paciente_racaCor_valor	2	:Preta
paciente_racaCor_valor	3	:Parda
paciente_racaCor_valor	4	:Amarela
paciente_racaCor_valor	99	:SemInformacao

Fonte: Elaborado pelo autor.

4.2.3. ARQUIVO DE CONFIGURAÇÃO INFOSHEET

O Infosheet é um arquivo de configuração dos metadados onde pode ser descritas informações como do autor, data de criação, formato dos arquivos utilizados, versão e outros. Ademais, os arquivos do *Dictionary Mapping, Codebook e Code Mapping*, encontram-se apontados no Infosheet. O arquivo *Code Mapping* descreve unidades de medidas mapeadas no Dictonary Mapping, porém para este trabalho não foi necessária a sua utilização.

TABELA 3 - Infosheet.

Attribute	**Value**
Type	http://purl.org/dc/dcmitype/Dataset
Title	CoberturaVacinal
Alternative Title	CoberturaVacinal
Comment	Criando um grafo de conhecimento usando a técnica SDD
Description	Os dados foram anotados do dataset CoberturaVacinalCovid19 fornecido pelo Ministério da Saúde para demonstrar a técnica SDD
Date Created	12/10/2021
Creators	Leonardo Heredia
Contributors	Evaldo de Oliveira
Publisher	Leonardo Heredia
Date of Issue	12/10/2021
Identifier	CoberturaVacinal
Version	1.0
Source	CoberturaVacinal/config/Infosheet.csv
File Format	csv
Dictionary Mapping	CoberturaVacinal/input/DM/SDD.csv
Codebook	CoberturaVacinal/input/CB/Codebook.csv
Code Mapping	CoberturaVacinal/config/code_mappings.csv

Fonte: Elaborado pelo autor

4.3. EXECUÇÃO DO SCRIPT SDD2RDF

Com a estrutura de anotação definida, por fim está pronto para ser executada a ferramenta sdd2rdf. Porém, antes de iniciar o processo de execução do script, necessita-se baixar o projeto sdd2rdf do repositório "semanticdatadictionary" localizado na plataforma Github[7] e realizar as configurações de ambiente para a execução da ferramenta. Deve ser feita a instalação das seguintes tecnologias e bibliotecas:

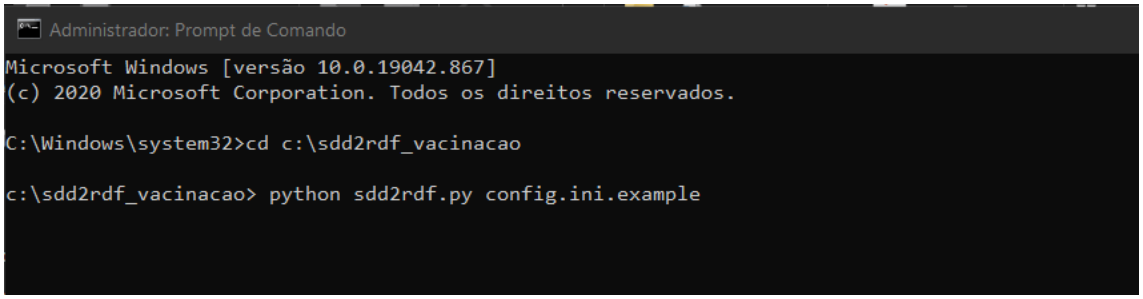
- Instalação Python 3[8]**:** linguagem de programação de alto nível necessária para a execução do script.

[7] https://github.com/tetherless-world/SemanticDataDictionary
[8] https://www.python.org/downloads/

- Instalação da biblioteca Pandas: biblioteca Python para análise de dados. Pode ser instalada através do comando “pip install pandas” no terminal windows.
- Instalação da biblioteca rdflib: biblioteca Python para utilização do formato RDF. Pode ser instalada através do comando “pip install rdflib” no terminal windows.
- Instalação da ferramenta Docker Desktop(Windows): ferramenta de virtualização para criação de ambientes de software.[9]
- Baixar uma instância do container BlazeGraph: banco de triplas (*triplestore*), para persistência das triplas geradas pelo script sdd2rdf, através do comando “docker pull metaphacts/blazegraph-basic” no terminal.

Preparado o ambiente é necessário navegar até a pasta onde se encontra o arquivo de download do projeto sdd2rdf e alterar os dados do arquivo “config.ini.example” , apontando para a pasta com os dados anotados (*Infosheet, Codebook, Code Mapping, Dictionary Mapping*). Com as modificações realizadas, o script está pronto para ser executado.

Figura 11: Execução do script sdd2rdf



Fonte: Elaborado pelo autor.

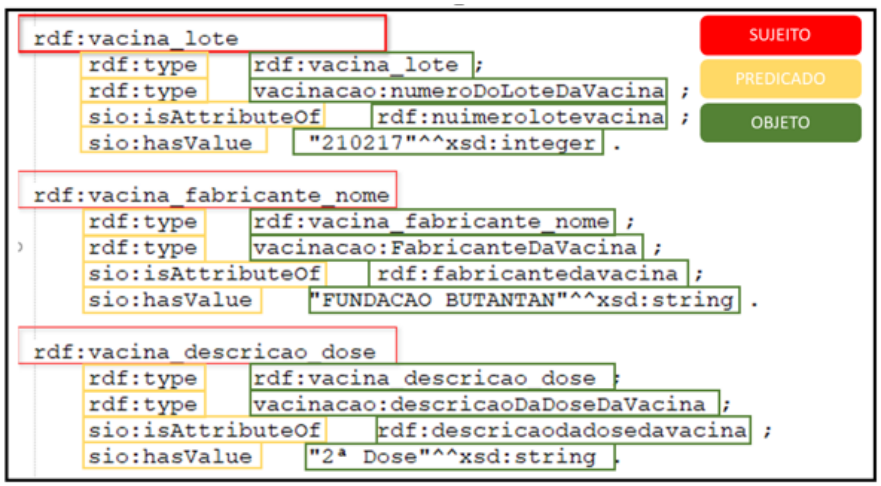
Terminada a execução do script um arquivo com a extensão “.trig” é gerado com as triplas armazenadas.

[9] https://docs.docker.com/desktop/windows/install/

4.4 REPRESENTAÇÃO DO CONHECIMENTO

Os arquivos com extensão .trig podem então ser consumidos no BlazeGraph e a extração de conhecimento pode ser feita com consultas SparQL através das Triplas geradas conforme pode ser visto nas figuras 11 e 12.

Figura 11: Triplas Geradas com a execução do script.



Fonte: Elaborado pelo autor

Figura 12: Query para extração do conhecimento do nome dos fabricantes de vacinas aplicadas.

Query SPARQL para extração dos nomes dos fabricantes das Vacinas

```
1 prefix vacinacao:<C:/sdd2rdf_vacinacao/CoberturaVacinal>
2
3 SELECT ?vacina_fabricante_nome
4 WHERE {?vacinacao <rdf:type> vacinacao:FabricanteDaVacina ;
5         <sio:hasValue> ?vacina_fabricante_nome .
6 }
7
8
9
```

Resultado da query SPARQL

vacina_fabricante_nome
SERUM INSTITUTE OF INDIA LTD
BioNTech/Fosun Pharma/Pfizer
FUNDACAO BUTANTAN
FUNDACAO OSWALDO CRUZ

Fonte: Elaborado pelo autor

5.0. CONSIDERAÇÕES FINAIS

Este trabalho apresentou um estudo de caso aplicando uma modelagem de dados com anotações semânticas com base nos dados disponibilizados pelo Ministério da Saúde a respeito da cobertura vacinal contra a COVID-19. Extrair informações de dados com estruturas já definidas a fim de se gerar conhecimento não é trivial e demanda esforço e conhecimento técnico para o mesmo.

A técnica SDD permite a extensão e integração de dados de vários domínios por meio de um padrão de metadados comum, estes metadados são baseados em ontologias e podem ser usados para consultar conjuntos de dados relevantes sem o conhecimento de qualquer um dos conjuntos de dados estrutura.

Em futuras pesquisas e trabalhos, através da possibilidade de inferir novos dados que as ontologias oferecem, o GC gerado pode ser ampliado com outros dados do domínio da área da saúde tais como taxa de ocupação de leitos, sintomas da COVID-19 e variantes, para profissionais de saúde estudarem e compreenderem questões de competência do domínio como as possíveis listadas abaixo:

- Qual a efetividade da vacina em pessoas com comorbidades?
- A taxa de infectados e hospitalizados, para casos moderados e graves, diminuiu ou se manteve nos leitos diante da região onde foi aplicada a vacina?
- Qual a efetividade das variantes Alfa, Beta, Gama e Delta mediante as vacinas aplicadas?
- Qual a efetividade das vacinas em suas respectivas faixas etárias?
- É necessário a aplicação de doses anuais contra a Covid-19 e suas variantes?

Desta forma, o grafo poderá auxiliar os profissionais a tomarem decisões de uma forma mais precisa e gerar conhecimento para possíveis tratamentos à esta doença recente e ainda muito pouco explorada.

6. Referências

ALMEIDA, M. B.; BAX, M. P. Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. Ciência da Informação, Brasília DF, v. 32, n. 3, p. 7-20, 2003.

FAPESP. Programa FAPESP de Pesquisa em eScience. Disponível em: https://fapesp.br/publicacoes/2015/folder_escience.pdf. Acesso em 26 de ago de 2021.

GUARINO, Nicola. Formal Ontology and Information Systems. In: Proceedings of the First Int. Conference on Formal Ontology in Information Systems, Trento, Italy, Junho 1998.

HOGAN, Aidan, BLOMQVIST , Eva, COCHEZ , Michael, D'AMATO, Claudia, MELO Gerard de, GUTIERREZ, Claudio, GAYO, José Emilio Labra, KIRRANE, Sabrina, NEUMAIER, Sebastian, POLLERES, Axel, NAVIGLI, Roberto, NGOMO, Axel-Cyrille Ngonga, RASHID, Sabbir M., RULA, Anisa, SCHMELZEISEN, Lukas, SEQUEDA,

ISOTANI, Seiji Dados Conectados Abertos. Disponível em https://ceweb.br/livros/dados-abertos-conectados//capitulo-3/. Acesso em 11 de novembro de 2021.

Juan, STAAB, Steffen, ZIMMERMANN, Antoine. Knowledge Graphs. arXiv preprint arXiv:2003.02320, 2020.

MARCHIORI, Patricia Zeni. A ciência e a gestão da informação: compatibilidades no espaço profissional. Ciência da informação, v. 31, n. 2, p. 72-79, 2002.

MEDEIROS, Claudia B. Gestão de Dados Científicos – da coleta à preservação. Disponível em https://tinyurl.com/4w44ubu. Acesso em 07 de agosto de 2021.

MIZOGUCHI, R.; VANWELKENHUYSEN, J.; IKEDA, M. Task ontology for reuse of problem solving knowledge. In: Proceedings OF ECAI'94 TOWARDS VERY LARGE KNOWLEDGE BASES, 1995, Amsterdam: IOS Press, 1995. p. 46-59.

MINISTERIO DA SAUDE, Campanha Nacional de Vacinação contra Covid-19. Disponível em https://dados.gov.br/dataset/covid-19-vacinacao . Acessado em 15 de outubro de 2021.

OREN, Eyal et al. What are semantic annotations. Relatório técnico. DERI Galway, v. 9, p. 62, 2006.

RASHID, Sabbir M. et al. The Semantic Data Dictionary Approach to Data Annotation & Integration. In: SemSci@ ISWC. 2017. p. 47-54.

RASHID, S. M., MCCUSKER, J. P., PINHEIRO, P., BAX, M. P., SANTOS, H., STINGONE, J. A., ... & MCGUINNESS, D. L. (2020). The Semantic Data Dictionary–An Approach for Describing and Annotating Data. Data Intelligence, 443-486.

WANG, Zhen et al. Knowledge graph embedding by translating on hyperplanes. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2014.