UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS DE CRATEÚS

CURSO DE GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

YURI CRISTIAN PEDROSA DE OLIVEIRA

UMA ABORDAGEM DE INTEGRAÇÃO SEMÂNTICA PARA O OBSERVATÓRIO DE DADOS ABERTOS DO SERTÃO DOS CRATEÚS

CRATEÚS

2024

YURI CRISTIAN PEDROSA DE OLIVEIRA

UMA ABORDAGEM DE INTEGRAÇÃO SEMÂNTICA PARA O OBSERVATÓRIO DE DADOS ABERTOS DO SERTÃO DOS CRATEÚS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Sistemas de Informação do Campus de Crateús da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Sistemas de Informação.

Orientadora: Profª. Msc. Lisieux Marie Marinho dos Santos Andrade

Coorientadora: Profª. Drª. Amanda Drielly Pires Venceslau

CRATEÚS

2024

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

O52a Oliveira, Yuri Cristian Pedrosa de.
Uma abordagem de integração semântica para o observatório de dados abertos do sertão dos Crateús / Yuri Cristian Pedrosa de Oliveira. – 2024.
53 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Crateús, Curso de Sistemas de Informação, Crateús, 2024.
Orientação: Profa. Ma. Lisieux Marie Marinho dos Santos Andrade.
Coorientação: Prof. Dr. Amanda Drielly Pires Venceslau.

1. Dados governamentais abertos. 2. Observatório de dados abertos. 3. Web Semântica. 4. Integração Semântica. 5. Integração de dados. I. Título.

CDD 005

YURI CRISTIAN PEDROSA DE OLIVEIRA

UMA ABORDAGEM DE INTEGRAÇÃO SEMÂNTICA PARA O OBSERVATÓRIO DE DADOS ABERTOS DO SERTÃO DOS CRATEÚS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Sistemas de Informação do Campus de Crateús da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Sistemas de Informação.

Aprovada em:

BANCA EXAMINADORA

Profª. Msc. Lisieux Marie Marinho dos Santos Andrade (Orientadora)
Universidade Federal do Ceará (UFC)

Profª. Drª. Amanda Drielly Pires Venceslau (Coorientadora)

Prof. Dr. Jose Wellington Franco da Silva
Universidade Federal do Ceará (UFC)

Prof. Msc. Marciel Barros Pereira
Universidade Federal do Ceará (UFC)

À minha família, por seu apoio incondicional e por acreditar em mim em todos os momentos. Mãe, seu cuidado e dedicação foram a base que me sustentou. Pai, sua presença constante me deu a segurança necessária para seguir em frente.

AGRADECIMENTOS

À minha orientadora, Profª. Msc. Lisieux Marie Marinho dos Santos Andrade, pela orientação, paciência e apoio durante todo o desenvolvimento deste trabalho.

À minha coorientadora, Profª. Drª. Amanda Drielly Pires Venceslau, pelos valiosos conselhos e contribuições essenciais para a realização deste projeto.

Ao Emanoel Bezerra Alves, por seu trabalho, amizade e apoio, que foram fundamentais durante esta jornada. Sua colaboração e incentivo constante fizeram a diferença em momentos críticos, ajudando a superar desafios e a alcançar este objetivo.

A todos os professores que contribuíram para minha formação acadêmica, pelo conhecimento transmitido e pelo exemplo de caráter e dedicação. Cada ensinamento foi um passo importante na construção deste projeto e na minha formação como profissional.

Aos meus amigos, que sempre me incentivaram a continuar, oferecendo suporte e encorajamento nos momentos mais desafiadores. Sua amizade e palavras de incentivo foram essenciais para manter a motivação e seguir em frente.

Aos meus familiares e a tantos outros que foram vítimas da COVID-19. Às famílias que sofreram perdas imensuráveis e aos profissionais de saúde que estiveram na linha de frente, enfrentando riscos diariamente para salvar vidas. Que suas memórias sejam sempre honradas e que encontremos força em sua coragem e resiliência.

"A verdadeira sabedoria está em reconhecer a própria ignorância."

(Sócrates)

RESUMO

O acesso à informação é um direito fundamental e a transparência governamental é um elemento essencial para garantir a participação cidadã na gestão pública. A Lei de Acesso à Informação do Brasil garante o direito constitucional ao acesso à informação, onde a disponibilização de informações por meio de dados abertos é um requisito para o cumprimento desta legislação. Na região oeste do estado do Ceará há o Observatório de Dados Abertos do Sertão dos Crateús, uma iniciativa em desenvolvimento na Universidade Federal do Ceará, que visa aumentar a transparência e o controle social dos dados públicos das cidades que compõem a região. No entanto, os dados fornecidos pelas fontes utilizadas no Observatório estão estruturados de forma heterogênea, dificultando a sua compreensão e reutilização. Portanto, a criação de uma estrutura unificada que facilite a interpretação e reutilização dos dados é o objetivo principal desta pesquisa. Para atingir este objetivo, foi implementada uma metodologia para integrar os dados heterogêneos do Observatório, através da Web Semântica, em que a expansão de ontologias, a coleta e tratamento de dados, o mapeamento para a ontologia expandida e a integração e correlação de dados foram componentes da abordagem. Os dados utilizados nesta pesquisa, são de bases de dados distintas e externas utilizadas pelo Observatório, refentes aos casos do COVID-19 e de índices de mortalidade de acordo com a Classificação Internacional de Doenças (CID-10). Para além da integração das informações, esta pesquisa buscou verificar a correlação dos casos COVID-19 com os índices de mortalidades registradas no ano de 2020 por outras doenças respiratórias. Após a aplicação da metodologia e a integração das informações, a análise estatística, por meio do coeficiente de correlação de Pearson, revelou uma correlação muito fraca entre o número de casos de COVID-19 e a mortalidade por outras doenças respiratórias nos cenários analisados. Com os resultados obtidos, conclui-se que a integração semântica de dados, por meio da metodologia utilizada, é viável e pode promover maior transparência das informações públicas e revelar a correlação de dados integrados em outros Observatórios de Dados Abertos.

Palavras-chave: Dados governamentais abertos. Observatório de dados abertos. Web Semântica. Integração Semântica. Integração de dados.

ABSTRACT

Access to information is a fundamental right, and government transparency is an essential element to ensure citizen participation in public management. Brazil's Access to Information Law guarantees the constitutional right to access information, where the provision of information through open data is a requirement for compliance with this legislation. In the western region of the state of Ceará, there is the Observatório de Dados Abertos do Sertão dos Crateús, an initiative under development at the Universidade Federal do Ceará, which aims to increase transparency and social control of public data from the cities that make up the region. However, the data provided by the sources used in the Observatory are structured in a heterogeneous manner, making their understanding and reuse difficult. Therefore, the creation of a unified structure that facilitates the interpretation and reuse of data is the main objective of this research. To achieve this objective, a methodology was implemented to integrate the heterogeneous data of the Observatory through Semantic Web, where ontology expansion, data collection and processing, mapping to the expanded ontology, and data integration and correlation were components of the approach. The data used in this research come from distinct and external databases used by the Observatory, related to COVID-19 cases and mortality rates according to the International Classification of Diseases (ICD-10). Beyond the integration of information, this research sought to verify the correlation of COVID-19 cases with the mortality rates recorded in 2020 from other respiratory diseases. After applying the methodology and integrating the information, statistical analysis using Pearson's correlation coefficient revealed a very weak correlation between the number of COVID-19 cases and mortality from other respiratory diseases in the analyzed scenarios. With the results obtained, it is concluded that the semantic integration of data, through the methodology used, is feasible and can promote greater transparency of public information and reveal the correlation of integrated data in other Open Data Observatories.

Keywords: Open Government Data. Open Data Observatory. Semantic Web. Semantic Integration. Data Integration

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE CÓDIGOS-FONTE

LISTA DE ABREVIATURAS E SIGLAS

ABox	Assertion Box
CID-10	Classificação Internacional de Doenças, 10ª Revisão
DATASUS	Departamento de Informática do Sistema Único de Saúde
DC	Dublin Core
e-SUS	Sistema de Informação em Saúde para a Atenção Básica
EKG	Enterprise Knowledge Graph
GISSA	Governança Inteligente dos Sistemas de Saúde
ICD-10-CM	International Classification of Diseases, Version 10 - Clinical Modification
LDM	Linked Data Mashup
RDF	Resource Decription Framework
RDFS	Resource Decription Framework Schema
SINASC	Sistema de Informações sobre Nascidos Vivos
SPARQL	SPARQL Protocol and RDF Query Language
SUS	Sistema Único de Saúde
TBox	Terminology Box
XML	Extensible Markup Language

SUMÁRIO

1 INTRODUÇÃO

A Web atual permite acesso a uma abundante gama de dados abertos, os quais são conteúdos, informações ou dados de acesso livre para usar, reutilizar e redistribuir, sem nenhuma restrição legal, tecnológica ou social (KNOWLEDGE, 2020). O Open Government Data é um movimento cujo objetivo é incentivar a disponibilização de dados governamentais sobre diversas áreas: Saúde, Educação, Clima, Finanças, dentre outras. No Brasil com a aprovação da Lei 12.527, de 18 de novembro de 2011, a Lei de Acesso à Informação [1], a administração pública tem se tornado cada vez mais transparente nas suas ações ao utilizar ferramentas tecnológicas para prestação de serviços à população, como a publicação de informações de modo a aproximar governo e cidadão.

Dessa forma, o governo brasileiro tem buscado investir nas Tecnologias da Informação e da Comunicação, onde a administração pública tem trabalhado na publicação de dados abertos, que podem ser acessados por qualquer indivíduo por meio da internet em portais governamentais. Alguns desses portais são: Portal da transparência[2], Portal Brasileiro de Dados Abertos[3], Departamento de Informática do Sistema Único de Saúde (DATASUS)[4].

Apesar de tais iniciativas, a cultura de publicação de dados abertos no Brasil é muito tímida e carente de melhor infraestrutura de acesso pelos usuários finais e de padrões, um exemplo são as bases de dados que além de não utilizarem um padrão para publicação, possuem modelos de dados distintos, fazendo com que cada base pareça não ter relação com as demais (SEGUNDO, 2015). Diante de tais dificuldades os conceitos de Web Semântica surgem como solução para estabelecer relações entre conjuntos de dados, não apenas tornando os mesmos acessíveis e processáveis por máquinas, mas passíveis de processos de organização que podem facilitar a geração de novos dados, apresentação de resultados, aumento do conhecimento para tomadas de decisão, novos modelos de dados gerados a partir do relacionamento e cruzamento de dados de várias esferas governamentais.

Sanando esta deficiência, a região do Sertão dos Crateús conta atualmente com um Observatório de Dados Abertos da Região do Sertão dos Crateús (ALVES, 2022), como uma solução inicial, assim como outros portais de dados, ainda não dispõe de semântica em suas informações disponibilizadas.

---

1 https://www.gov.br/aeb/pt-br/acesso-a-informacao/lei-de-acesso-a-informacao
2 https://www.portaldatransparencia.gov.br/
3 https://dados.gov.br/
4 https://datasus.saude.gov.br/

Como solução para essa problemática, a presente pesquisa apresenta uma abordagem de integração de dados no Observatório de Dados Abertos da Região do Sertão dos Crateús, destacando um estudo de caso no domínio de saúde.

1.1 Objetivos

1.1.1 Objetivo Geral

Desenvolver uma abordagem de integração semântica de dados utilizando conceitos e tecnologias da Web Semântica para as informações presentes no Observatório de Dados Abertos da Região do Sertão dos Crateús, com foco nos dados relacionados aos casos de COVID-19 e mortalidade por outras doenças respiratórias.

1.1.2 Objetivos Específicos

- Estudar conceitos e tecnologias semânticas;
- Analisar e normatizar dados do Observatório de Dados Abertos da Região do Sertão dos Crateús;
- Aplicar a abordagem de integração ao Observatório;
- Discutir o emprego dos procedimentos propostos pela abordagem de integração.

1.2 Organização do Trabalho

Com a finalidade de proporcionar uma melhor compreensão deste trabalho, a presente pesquisa foi estruturada em seis capítulos que abrangem os aspectos teóricos e práticos da pesquisa. O Capítulo 2 aborda os conceitos e tecnologias da Web Semântica, incluindo ontologias, RDF, e SPARQL, além do conceito de Mashup de Dados. O Capítulo 3 analisa estudos e frameworks existentes para a integração de dados e a construção de mashups. No Capítulo 4, é detalhada a abordagem proposta, incluindo a arquitetura da solução e as etapas metodológicas para a integração e análise dos dados. O Capítulo 5 descreve o processo de integração de dados, desde a definição e expansão da ontologia até a coleta, tratamento, mapeamento e correlação dos dados, explicando a aplicação da metodologia no estudo de caso. O Capítulo 6 apresenta e discute os resultados obtidos, incluindo o estudo de caso que compreende a análise das correlações entre os dados de COVID-19 e outras doenças respiratórias. Finalmente, o Capítulo 7 resume

os principais achados da pesquisa, discute as implicações dos resultados e sugere direções para trabalhos futuros, proporcionando uma visão clara e detalhada do desenvolvimento da pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, serão apresentados conceitos essenciais para a compreensão do trabalho. A organização segue a seguinte estrutura, na Seção 2.1 será apresentada definição de Web Semântica e suas tecnologias, as quais são de suma importância para compreensão do contexto em que a presente pesquisa está inserida. Por fim, a Seção 2.2, onde será discutido o conceito de Mashup de dados e o seu processo de criação.

2.1 Tecnologias da Web Semântica

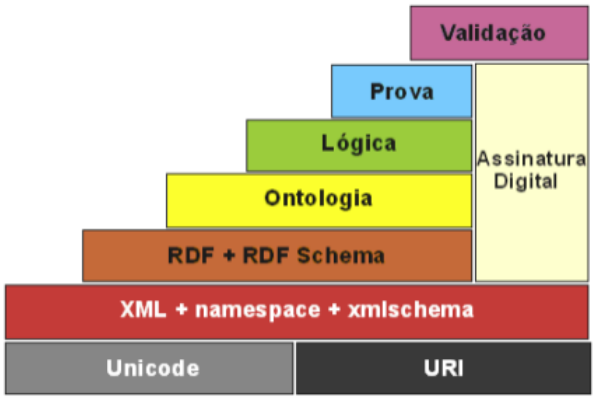
Segundo Berners-Lee et al. (2001), a Web Semântica é uma extensão da web atual, onde a informação possui um significado claro e bem definido, possibilitando uma melhor interação entre computadores e pessoas. Nesse contexto, a Web Semântica surge com o propósito de fornecer tecnologias que permitem a compreensão por humanos e computadores, das informações fornecidas pela Web.

A Web Semântica é muito diversificada, não apenas em métodos e objetivos sendo pesquisados e aplicados, mas também porque existem muitas perspectivas alternativas (HITZLER, 2021). Segundo Lobo et al. (2022), as tecnologias da Web Semântica desempenham um papel significativo na integração e descoberta de dados em vários domínios, incluindo a interseção com tecnologias emergentes como blockchain, aumentando a confiança e a segurança no intercâmbio de dados digitais.

A ideia básica da Web Semântica é utilizar padrões para descrever semanticamente objetos do mundo real publicados na Web e atribuir links entre eles, permitindo que um computador compreenda o significado das informações e consiga fazer descobertas de conteúdo em tempo de execução (AGHAEI et al., 2012). A arquitetura da Web Semântica define as tecnologias e padrões necessários para que os conteúdos das páginas web possam ser compreendidos por computadores, segundo a visão do W3C (2000), tal arquitetura é dividida em camadas, como apresentada na Figura 1:

As camadas Unicode e URI (Uniform Resource Identifier) estabelecem um conjunto de caracteres e fornecem meios para a identificação de objetos na Web Semântica, onde Unicode permite que textos e imagens possam ser lidos e interpretados por computadores independente de localização por meio de uma URI que fornece um endereço global para cada recurso disponível na web. A camada XML, Namespace (NS), XML Schema estabelece que se pode integrar

Figura 1 – Camadas da Web Semântica



Fonte: Adaptado de Lima e Carvalho (2004).

definições da Web Semântica com outros padrões baseados em Extensible Markup Language (XML). A camada RDF e RDF Schema é a camada onde se tem a capacidade de oferecer tipos para recursos e links, pois permite declarações sobre objetos com URIs e vocabulários (LIMA; CARVALHO, 2004).

A camada de Ontologia fornece um vocabulário compartilhado e suporta a evolução de vocabulários assim como pode definir relações entre conceitos diferentes (LIMA; CARVALHO, 2004), deste modo permite que máquinas possam analisar dados e realizar inferências.

A camada de Assinatura Digital visa garantir a procedência de um documento, o que é fundamental para decidir se a informação é confiável ou não, onde blocos de dados criptografados são utilizados para garantir a autenticidade das fontes e a confiabilidade das informações consultadas pelos agentes [1].

A camada da Lógica tem como objetivo especificar a escrita de regras, para facilitar a construção de inferências, as quais os agentes poderão utilizar para relacionar e processar informações, executando serviços inteligentes. Por fim, a camada da Prova executa estas regras, onde os agentes têm mais poder para raciocinar sobre conceitos e relacioná-los na camada de ontologia com o propósito de provar que as informações trocadas são verdadeiras. Em conjunto com a camada de Validação e utilizando assinatura digital, os mecanismos que permitem às aplicações confiar ou não nas provas realizadas (LIMA; CARVALHO, 2004).

[1] Os agentes são programas com a função de coletar conteúdos de várias fontes, processar estas informações e compartilhar os resultados com outros programas.

2.1.1 Ontologias

Para Berners-Lee et al. (2001) os computadores necessitam ter acesso a coleções estruturadas de informações e de conjuntos de regras de inferência que auxiliem no processo de dedução automática para o raciocínio automatizado, ou seja, a representação do conhecimento.

As ontologias desempenham um papel central no contexto da Web Semântica como um veículo principal para integração, compartilhamento e descoberta de dados (HITZLER, 2021). Recentemente, a importância das ontologias destaca-se na resolução de problemas de interoperabilidade semântica em sistemas, facilitando a integração de diferentes fontes de dados e melhorando a precisão das inferências realizadas por sistemas automatizados (TIWARI et al., 2023).

A utilização de ontologias é uma das maneiras de se construir uma relação organizada entre termos dentro de um domínio, favorecendo o compartilhamento da mesma estrutura de informações. Isso permite a interpretação dos dados pelas ferramentas de recuperação da informação. Na Web Semântica, uma ontologia é uma conceitualização parcial de um dado domínio do conhecimento, compartilhado por uma comunidade de usuários, que tem sido definido em uma linguagem formal e processável por máquina para a proposta explícita de compartilhar informações semânticas de dados por meio de sistemas automatizados (JACOB, 2003).

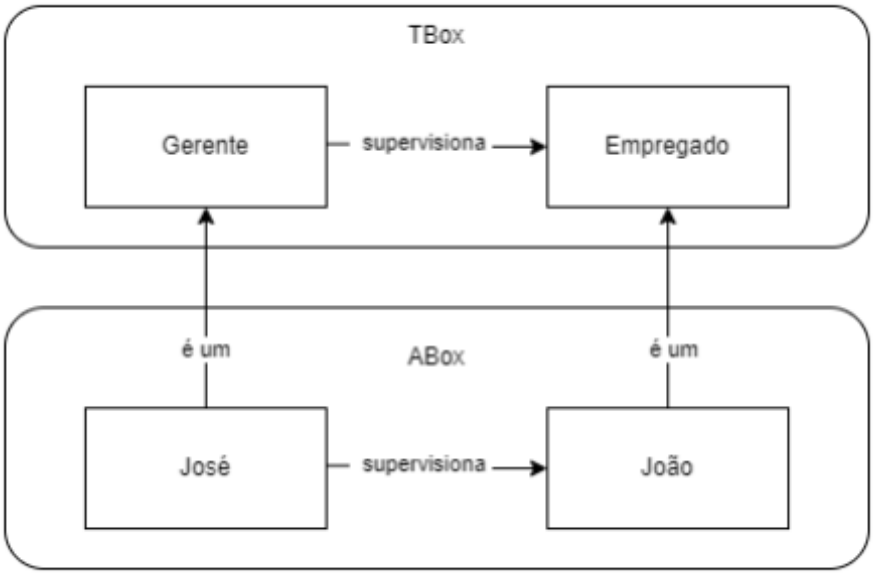
Uma ontologia para a Web Semântica é composta por duas partes principais: a Terminology Box (TBox) e a Assertion Box (ABox), tal como descrito por Noy e McGuinness (2001). TBox é a parte da ontologia que contém a definição dos conceitos e relações do domínio, incluindo classes, propriedades e axiomas. TBox é a descrição dos tipos de entidade e relações que existem no mundo representado pela ontologia. Ela serve como uma espécie de dicionário conceitual, onde são definidos os termos e conceitos utilizados no domínio em questão.

Por outro lado, segundo Noy e McGuinness (2001), ABox é a parte da ontologia que contém as instâncias dos conceitos e relações do domínio. Ela é a estrutura de fatos da ontologia. ABox é uma descrição dos indivíduos e suas relações no mundo representado pela ontologia, ela contém informações específicas sobre os objetos do domínio.

Considere uma ontologia de exemplo para representar informações sobre uma empresa. A TBox pode conter classes como “Empregado” e “Gerente”, e como propriedades “é um"e “supervisiona”. Essas classes e propriedades são conceitos gerais relacionados ao domínio de animais. A ABox pode conter instâncias dessas classes, como “João” e “José”, e relações

entre essas instâncias, como “José é um Gerente” e “José supervisiona João”. Essas instâncias e relações são informações específicas sobre a empresa em questão. A Figura 2 ilustra tal exemplo de TBox e ABox em uma ontologia.

Figura 2 – Exemplo de TBox e ABox em uma ontologia de empresa



Fonte: Elaborado pelo autor.

Em resumo, TBox é a estrutura de conceitos da ontologia, que descreve os tipos de entidades e relações do domínio, e a ABox é a estrutura de fatos da ontologia, que contém informações específicas sobre os indivíduos e suas relações no mundo representado pela ontologia. É importante destacar que a TBox e a ABox são complementares, pois TBox fornece a estrutura conceitual necessária para entender as relações entre as informações contidas na ABox. Juntas, a TBox e a ABox fornecem uma representação completa e coerente do domínio representado pela ontologia, tornando-a uma ferramenta poderosa para a representação e processamento de informações na Web Semântica.

2.1.2 Resource Description Freamework

A tecnologia adotada pelo W3C, para definição de padrões relacionais entre os dados, é o Resource Decription Framework (RDF), considerado como a base fundamental da Web Semântica. Sendo possível fazer várias ligações e inferências utilizando os meios disponibilizados neste padrão (DONG et al., 2014). Corresponde uma linguagem baseada em XML com objetivo de promover uma padronização para descrever semanticamente dados na Web. (LASSILA et al., 1998).

O XML possibilita a adição de estruturas ao documento, porém não tem a capacidade de atribuir significado a tais estruturas. Diante desta limitação, o RDF impõe uma estrutura que proporciona a expressão não ambígua da semântica e, desse modo, possibilita a codificação, o intercâmbio e o processamento consistente de metadados padronizados (MILLER, 1998).

Furgeri (2006) destaca que o RDF

> elimina o problema da representação da informação em forma de árvore, criando uma estrutura mais flexível em forma de grafos, possibilitando a formação de uma cadeia de informações e estabelecendo uma rede de conhecimento.

O RDF tem como estrutura o chamado grafo RDF, que é composto por um conjunto de triplas subdividas em sujeito, predicado e objeto, onde o predicado evidencia um relacionamento entre sujeito e objeto. O grafo RDF é um grafo direcionado, onde a direção do arco é significativa: sempre aponta para o objeto (W3C, 2004). A Figura 3 ilustra um exemplo de uma tripla de um grafo RDF sobre o contexto de empresa apresentado na Seção 2.1.1:

Figura 3 – Tripla de um grafo RDF de uma empresa

Fonte: Elaborado pelo autor

Resource Decription Framework Schema (RDFS) é construído sobre o modelo RDF básico, de forma a complementá-lo por meio de mecanismos usados para descrever grupos de recursos relacionados e os relacionamentos entre esses recursos (W3C, 2014). Essa linguagem estende o poder expressivo dos modelos RDF, permitindo que o significado dos objetos seja descrito usando o mesmo modelo que descreve os dados. Por exemplo, o RDFS permite que os recursos sejam instâncias de uma ou mais classes. Uma classe é uma abstração que agrupa recursos com características semelhantes e pode ser organizada hierarquicamente.

A principal vantagem do RDF é sua capacidade de criar uma rede de conhecimento interligada, permitindo inferências e descobertas automáticas de novos dados a partir de dados existentes (TANG et al., 2022).

2.1.3 SPARQL

SPARQL Protocol and RDF Query Language (SPARQL) é a linguagem padrão para consulta e manipulação de dados RDF na Web semântica. Ele permite a execução de consultas complexas e distribuídas, utilizando uma estrutura baseada em triplas que facilita a recuperação e integração de dados de diversas fontes (GUAN; LIANG, 2023).

De acordo com W3C (2013) o padrão inclui as seguintes especificações: uma linguagem de consulta para RDF; uma especificação que define extensões para a linguagem de consulta SPARQL para executar consultas distribuídas em diferentes terminais SPARQL; uma especificação que define a semântica de consultas SPARQL sob mecanismos de ligação, como RDFS; um acordo define métodos para passar consultas SPARQL arbitrárias e solicitações de atualização para serviços SPARQL; e por fim uma especificação que define métodos de pesquisa e descoberta; e um vocabulário para descrever serviços SPARQL e suítes de teste.

Os parâmetros da consulta SPARQL são definidos através de triplas do tipo sujeito, predicado e objeto, as quais vão de encontro com a estrutura das triplas RDF, porém esses padrões podem ser variáveis onde, dentro da consulta, atuam como espaços reservados que estão associados a termos RDF para construção de uma solução. A linguagem especifica quatro variações de consulta para diferentes finalidades (PRUD'HOMMEAUX, 2008):

- SELECT: Usado para extrair valores brutos de um endpoint SPARQL, e retorna (em um formato de tabela) os valores ligados a todas ou um subconjunto das variáveis nas soluções encontradas para consulta;
- CONSTRUCT: Utilizado para extrair informações do endpoint SPARQL e retorna um grafo RDF válido construído através da substituição de variáveis em um conjunto de templates de tripla;
- ASK: Utilizado para fornecer um resultado simples para uma consulta em um endpoint SPARQL. Isto é, retorna um valor de verdadeiro ou falso indicando se um padrão de tripla possui algum resultado;
- DESCRIBE: Utilizado para extrair um grafo RDF do endpoint SPARQL, o conteúdo do que é deixado para o endpoint para decidir com base no que o mantenedor considerar informações como útil.

A Figura 4 representa um exemplo de uma consulta SPARQL sobre grafos RDF para consulta de informações à respeito de livros. Inicialmente Prud'hommeaux (2001) define as fontes dos dados, usando três prefixos: "rdf:" é o prefixo para o namespace do RDF, "dc:" é

o prefixo para o namespace do Dublin Core (DC), e ":" é um prefixo personalizado para um namespace específico. A cláusula SELECT especifica as variáveis que desejamos selecionar como resultado da consulta. Neste caso, as variáveis são ?book e ?title. A cláusula WHERE especifica as condições que devem ser atendidas para que uma solução seja incluída no resultado da consulta. As quatro linhas dentro da cláusula WHERE, são as condições para selecionar os livros e títulos:

- ?book rdf:predicate dc:title. Esta linha especifica que o livro tem um predicado de título, usando o namespace do DC para o título.
- ?book rdf:subject ?t . Esta linha especifica que o livro tem um sujeito, com a variável ?t.
- ?book rdf:object ?title . Esta linha especifica que o livro tem um objeto, com a variável ?title.
- ?t saidBy "Bob". Esta linha especifica que o sujeito ?t é dito por "Bob".

A consulta seleciona todos os livros cujo título é especificado e cujo sujeito é dito por "Bob".

Figura 4 – Exemplo de uma consulta SPARQL sobre grafos RDF.

```
PREFIX rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dc:   <http://purl.org/dc/elements/1.1/>
PREFIX :     <http://example/ns#>

SELECT ?book ?title
WHERE
{ ?t rdf:subject    ?book  .
  ?t rdf:predicate  dc:title .
  ?t rdf:object     ?title .
  ?t :saidBy        "Bob" .
}
```

Fonte: Prud'hommeaux (2001)

2.2 Mashup de Dados

Segundo Thor et al. (2007), mashups são aplicações web interativas que combinam conteúdo de múltiplos serviços ou fontes em um novo serviço ou fonte de dados. Um mashup de dados é uma visão materializada construída através da transformação e integração de dados presentes em diferentes bases.

O processo de criação de mashup é uma tarefa complexa se as fontes de dados não foram integradas semanticamente, tendo como principais desafios: seleção das fontes de dados, extração e tradução dos dados provenientes de fontes distintas e possivelmente heterogêneas para

uma vocabulário comum, identificação de relacionamentos entre recursos de diferentes fontes e combinação e fusão de múltiplas representações de um mesmo objeto em uma representação concisa e unificada, e resolução das inconsistências existentes para melhorar a qualidade dos dados (VIDAL et al., 2015).

De acordo com Bleiholder e Naumann (2009), o mashup de dados é dividido em três etapas:

- Integração de esquema: O objetivo dessa etapa é diminuir a quantidade de comparações entre os recursos de diferentes datasets, comparando apenas recursos de elementos iguais ou similares do esquema.
- Detecção de duplicação: O objetivo dessa etapa é identificar e ligar recursos que correspondem ao mesmo objeto do mundo real.
- Fusão: É o processo de combinação das propriedades de múltiplos recursos que representam o mesmo objeto na realidade. Isso resulta em uma descrição mais completa e precisa do objeto, onde os conflitos e irregularidades nas propriedades das entidades são resolvidos. A maior dificuldade desta tarefa é aplicar eficientemente a estratégia de resolução de conflitos mais adequada para as propriedades a serem utilizadas.

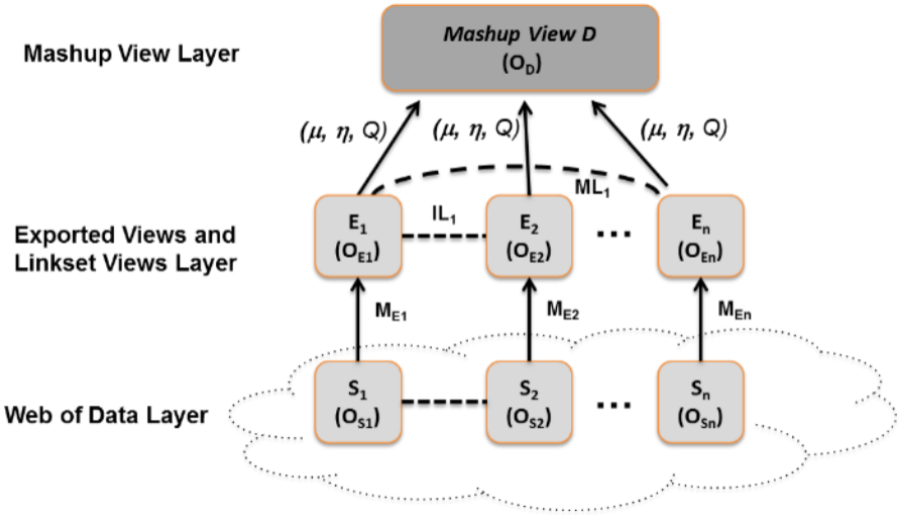
Vidal et al. (2015) apresenta um framework baseado em ontologia para especificar visões de mashup, onde tal especificação é definida por uma tupla $\lambda = (D, O_D, E_i, IL_i, ML_i, \mu, \eta, Q)$, sendo:

- D o nome da visão de mashup;
- O_D a ontologia da visão de mashup;
- E_i são especificações das visões exportadas com ontologias O_{E_i}, cujos vocabulários são subconjuntos do vocabulário O_D;
- IL_i e ML_i são especificações das visões de conjuntos de ligações exportadas entre E_i e E_n;
- μ é um conjunto de regras de fusão de O_{E_i} para O_D;
- η é um símbolo da função de normalização, cujo a interpretação definida como remapear IRIs [2] de visões exportadas para IRIs das visões;
- Q é um conjunto de métricas de avaliação de qualidade das fontes de dados.

A Figura 5 descreve a arquitetura do framework de Vidal et al. (2015), este possuindo três camadas: Visão de Mashup (Mashup View), Visões Exportadas e Visões de Conjunto de Ligações (Exported Views and Linkset Views) e Web de Dados (Web of Data).

[2] Internationalized Resource Identifier, definido pela Internet Engineering Task Force (IETF), como um novo padrão para estender o esquema existente do URI (GANGEMI; PRESUTTI, 2006).

Figura 5 – Três Camadas do Framework Baseado em Ontologia



Fonte: Vidal et al. (2015)

Na camada Visão de Mashup, a ontologia da visão de mashup (O_D) especifica os conceitos da aplicação de mashup, o qual é o vocabulário comum para integração de dados exportados pelas fontes de dados.

Na camada Web de Dados, cada fonte de dados (S_i) é descrita por uma ontologia fonte (O_{S_i}) e exporta uma ou mais visões (E_i), chamadas de Visões Exportadas.

Na camada Visões Exportadas e Visões de Conjunto de Ligações, cada uma dessas visões E_i tem uma ontologia (O_{E_i}) e um conjunto de regras (M_{E_i}), que mapeiam conceitos de O_{S_i} em O_{E_i}. Também são definidas as regras para descobertas de Links Semânticos (IL_i e/ou ML_i), que definem a similaridade entre duas representações distintas de um mesmo objeto do mundo real.

Compreendido os conceitos fundamentais sobre Web Semântica e o processo de construção de mashup de dados , no próximo capítulo serão apresentados trabalhos correlatos à presente pesquisa.

3 TRABALHOS RELACIONADOS

Nesta seção, serão apresentados alguns trabalhos relacionados ao tema da construção de mashups de dados utilizando tecnologias de Web Semântica. A literatura na área tem explorado diversas abordagens para integrar e reutilizar dados de múltiplas fontes, proporcionando uma visão unificada e enriquecida semanticamente.

Os trabalhos revisados a seguir foram escolhidos mediante à relevância e contribuição para o campo. Eles abordam metodologias estruturadas para a construção e manutenção de grafos de conhecimento comercial e frameworks semânticos que facilitam a criação de mashups de dados por usuários não técnicos. A revisão desses trabalhos fornece uma base sólida para a compreensão de técnicas e ferramentas atualmente disponíveis. Além disso, eles destacam práticas e os problemas enfrentados no processo de criação de soluções semânticas para a integração de dados.

3.1 MAURA: Um Framework baseado em Mediador Semântico para construção eficiente de Linked Data Mashups

Cavalcante (2017) propõe o framework MAURA, baseado em mediador semântico para construção e reutilização de Linked Data Mashups (LDMs). Esse framework permite que usuários criem mashups personalizados sem a necessidade de conhecimentos específicos em Integração de Dados ou Web Semântica.

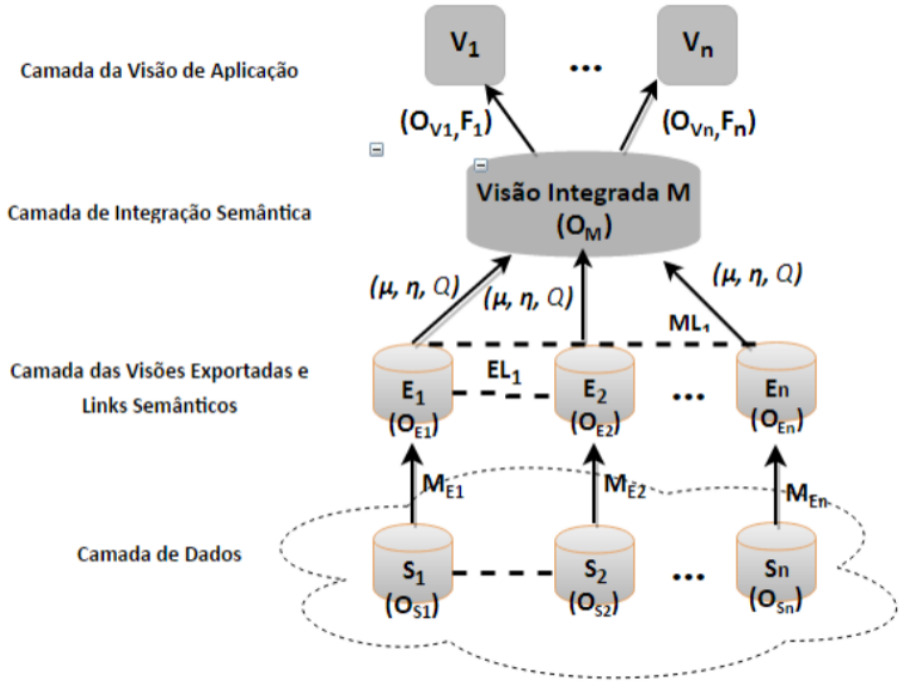
MAURA reutiliza uma especificação de LDM na criação de novos mashups, materializando apenas os dados relevantes para o usuário. Um mediador convencional reescreve uma consulta em subconsultas, executa-as sobre diversas bases de dados, trata os dados e retorna uma visão materializada. No mediador semântico proposto, o processo de reescrita é realizado sobre a especificação de um mashup. Essa especificação é usada para construir mashups virtuais em tempo de execução, materializados apenas quando requisitados pelo usuário, chamados de Visões de Aplicação de Mashup.

A arquitetura do framework MAURA é dividida em quatro camadas, conforme ilustrado na Figura 6:

- Camada de Dados: A camada de dados constitui a base da arquitetura, onde os dados brutos de diferentes fontes são armazenados. Esta camada pode incluir bases de dados relacionais, documentos RDF e outras fontes de dados que precisam ser integradas;

- Camada de Visões Exportadas e Links Semânticos: Nesta camada, os dados brutos são transformados em visões exportadas, as quais são representações semânticas dos dados de uma fonte específica, conforme a ontologia de domínio. Além disso, são estabelecidos links semânticos entre diferentes visões exportadas, conectando dados que representam o mesmo objeto do mundo real, mesmo que provenientes de fontes distintas;
- Camada de Integração Semântica: No topo da arquitetura, a camada de integração semântica coordena a combinação das visões de aplicação com as ontologias de domínio e as regras de integração. Esta camada realiza a mediação semântica e a materialização dos dados, assegurando que as visões de aplicação sejam consistentes e coerentes com a semântica definida. A camada de integração semântica permite a criação eficiente de novos mashups, reutilizando especificações existentes e incrementando-as com novas fontes de dados conforme necessário;
- Camada de Visão de Aplicação: A camada de visão de aplicação permite a especificação e criação de visões personalizadas dos dados integrados, ajustadas aos requisitos dos usuários finais. Nesta camada, os usuários podem definir parâmetros específicos, como filtros e condições, para criar visões que atendam às suas necessidades. As visões de aplicação são geradas a partir das visões exportadas e dos links semânticos, garantindo que apenas os dados relevantes sejam apresentados.

Figura 6 – Arquitetura de 4 Camadas do Mediador Semântico.



Fonte: Cavalcante (2017).

Para realizar o processo de construção de uma visão de aplicação em tempo de execução, o mediador segue os seguintes passos:

- Especifica-se um LDM (M) sobre as fontes heterogêneas, onde é armazenada a especificação previamente criada em um formato próprio;
- Usuário define os parâmetros para construção da visão de aplicação, sendo esses uma ontologia (O_{vi}) e um conjunto de filtros (F_{vi});
- O mediador aplica os parâmetros na especificação existente, gerando uma nova;
- O mediador materializa a especificação gerada e retorna ao usuário, como uma visão de aplicação de mashup (V_i).

Existem diversas vantagens associadas a essa abordagem. Primeiramente, o processo de integração semântica das fontes de dados é realizado apenas uma vez e pode ser atualizado sempre que necessário. Além disso, apenas as informações especificadas pelo usuário são materializadas, economizando recursos computacionais. Outra vantagem é que não é preciso possuir conhecimentos específicos em Web Semântica, pois o uso de parâmetros permite que usuários comuns criem mashups de forma transparente, sem precisar entender o processo de integração semântica. A especificação de um mashup pode ser disponibilizada na Linked Open Data, permitindo a reutilização por outros pesquisadores.

Este framework foi aplicado no projeto Governança Inteligente dos Sistemas de Saúde (GISSA) no contexto da Rede Cegonha, um programa do Ministério da Saúde do Brasil. A Rede Cegonha tem como objetivo melhorar a saúde materno-infantil e reduzir a mortalidade infantil no Brasil, oferecendo cuidados abrangentes e contínuos às gestantes e recém-nascidos. Em Tauá, Ceará, o MAURA foi utilizado para integrar dados de saúde materno-infantil, enfrentando o desafio de consolidar informações provenientes de diferentes bases do SUS, como o Sistema de Informações sobre Nascidos Vivos (SINASC) e o Sistema de Informação em Saúde para a Atenção Básica (e-SUS).

A aplicação prática do MAURA na Rede Cegonha resultou em melhorias significativas na gestão de dados de saúde materno-infantil. Os gestores de saúde de Tauá puderam usar os mashups semânticos para identificar áreas de alto risco para intervenções prioritárias. Com dados integrados e semanticamente enriquecidos, foi possível planejar intervenções mais eficazes, alocando recursos de forma mais direcionada e eficiente. A análise detalhada dos dados permitiu a tomada de decisões mais informadas, contribuindo para a redução da mortalidade infantil e a melhoria geral da saúde materno-infantil.

3.2 SISIFO: Uma Abordagem Semântica para Construção de Enterprise Knowledge Graphs

Rolim (2020) apresenta uma metodologia que usa camadas semânticas para construir e manter um Enterprise Knowledge Graphs (EKGs). O objetivo deste trabalho é resolver problemas comuns com a integração e a gestão de dados empresariais, bem como fornecer uma abordagem estruturada e eficaz para a criação de EKGs, que visa organizar e tornar o conhecimento mais acessível dentro de uma empresa.

A metodologia SISIFO é dividida em várias etapas. Primeiro, são definidas questões de competência que orientam o escopo e os objetivos do EKG . Em seguida, uma ontologia é modelada representando os conceitos e relações do domínio do EKG . Posteriormente, cada ontologia é especificada em detalhes, incluindo mapeamentos e metadados. Finalmente, as visões do EKG são publicadas e validadas por meio de consultas e estudos de caso para garantir que o EKG esteja corretamente estruturado e atenda às necessidades informacionais do contexto corporativo.

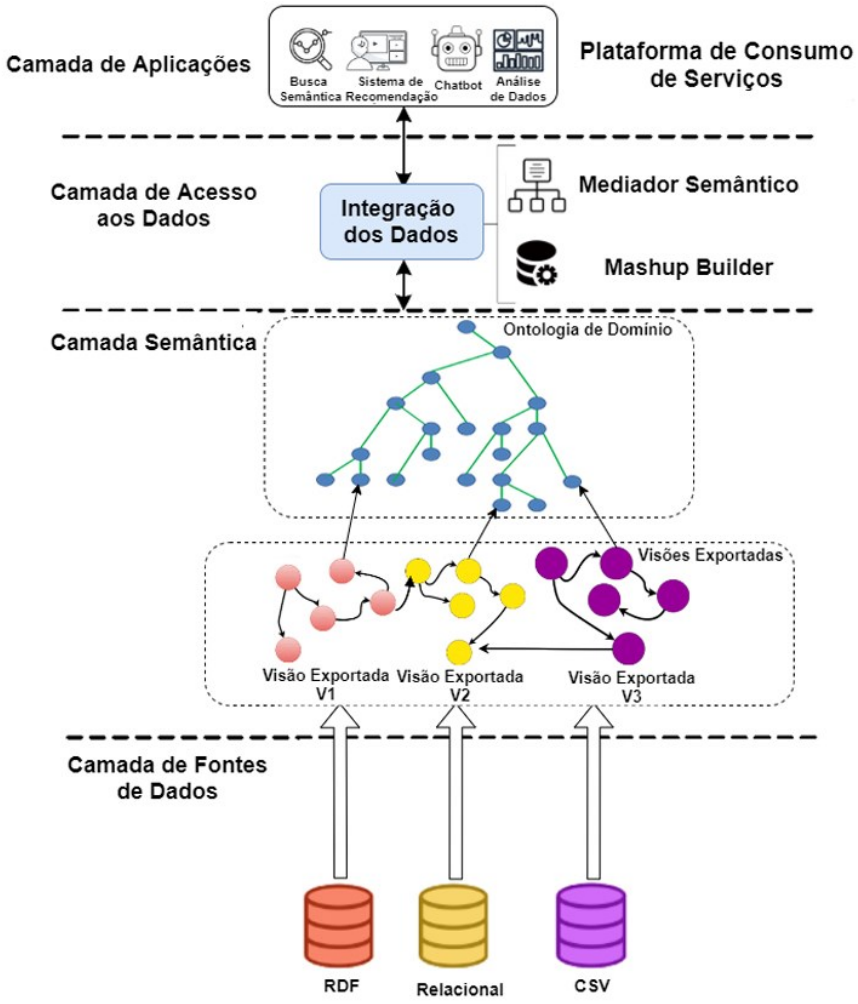
A arquitetura de SISIFO é dividida em quatro camadas principais, conforme ilustrado na Figura 7:

- Camada de Fontes de Dados: Abrange diversas fontes de dados presentes na empresa, que podem estar em diferentes formatos como bancos de dados relacionais, arquivos CSV, documentos JSON, etc.
- Camada Semântica: Publica os dados por meio de uma visão semântica unificada usando uma ontologia de domínio, permitindo consultas SPARQL para acesso aos dados integrados.
- Camada de Acesso aos Dados: Inclui o Mediador Semântico e o Construtor de Mashups, que facilitam a consulta e integração dos dados semanticamente.
- Camada de Aplicações: Permite a construção de aplicações que consomem os dados integrados, proporcionando uma plataforma self-service para usuários finais.

A aplicação prática desta metodologia foi testada no domínio fiscal com empresas, parceiros e contribuintes. Os resultados mostraram que SISIFO é eficaz na construção e gestão de EKGs e fornece uma maneira formal e estruturada de especificar e representar semanticamente os grafos de conhecimento. A abordagem semi-automática proposta por Rolim (2020) torna o processo mais eficiente e menos propenso a erros.

Os benefícios das técnicas de Web Semântica e integração de dados são inúmeros. A

Figura 7 – Arquitetura de SISIFO para construção de Enterprise Knowledge Graphs.



Fonte: Rolim (2020).

Web Semântica organiza e torna o conhecimento mais corporativo ao formalizar e representar conceitos e suas relações. A integração de dados unifica semanticamente dados de fontes heterogêneas e fornece uma visão única do conhecimento. Além disso, a metodologia de Rolim (2020) reduz o processo de construção e manutenção de EKGs e erros humanos, e permite a reutilização e extensão de dados.

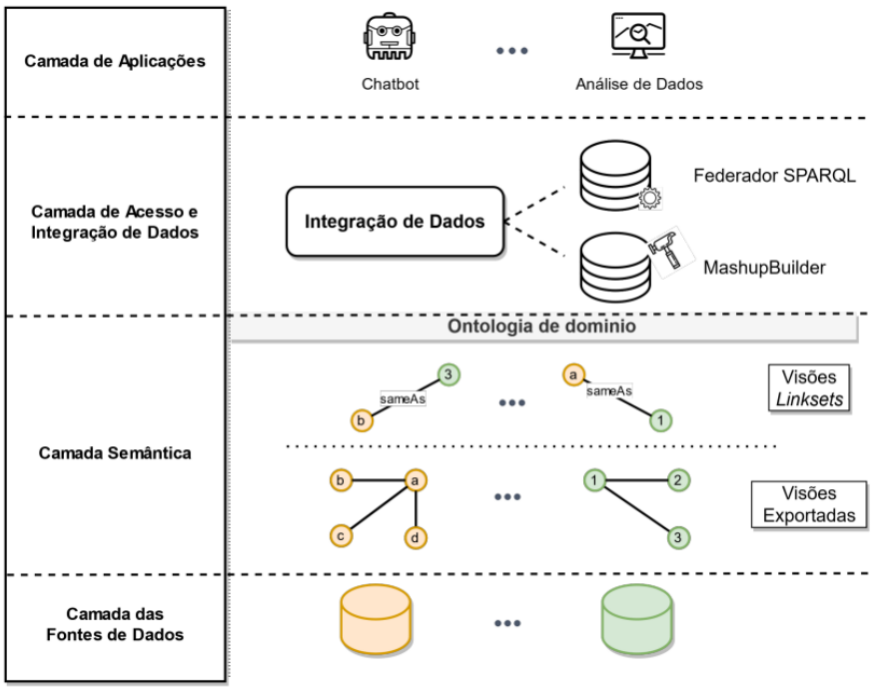
3.3 Uma abordagem para construção de Mashup de Dados especificados como uma visão sobre um EKG

Cruz (2021) apresenta uma abordagem semi-automática para construção de um mashup de dados como uma visão sobre um EKG. A arquitetura de EKG considerada por este trabalho é dividida em quatro camadas de abstração, onde a camada inferior representa um nível de abstração menor em relação a camada superior, sendo elas:

- Camada das Fontes de Dados: Composta pelas fontes de dados selecionadas para serem integradas pelo EKG.
- Camada Semântica: É responsável por resolver os problemas de interoperabilidade existentes entre as fontes de dados e explicitar as ligações existentes entre os recursos.
- Camada de Acesso e Integração de Dados: Fornece o acesso a uma visão integrada sobre visões publicadas pela camada semântica e, consequentemente, uma visão integrada sobre as fontes de dados.
- Camada de Aplicações: Nessa última camada, pode-se ter aplicações de busca semântica, chatbots, sistemas de recomendação, ferramentas de análise de dados que realizam suas requisições de informação em termos da ontologia de domínio.

A Figura 8 representa a arquitetura EKG descrita da camada com menor nível de abstração, Camada das Fontes de Dados, até a camada com maior nível de abstração, Camada de Aplicações.

Figura 8 – Arquitetura de um EKG implementado a partir de uma visão semântica.



Fonte: Cruz (2021).

O processo de construção de um mashup sobre um EKG foi dividido em três etapas: Especificação da visão de mashup como uma consulta facetada, Decomposição da visão de mashup sobre a visão semântica do EKG, e Materialização da visão de mashup (Construção do plano de consulta para recuperação dos sujeitos relevantes para Visão de Mashup de Dados,

Execução da consulta para recuperação dos sujeitos relevantes, Construção e execução das consultas para extração das propriedades dos sujeitos relevantes, Definição e aplicação das regras de fusão).

Além da especificação de consulta facetada e regras de fusão, o restante do processo pode acontecer automaticamente, resultando em um mashup de dados, que pode ser entendido como um KG (Knowledge Graph), que possui a estrutura definida pela visão de mashup.

Cruz (2021) utilizou essa abordagem no portal SemanticSUS, que integra dados do Sistema Único de Saúde (SUS) para facilitar a análise e exploração de dados de saúde, permitindo a criação de mashups para estudos específicos, como a mortalidade infantil e neonatal. A abordagem mostrou-se eficaz na integração de dados heterogêneos e na criação de uma estrutura unificada que facilita a interpretação e reutilização dos dados.

3.4 Comparação e Relevância para a Pesquisa

Cada um desses trabalhos traz contribuições essenciais para a construção da solução proposta nesta pesquisa. No contexto do Observatório de Dados Abertos, o uso de mashups semânticos e a criação de grafos de conhecimento são cruciais para possibilitar a interoperabilidade entre diferentes fontes de dados, promovendo uma integração eficiente. Cavalcante (2017) e Cruz (2021) se destacam por suas aplicações no domínio da saúde, enquanto Rolim (2020) oferece uma visão mais ampla e estruturada para a criação e manutenção de grafos de conhecimento.

Essas abordagens oferecem diferentes arquiteturas para integrar dados heterogêneos usando ontologias, RDF e SPARQL, permitindo consultas complexas e facilitando a descoberta de novos conhecimentos a partir da integração semântica. A Tabela 1 resume as principais características de cada trabalho relacionado e sua relevância para esta pesquisa.

Tabela 1 – Comparação dos Trabalhos Relacionados

Critério	Cavalcante (2017)	Rolim (2020)	Cruz (2021)
Mashups semânticos	X		X
Grafos de Conhecimento		X	X
Tecnologias RDF/SPARQL	X	X	X
Aplicação em Saúde	X		X
Foco na Integração de Dados	X	X	X

A relevância dessas técnicas para o trabalho atual está na capacidade de oferecer

uma solução para integrar dados heterogêneos, explorando a Web Semântica para garantir a interoperabilidade, transparência e reutilização dos dados públicos, que são os principais objetivos desta pesquisa.

4 METODOLOGIA

A crescente evolução e popularização das tecnologias de informação e comunicação tem contribuído diretamente para geração e disponibilização de dados, porém em formato pouco amigável para recuperação e consulta pela população e para interligação com outros datasets. Um exemplo é o trabalho de Alves (2022) que apresenta o Observatório de Dados Abertos da Região do Sertão dos Crateús, uma plataforma cujo objetivo é ser um mecanismo facilitador de disponibilização de dados abertos, para a população da região.

Atualmente na pesquisa de Alves (2022) são apresentados dados relacionados à COVID-19 e acerca da mortalidade na região por doenças catalogadas em uma lista de tabulação do Brasil com base na Classificação Internacional de Doenças, 10ª Revisão (CID-10)[1], utilizando como fonte de dados o repositório o covid19br[2] e a plataforma Mortalidade no Estado do Ceará[3].

Ao acessar a plataforma é possível verificar dados de algumas instituições governamentais, contudo todo o contexto de ligações semânticas, o uso de tecnologias e conceitos da Web Semântica discutidos nos capítulos anteriores não estão presentes. Desta forma, para atingir os objetivos da presente pesquisa, a seguir serão apresentados os elementos metodológicos pretendidos.

4.1 Arquitetura da Solução

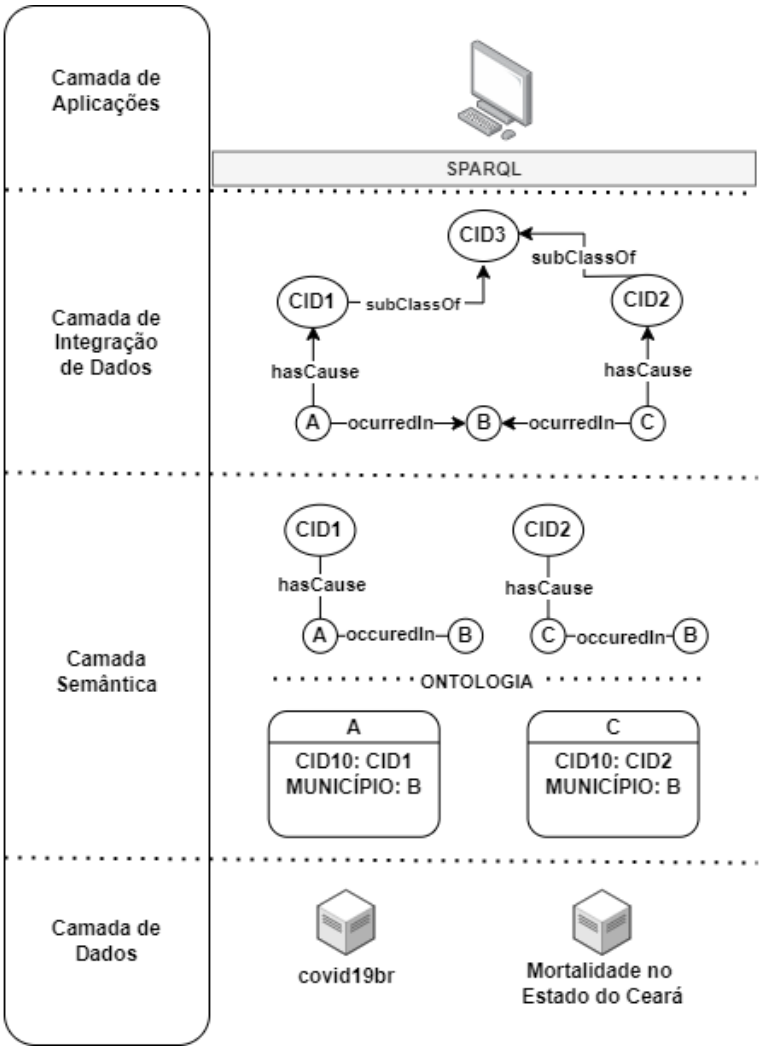
A arquitetura proposta por este trabalho, baseada nas apresentadas por Cavalcante (2017), Rolim (2020) e Cruz (2021), é dividida em quatro camadas: Camada das Fontes de Dados, Camada Semântica, Camada de Integração de Dados e Camada de Aplicações, onde cada camada representa um diferente nível de abstração. A Figura 9 representa uma visão geral dessa arquitetura.

4.1.1 Camada de Dados

A camada de dados é composta pelas fontes de dados selecionadas para serem integradas. O repositório covid19br disponibiliza um grande quantidade de dados, em formato CSV, relacionados à COVID-19 no Brasil, como número de casos e óbitos por município, provenientes da junção dos dados oficias de todas as secretarias de saúde de cada estado. A

[1] http://tabnet.datasus.gov.br/cgi/sih/mxcid10lm.htm
[2] https://github.com/wcota/covid19br
[3] http://extranet.saude.ce.gov.br/tabulacao/deftohtm.exe?sim/obito.def

Figura 9 – Arquitetura proposta



Fonte: Elaborado pelo autor.

plataforma Mortalidade no Estado do Ceará, assim como o repositório, disponibiliza um grande quantidade de dados relacionados à COVID-19 provenientes da junção dos dados oficias de todas as secretarias de saúde, com o diferencial de que as informações disponibilizadas, em formato tabulado, estarem relacionadas as doenças catalogadas no CID-10.

Na Figura 9, essa camada é representada na parte inferior, onde são ilustradas as fontes de dados "covid19br"e "Mortalidade no Estado do Ceará". Estas fontes fornecem os dados brutos que serão processados e integrados nas camadas superiores.

4.1.2 Camada Semântica

A camada semântica é responsável por resolver os problemas de interoperabilidade existentes entre as fontes de dados. Cruz (2021) considera para esta camada os seguintes componentes: ontologia de domínio, um conjunto de visões exportadas e um conjunto de visões

linkset, mas para este trabalho serão considerados apenas os dois primeiros para esta camada, deixando a Camada de Integração de Dados responsável por explicitar as ligações existentes entre os recursos.

A ontologia de domínio é parte crucial para resolução dos problemas de interoperabilidade, sendo responsável por estabelecer um vocabulário comum entre as informações, como visto no capítulo 2. As visões exportadas são visões RDF definidas sobre as fontes de dados especificadas na camada anterior, sendo uma ou mais visões por fonte de dados.

Na Figura 9, essa camada é representada pelo nível intermediário onde as informações das fontes de dados são estruturadas semanticamente. Aqui, as instâncias "CID1"e "CID2"são mostradas com suas respectivas relações hasCause, e occurredIn associadas ao município "B". Isso demonstra como os dados brutos são transformados em uma representação semântica utilizando uma ontologia.

4.1.3 Camada de Integração de Dados

A camada de integração de dados é responsável por definir links existentes entre recursos de diferentes visões exportadas provenientes da Camada Semântica. Esses links têm a função de explicitar uma relação de equivalência entre dois recursos distintos, indicando que aqueles recursos se referem a um mesmo objeto ou estão correlacionados.

Na Figura 9, essa camada é ilustrada pelo nível onde "CID1"e "CID2"são integrados com a ajuda da ontologia, mostrando que ambas as classes são subclasses de "CID3". Além disso, as propriedades hasCause e occurredIn são usadas para conectar instâncias a causas e locais, respectivamente, estabelecendo uma rede de informações que facilita a análise cruzada de dados de diferentes fontes. Por exemplo, uma instância de "CID1"pode ter a propriedade hasCause ligando-a a um agente causador específico, enquanto occurredIn pode mapear essa instância a um município específico. Isso permite a organização hierárquica e a integração dos dados, permitindo que as relações entre diferentes fontes de dados sejam claramente definidas e mapeadas.

4.1.4 Camada de Aplicações

A última camada é composta pelas aplicações que podem dispor de consultas SPARQL, definidas sobre um Mashup de dados. O grafo de conhecimento fornecido pela Camada de Integração de Dados pode ser consultado e os dados resultantes de tais consultas são

consumidos e disponibilizados.

Na Figura 9, essa camada é representada no topo, onde um computador executa consultas SPARQL sobre a ontologia integrada. Isso demonstra como as aplicações podem utilizar essas consultas para extrair informações valiosas dos dados integrados, permitindo análises avançadas e suportando a tomada de decisões.

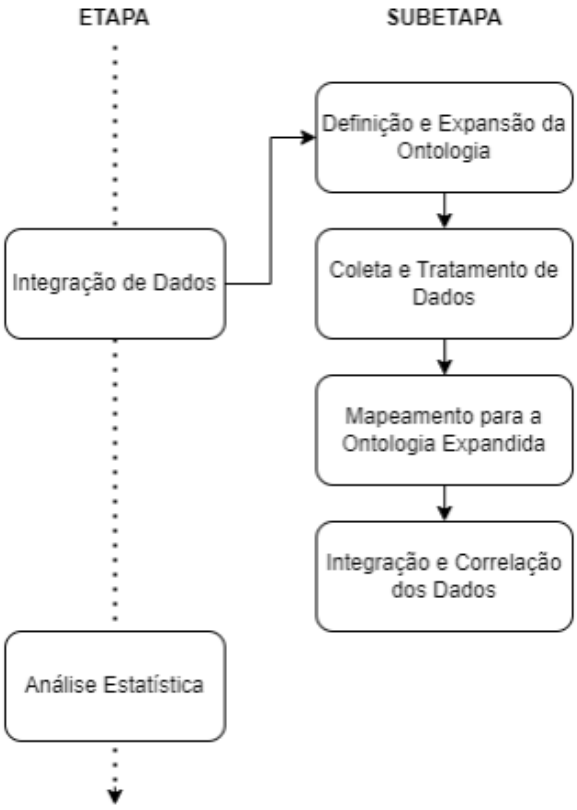
Diante do que foi apresentado, a arquitetura proposta é estabelecida como a estrutura base para a próxima fase do trabalho, servindo como um guia para a análise que se segue, fornecendo um caminho estruturado para investigar as relações entre a COVID-19 e a mortalidade por outras doenças respiratórias. A seguir será apresentado o processo de desenvolvimento e seus resultados, onde a utilidade e a eficácia da arquitetura serão exploradas através da coleta, integração e análise de dados na Região do Sertão dos Crateús.

5 ABORDAGEM PROPOSTA

Este trabalho adota uma abordagem predominantemente quantitativa, complementada por elementos qualitativos na fase de definição e expansão da ontologia. A escolha de uma abordagem quantitativa se justifica pela natureza dos dados disponíveis, os quais são predominantemente numéricos e podem ser analisados estatisticamente para identificar padrões e correlações. No entanto, a fase de expansão da ontologia envolve um processo qualitativo de definição conceitual para garantir a coerência semântica dos dados.

A abordagem é dividida em duas etapas principais: a Integração de Dados e a Análise estatística, a Figura 10 é a representação de tais etapas e suas respectivas subetapas:

Figura 10 – Etapas e subetapas da abordagem



Fonte: Elaborado pelo autor.

5.1 Integração de Dados

A integração de dados é a etapa principal na metodologia de pesquisa, envolvendo um conjunto de processos que garantem a coleta, organização e preparação dos dados para

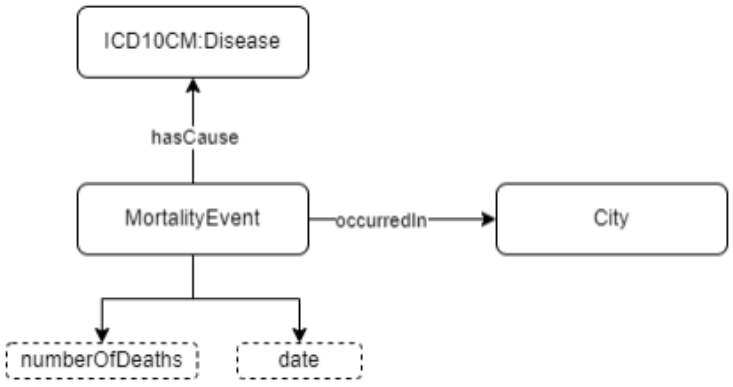
análises subsequentes. Esta fase é subdividida em quatro etapas: Definição e Expansão da Ontologia, Coleta e Tratamento de Dados, Mapeamento para a Ontologia Expandida e Integração e Correlação dos Dados, essas etapas tem como base a arquitetura proposta.

5.1.1 Definição e Expansão da Ontologia

Antes de iniciar a coleta de dados, é essencial estabelecer uma base sólida para a estruturação semântica das informações. Para isso, optou-se por reutilizar e expandir a ontologia International Classification of Diseases, Version 10 - Clinical Modification (ICD-10-CM)[1], que organiza e codifica informações de saúde de forma hierárquica através do CID-10, permitindo uma classificação detalhada e padronizada. A ferramenta Protégé[2], que é uma ferramenta específica para a criação, manipulação e manutenção de ontologias, oferecendo um ambiente robusto e especializado para esse fim, é importante para adicionar novas classes que representam eventos de mortalidade. Essa expansão é crucial para permitir a correlação entre os dados de mortalidade e outros fatores relevantes para o estudo.

A Figura 11 ilustra essa expansão, sendo a adição de duas classes: City e MortalityEvent, onde MortalityEvent possui duas propriedades de dados que representam a data do evento (date) e o número de óbitos (numberOfDeaths), e a adição de duas propriedades de objetos: hasCause que estabelece uma relação entre o evento e a sua causa, representada por uma instancia da classe Disease da ontologia ICD-10-CM, e occurredIn que estabelece uma relação entre evento e cidade, representada por uma instancia da classe City.

Figura 11 – Ontologia Expandida



Fonte: Elaborado pelo autor.

[1] https://bioportal.bioontology.org/ontologies/ICD10CM/
[2] https://protege.stanford.edu/

5.1.2 Coleta e Tratamento de Dados

Após a expansão da ontologia, a coleta de dados, representada pela Camada de Dados, é realizada por meio de scripts automatizados em Python, que extraem as informações das fontes de dados. A escolha do Python deveu-se à sua eficiência e à vasta gama de bibliotecas disponíveis para manipulação de dados, como Pandas, que é uma biblioteca poderosa para manipulação e análise de dados. Os dados coletados são submetidos a um processo de limpeza e normalização, que inclui a remoção de dados inválidos, a padronização de formatos de datas e a verificação de consistência, garantindo a qualidade e a consistência dos dados para o mapeamento subsequente.

5.1.3 Mapeamento para a Ontologia Expandida

Com os dados devidamente tratados, é realizado o mapeamento para a ontologia expandida, representada pela Camada Semântica, onde cada base de dados terá uma visão RDF exportada. A biblioteca RDFLib[3] foi utilizada para associar os dados coletados às classes e propriedades da ontologia, garantindo a integração semântica e a interoperabilidade dos dados. Esse mapeamento é essencial para a construção de um conjunto de dados coeso que possa ser efetivamente analisado.

5.1.4 Integração e Correlação dos Dados

Em posse da ontologia e das visões exportadas, inicia-se a etapa de integração e correlação dos dados, representada pela Camada de Integração de Dados. Os dados são importados para o GraphDB[4], um banco de dados de grafos especializado em dados semânticos. A escolha do GraphDB é motivada pela sua capacidade de realizar inferências, consultas complexas e pela sua compatibilidade com as tecnologias da Web Semântica, como RDF e SPARQL.

5.2 Análise estatística

A análise estatística desempenha um papel crucial na compreensão do impacto da COVID-19 na mortalidade por outras doenças na Região do Sertão dos Crateús, esta etapa é representada pela Camada de Aplicações. No GraphDB, são realizadas consultas SPARQL para

[3] https://rdflib.readthedocs.io/
[4] https://graphdb.ontotext.com/

correlacionar os dados de saúde pública, explorando as relações semânticas estabelecidas pela ontologia.

Para esta análise, foram considerados três cenários distintos, cada um representando um período específico do ano de 2020: o ano completo de 2020, os quatro primeiros meses do ano e os meses que registraram picos de óbitos por COVID-19. Essa estruturação permitiu uma investigação minuciosa dos efeitos da pandemia sobre as taxas de mortalidade, bem como a consideração de subnotificações.

A análise do ano completo de 2020 busca fornecer uma visão abrangente das tendências de mortalidade ao longo dos doze meses. Esse panorama global permite identificar se houve um aumento nas mortes por doenças respiratórias que pudesse ser associado à trajetória da pandemia, levando em conta que a sobrecarga do sistema de saúde e as mudanças no comportamento da população poderiam ter contribuído para um cenário de subnotificação.

Focando nos quatro primeiros meses do ano, de Janeiro a Abril, procura-se entender as respostas iniciais e as adaptações do sistema de saúde diante do desafio emergente da COVID-19. Esse período crítico é analisado para detectar mudanças imediatas nas taxas de mortalidade por outras doenças respiratórias.

Durante os meses de pico de óbitos por COVID-19 na região, a análise estatística concentra-se em avaliar se o aumento expressivo no número de mortes causadas por COVID-19 estava correlacionado com um aumento nas mortes por outras doenças respiratórias. A preocupação com subnotificações foi particularmente relevante nesse cenário, pois a alta demanda por serviços de saúde pode ter mascarado o verdadeiro impacto da pandemia na mortalidade geral.

Para a análise da correlação dos dados é aplicado o coeficiente de correlação de Pearson, que é uma medida estatística que expressa o grau de relação linear entre duas variáveis quantitativas. O cálculo do coeficiente de Pearson é baseado na covariância das variáveis em questão, normalizada pelos seus respectivos desvios padrão.

A fórmula para o cálculo é descrita na equação 5.1:

$$r = \rho \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5.1}$$

onde x_i e y_i são os valores individuais das variáveis X e Y, respectivamente, $\bar{x}$ e $\bar{y}$ são as médias dos valores de X e Y, e n é o número total de observações.

A interpretação do coeficiente de Pearson é direta: um valor próximo de +1 sugere que, à medida que uma variável aumenta, a outra também aumenta em uma relação proporcional. Por outro lado, um valor próximo de -1 indica que o aumento em uma variável está associado a uma diminuição proporcional na outra. Um valor próximo de 0 indica que não há uma relação linear significativa entre as variáveis.

A abordagem desta pesquisa é estruturada para investigar a aplicação de tecnologias de Web Semântica na integração e correlação de dados heterogêneos, com foco no Observatório de Dados Abertos da Região do Sertão dos Crateús. Para orientar o desenvolvimento desta pesquisa, foram formuladas as seguintes questões de pesquisa:

1. Como a integração semântica de dados pode ser utilizada para unificar informações de diferentes fontes de dados heterogêneas no contexto do Observatório de Dados Abertos da Região do Sertão dos Crateús?
2. Qual a viabilidade da utilização da Web Semântica para correlacionar dados de diferentes fontes, como os casos de COVID-19 e mortalidade por outras doenças respiratórias?
3. A metodologia de integração semântica pode revelar correlações significativas entre os dados de saúde pública da região do Sertão dos Crateús?

Essas questões guiaram as etapas metodológicas de coleta, tratamento e integração dos dados, permitindo uma análise estatística e semântica robusta, com o objetivo de responder a essas perguntas de forma clara e objetiva.

6 RESULTADOS

A presente seção relata os resultados obtidos a partir da metodologia aplicada para analisar a correlação entre o aumento dos casos de COVID-19 e a mortalidade por outras doenças respiratórias na Região do Sertão dos Crateús. Os dados foram coletados, integrados e submetidos à análise estatística conforme descrito no Capítulo 5.

Após o processo de limpeza e normalização, que incluiu a remoção de entradas incompletas, a padronização de formatos de datas e a verificação de consistência entre os registros obteve-se um total de 294.586 registros de mortalidade válidos referentes ao ano de 2020 e à região de estudo. Esses registros foram então mapeados para a ontologia expandida, utilizando a biblioteca RDFLib para associar cada evento de mortalidade às suas respectivas causas e localizações geográficas, conforme definido pelas novas classes e propriedades introduzidas. A integração dos dados foi realizada com sucesso, resultando em um conjunto de informações estruturadas e prontas para a análise estatística.

A integração semântica dos dados foi facilitada pelo uso do GraphDB, que permitiu a importação dos dados mapeados e a realização de consultas SPARQL. Isso possibilitou explorar as relações semânticas estabelecidas pela ontologia, como correlacionar eventos de mortalidade específicos com surtos de COVID-19 em diferentes cidades da Região do Sertão dos Crateús.

A consulta do Código-fonte 1 foi realizada para obter o número de óbitos causados por COVID-19 em 2020. O resultado dessa consulta pode ser visto na Tabela 2.

Código-fonte 1 – Busca de eventos causados por COVID-19 em 2020

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX icd10: <http://purl.bioontology.org/ontology/ICD10CM/>
PREFIX icd10MO: <https://www.ufc.br/ontologies/ICD10CM-MO/>

SELECT ?date (SUM(?deaths) as ?totalDeaths)
WHERE {
    ?disease rdfs:subClassOf* icd10:U07.1 .
    ?event rdf:type icd10MO:MortalityEvent ;
        icd10MO:hasCause ?disease ;
```

```
        icd10MO:numberOfDeaths ?deaths ;
        icd10MO:date ?date .
    FILTER (?date >= "2020-01-31"^^xsd:date && ?date <=
        "2020-12-31"^^xsd:date)
}
GROUP BY ?date
ORDER BY ?date
```

Tabela 2 – Número de óbitos causados por COVID-19 no ano de 2020

Mês	Número de óbitos
Janeiro	0
Fevereiro	0
Março	1
Abril	5
Maio	23
Junho	54
Julho	24
Agosto	40
Setembro	49
Outubro	38
Novembro	17
Dezembro	11

Fonte: Elaborado pelo autor.

A consulta do Código-fonte 2 foi realizada para obter o número de óbitos causados por doenças respiratórias em 2020. O resultado dessa consulta pode ser visto na Tabela 3.

Código-fonte 2 – Busca de eventos causados por doenças respiratórias em 2020

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX icd10: <http://purl.bioontology.org/ontology/ICD10CM/>
PREFIX icd10MO: <https://www.ufc.br/ontologies/ICD10CM-MO/>

SELECT ?date (SUM(?deaths) as ?totalDeaths)
WHERE {
```

```
    ?disease rdfs:subClassOf* icd10:J00-J99 .
    ?event rdf:type icd10MO:MortalityEvent ;
        icd10MO:hasCause ?disease ;
        icd10MO:numberOfDeaths ?deaths ;
        icd10MO:date ?date .
    FILTER (?date >= "2020-01-31"^^xsd:date && ?date <=
        "2020-12-31"^^xsd:date)
}
GROUP BY ?date
ORDER BY ?date
```

Tabela 3 – Número de óbitos causados por doenças respiratórias no ano de 2020.

Mês	Número de óbitos
Janeiro	23
Fevereiro	27
Março	43
Abril	30
Maio	26
Junho	15
Julho	22
Agosto	20
Setembro	29
Outubro	34
Novembro	19
Dezembro	29

Fonte: Elaborado pelo autor.

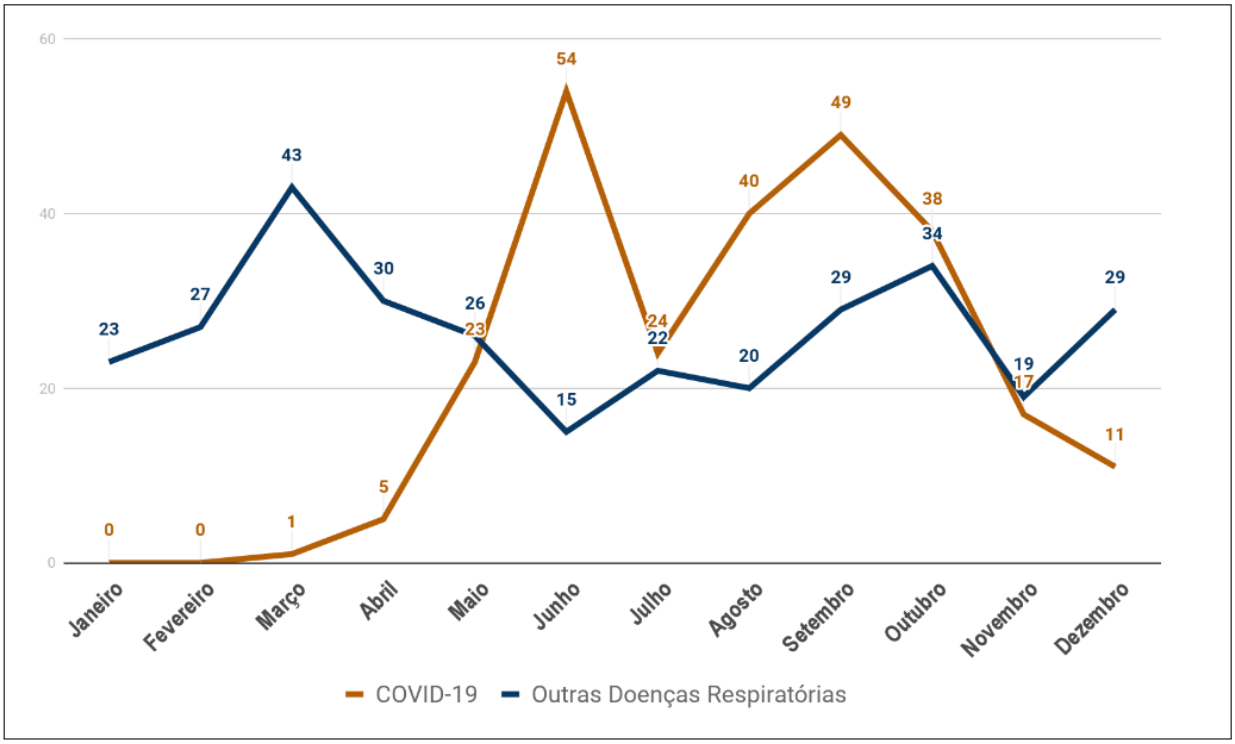
Em posse do resultado de ambas as consultas, é possível realizar a análise do número de óbitos causados por COVID-19 e comparação com o número de óbitos causados por doenças respiratórias no ano de 2020. A Figura 12 apresenta o gráfico com os resultados obtidos.

Os resultados da análise estatística são apresentados conforme os cenários estabelecidos:

- Ano Completo de 2020: Para este cenário foram obtidos os resultados apresentados na Tabela 4.

  O coeficiente de correlação de Pearson para este cenário foi de -0.39670264, o que indica uma correlação inversa muito fraca entre o número de casos de outras doenças e o número

Figura 12 – Gráfico de óbitos causados por COVID-19 e outras doenças respiratórias



Fonte: Elaborada pelo autor.

Tabela 4 – Número de óbitos causados por COVID-19 e outras doenças respiratórias no ano de 2020.

Mês	COVID-19	Outras doenças respiratórias
Janeiro	0	23
Fevereiro	0	27
Março	1	43
Abril	5	30
Maio	23	26
Junho	54	15
Julho	24	22
Agosto	40	20
Setembro	49	29
Outubro	38	34
Novembro	17	19
Dezembro	11	29

Fonte: Elaborado pelo autor.

de casos de COVID-19.

- Primeiros Quatro Meses de 2020: Para este cenário foram obtidos os resultados apresentados na Tabela 5.

  O coeficiente de correlação de Pearson para este cenário foi de -0.008018496, apontando para uma ausência quase total de correlação entre as duas variáveis no período considerado.

- Meses de Pico da COVID-19:

Tabela 5 – Número de óbitos causados por COVID-19 e outras doenças respiratórias nos primeiros meses ano de 2020.

Mês	COVID-19	Outras doenças respiratórias
Janeiro	0	23
Fevereiro	0	27
Março	1	43
Abril	5	30

Fonte: Elaborado pelo autor.

Para este cenário foram obtidos os resultados apresentados na Tabela 6.

Tabela 6 – Número de óbitos causados por COVID-19 e outras doenças respiratórias nos meses pico de COVID-19 em 2020.

Mês	COVID-19	Outras doenças respiratórias
Junho	54	15
Agosto	40	20
Setembro	49	29
Outubro	38	34

Fonte: Elaborado pelo autor.

O coeficiente de correlação de Pearson para este cenário foi de 0.137513272, o que indica uma correlação direta, porém ainda muito fraca, entre o número de casos de outras doenças e o número de casos de COVID-19.

Os resultados obtidos a partir da aplicação da abordagem descrita no Capítulo 5 permitiram responder às questões de pesquisa formuladas. Em relação à primeira questão, verificou-se que a integração semântica de dados foi eficaz em unificar informações heterogêneas provenientes de diferentes fontes do Observatório de Dados Abertos da Região do Sertão dos Crateús. A utilização de ontologias permitiu criar uma estrutura coesa e interoperável, facilitando a consulta e manipulação dos dados.

Quanto à segunda questão, a abordagem semântica se mostrou viável para correlacionar dados de fontes distintas. No entanto, ao aplicar a análise estatística com o coeficiente de correlação de Pearson, foi identificado que a correlação entre os casos de COVID-19 e a mortalidade por outras doenças respiratórias foi muito fraca. Isso sugere que, apesar da integração semântica dos dados, a relação entre essas variáveis não apresentou significância estatística nos cenários analisados.

Finalmente, em relação à terceira questão, concluiu-se que a metodologia de integração semântica é aplicável para futuros estudos e outros contextos, especialmente no campo

de dados de saúde pública. Embora a correlação específica investigada não tenha sido forte, o modelo de integração pode ser replicado e ampliado para outros tipos de dados, permitindo uma análise mais abrangente e detalhada em diferentes domínios.

7 CONCLUSÃO E CONSIDERAÇÕES FINAIS

O presente trabalho teve como objetivo avaliar a existência de uma correlação entre o aumento dos casos de COVID-19 e a mortalidade por outras doenças respiratórias na Região do Sertão dos Crateús durante o ano de 2020. Através de uma metodologia detalhada que envolveu a integração de dados e a análise estatística, foi possível compilar um conjunto de informações estruturadas que serviram como alicerce para uma análise criteriosa dos impactos da pandemia.

A integração dos dados, que resultou em 294.586 registros válidos, foi um ponto crucial deste estudo. A expansão da ontologia ICD-10-CM e o uso de ferramentas como Protégé, RDFLib e GraphDB permitiram uma abordagem semântica detalhada para o mapeamento e correlação dos dados. Esta abordagem destaca a importância da Web Semântica e dos dados abertos na pesquisa em saúde pública, pois permite uma integração e interpretação mais eficientes dos dados, facilitando a descoberta de conhecimento e a interoperabilidade entre diferentes sistemas e repositórios de dados.

Os resultados da análise estatística, baseados no coeficiente de correlação de Pearson, revelaram uma correlação muito fraca entre o número de casos de COVID-19 e a mortalidade por outras doenças respiratórias nos cenários estabelecidos. Os coeficientes de correlação obtidos, próximos de zero, indicam que não há uma relação linear significativa entre as variáveis estudadas.

7.1 Considerações Finais

A ausência de uma correlação forte entre o número de casos de COVID-19 e a mortalidade por outras doenças respiratórias sugere que fatores adicionais podem estar influenciando as taxas de mortalidade. As limitações do estudo, como a falta de variáveis que poderiam influenciar a mortalidade por doenças respiratórias são aspectos que podem ter afetado as conclusões. Além disso, a análise foi restrita a uma região geográfica específica e a um período limitado, o que pode não refletir a situação em outras regiões ou períodos.

Este estudo ressalta a importância dos dados abertos e disponíveis publicamente no avanço da pesquisa científica. A transparência e o compartilhamento de dados de saúde são cruciais para a realização de estudos como este, permitindo que pesquisadores ao redor do mundo tenham acesso a informações que podem ser utilizadas para entender melhor as dinâmicas das doenças e suas inter-relações.

As estratégias de intervenção e prevenção em saúde pública devem ser informadas por estudos como este, que utilizam dados abertos e a Web Semântica para fornecer insights valiosos. A continuidade do apoio à disponibilidade de dados abertos e ao uso da Web Semântica é essencial para a pesquisa em saúde e para o desenvolvimento de políticas públicas eficazes.

Para pesquisas futuras, recomenda-se a inclusão de variáveis adicionais como fatores socioeconômicos, acesso a serviços de saúde e cobertura vacinal. Além disso, seria útil realizar análises longitudinais para acompanhar as tendências ao longo do tempo e estudos qualitativos para compreender melhor as experiências dos indivíduos e das comunidades afetadas.

Em suma, este trabalho contribui para o conhecimento sobre os efeitos da COVID-19 e reforça a necessidade de abordagens metodológicas rigorosas, a importância da Web Semântica e o valor inestimável dos dados abertos na pesquisa em saúde. Os resultados obtidos sublinham a necessidade de políticas de saúde pública bem fundamentadas e baseadas em evidências para enfrentar os desafios impostos por pandemias e outras crises de saúde.

REFERÊNCIAS

AGHAEI, S.; NEMATBAKHSH, M. A.; FARSANI, H. K. Evolution of the world wide web: From web 1.0 to web 4.0. International Journal of Web Semantic Technology, v. 3, n. 1, p. 26, 2012.

ALVES, E. B. Observatório de dados abertos da região do sertão dos Crateús. Monografia (Bacharelado em Ciência da Computação) — Universidade Federal do Ceará, Ceará, 2022. Disponível em: <http://www.repositorio.ufc.br/handle/riufc/64368>. Acesso em: 14 jun. 2022.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. Scientific american, JSTOR, v. 284, n. 5, p. 34–43, 2001.

BLEIHOLDER, J.; NAUMANN, F. Data fusion. ACM Comput. Surv., Association for Computing Machinery, New York, NY, USA, v. 41, n. 1, jan 2009. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/1456650.1456651>.

CAVALCANTE, G. M. L. MAURA: Um Framework baseado em Mediador Semântico para construção eficiente de Linked Data Mashups. Dissertação (Mestrado em Ciência da Computação) — Instituto Federal de Educação, Ciência e Tecnologia do Ceará, 2017.

CRUZ, M. M. L. da. Uma abordagem para construção de Mashup de Dados especificados como uma visão sobre um EKG. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal do Ceará, 2021.

DONG, X.; GABRILOVICH, E.; HEITZ, G.; HORN, W.; LAO, N.; MURPHY, K.; STROHMANN, T.; SUN, S.; ZHANG, W. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. [S.l.: s.n.], 2014. p. 601–610.

FURGERI, S. O papel das linguagens de marcação para a ciência da informação. Transinformação, 2006. ISSN 0103-3786. Disponível em: <https://www.redalyc.org/articulo.oa?id=384334744006>.

GANGEMI, A.; PRESUTTI, V. The bourne identity of a web resource. 01 2006.

GUAN, W.; LIANG, Y. Complex sparql queries based on ontology and rdf. In: ATIQUZZAMAN, M.; YEN, N.; XU, Z. (Ed.). Proceedings of the 4th International Conference on Big Data Analytics for Cyber-Physical System in Smart City - Volume 1. Singapore: Springer Nature Singapore, 2023. p. 205–213. ISBN 978-981-99-0880-6.

HITZLER, P. A review of the semantic web field. Commun. ACM, Association for Computing Machinery, New York, NY, USA, v. 64, n. 2, p. 76–83, jan 2021. ISSN 0001-0782. Disponível em: <https://doi.org/10.1145/3397512>. Acesso em: 21 abr. 2024.

JACOB, E. K. Ontologies and the semantic web. Bulletin of the American Society for Information Science and Technology, American Society for Information Science & Technology, v. 29, n. 4, p. 19–19, 2003.

KNOWLEDGE, O. Open knowledge: What is open? 2020. Disponível em: <https://okfn.org/opendata/>. Acesso em: 08 nov. 2022.

LASSILA, O.; SWICK, R. R. et al. Resource description framework (rdf) model and syntax specification. Citeseer, 1998.

LIMA, J. C. de; CARVALHO, C. L. de. Uma Visão da Web Semântica. [S.l.], 2004.

LOBO, P. R.; DAGA, E.; ALANI, H.; FERNANDEZ, M. Semantic web technologies and bias in artificial intelligence: A systematic literature review. Semantic Web, v. 14, p. 1–26, 09 2022.

MILLER, E. An introduction to the resource description framework. D-lib Magazine, ERIC, 1998.

NOY, N. F.; MCGUINNESS, D. L. Ontology development 101: A guide to creating your first ontology. [S.l.], 2001.

PRUD'HOMMEAUX, A. S. E. SPARQL Query Language for RDF. 2001. Disponível em: <https://www.w3.org/2001/sw/DataAccess/rq23/>. Acesso em: 19 dez. 2022.

PRUD'HOMMEAUX, A. S. E. SPARQL Query Language for RDF. 2008. Disponível em: <https://www.w3.org/TR/rdf-sparql-query/>. Acesso em: 18 mai. 2022.

ROLIM, T. V. Sisifo: uma abordagem semântica para construção de Enterprise Knowledge Graphs. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal do Ceará, Fortaleza, 2020. 93 f.

SEGUNDO, J. E. S. Web semântica, dados ligados e dados abertos: uma visão dos desafios do brasil frente às iniciativas internacionais. In: XVI Encontro Nacional de Pesquisa em Pós-Graduação em Ciência da Informação. [S.l.: s.n.], 2015.

TANG, Z.; PEI, S.; PENG, X.; ZHUANG, F.; ZHANG, X.; HOEHNDORF, R. TAR: Neural Logical Reasoning across TBox and ABox. 2022. Disponível em: <https://arxiv.org/abs/2205.14591>.

THOR, A.; AUMUELLER, D.; RAHM, E. Data integration support for mashups. 2007.

TIWARI, S. M.; ORTIZ-RODRíGUEZ, F.; VILLAZON, B. Guest editorial: Current trends in semantic web and knowledge graphs. International Journal of Web Information Systems, v. 19, p. 121–122, 11 2023.

VIDAL, V. M. P.; CASANOVA, M. A.; ARRUDA, N.; ROBERVAL, M.; LEME, L. P.; LOPES, G. R.; RENSO, C. Specification and incremental maintenance of linked data mashup views. In: ZDRAVKOVIC, J.; KIRIKOVA, M.; JOHANNESSON, P. (Ed.). Advanced Information Systems Engineering. Cham: Springer International Publishing, 2015. p. 214–229. ISBN 978-3-319-19069-3.

W3C. Semantic Web - XML2000. 2000. Disponível em: <https://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html>. Acesso em: 17 jun. 2022.

W3C. World Wide Web Consortium (W3C). 2004. Disponível em: <https://www.w3.org/TR/rdf-concepts/>. Acesso em: 18 mai. 2022.

W3C. SPARQL 1.1 Overview. 2013. Disponível em: <https://www.w3.org/TR/sparql11-overview/>. Acesso em: 18 mai. 2022.

W3C. RDF Schema 1.1. 2014. Disponível em: <https://www.w3.org/TR/rdf-schema/>. Acesso em: 18 mai. 2022.