UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE COMPUTAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MESTRADO ACADÊMICO EM CIÊNCIA DA COMPUTAÇÃO

JOSÉ RENATO DA SILVA FREITAS

EKG CONTEXT EXPLORER: UMA FERRAMENTA GRÁFICA PARA EXPLORAÇÃO BASEADA EM CONTEXTO DA VISÃO SEMÂNTICA DE UM EKG

FORTALEZA

2024

JOSÉ RENATO DA SILVA FREITAS

EKG CONTEXT EXPLORER: UMA FERRAMENTA GRÁFICA PARA EXPLORAÇÃO BASEADA EM CONTEXTO DA VISÃO SEMÂNTICA DE UM EKG

Dissertação apresentada ao Curso de Mestrado Acadêmico em Ciência da Computação do Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Sistemas de Informação

Orientadora: Prof.ª Dr.ª Vânia Maria Ponte Vidal

FORTALEZA

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)

F936e Freitas, José Renato da Silva.
EKG Context Explorer : uma ferramenta gráfica para exploração baseada em contexto da Visão Semântica de um EKG / José Renato da Silva Freitas. – 2024.
72 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Ciência da Computação, Fortaleza, 2024.
Orientação: Profa. Dra. Vânia Maria Ponte Vidal.

1. Enterprise Knowledge Graph. 2. Integração semântica em larga escala. 3. Visão semântica. I. Título.
CDD 005

JOSÉ RENATO DA SILVA FREITAS

EKG CONTEXT EXPLORER: UMA FERRAMENTA GRÁFICA PARA EXPLORAÇÃO BASEADA EM CONTEXTO DA VISÃO SEMÂNTICA DE UM EKG

Dissertação apresentada ao Curso de Mestrado Acadêmico em Ciência da Computação do Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Sistemas de Informação

Aprovada em: 23 de fevereiro de 2024

BANCA EXAMINADORA

Prof.ª Dr.ª Vânia Maria Ponte Vidal (Orientadora)
Universidade Federal do Ceará (UFC)

Prof. Dr. José Maria da Silva Monteiro Filho
Universidade Federal do Ceará (UFC)

Prof.ª Dr.ª Melissa Lemos Cavaliére
Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)

Dedico este trabalho à minha esposa, Eliene, pelo apoio incondicional. Aos meus filhos, Manuel Neto e Ravi. Aos meus pais, Lemos e Lucinda, e aos meus irmãos, Júnior, Renan e Nara.

AGRADECIMENTOS

A Deus, por tudo. Pela comida em nossos pratos, todos os dias. Agradeço-o pela vida com saúde e por todas as oportunidades de felicidade. Agradeço-o pela minha família, Vieira e Freitas. Por meus pais, tios e tias que são exemplos para mim.

À Prof.ª Dr.ª Vânia Maria Ponte Vidal por todas as orientações nos projetos, nos artigos, em minha pesquisa e pela imensa contribuição na minha dissertação de mestrado. Uma professora e orientadora perseverante, dedicada e compromissada, zelando pelo primor em tudo que faz, tentando sempre tirar o máximo de mim.

Ao Prof. Dr. José Maria e à Prof.ª Dr.ª Melissa Lemos Cavaliere. Gratidão pela disponibilidade para participar dessa etapa importante na minha vida.

Ao Jonatas e à Gláucia da secretaria do Programa de Mestrado e Doutorado em Ciência da Computação da UFC (MDCC). Agradeço por sempre estarem de prontidão para esclarecer dúvidas, fornecer orientações e facilitar os processos administrativos.

Aos doutorandos em Ciência da Computação da UFC Túlio Vidal Rolim, Narciso Moura Arruda Junior e Caio Viktor da Silva Ávila que muito me ajudaram, com sugestões, experiências e ensinamentos. Vocês foram verdadeiros colegas e, sem dúvidas, fundamentais na minha jornada durante todo o Mestrado. Desejo que vocês alcancem sucesso profissional e pessoal.

Agradecimento mais que especial à minha esposa, Eliene. Sou grato por toda paciência, amor, suporte e companheirismo a mim concedidos. A alegria em seus olhos quando soube do meu ingresso no Mestrado foi combustível suficiente para chegarmos até aqui.

Ao Doutorando em Engenharia Elétrica, Ednardo Moreira Rodrigues, e seu assistente, Alan Batista de Oliveira, aluno de graduação em Engenharia Elétrica, pela adequação do *template* utilizado neste trabalho para que o mesmo ficasse de acordo com as normas da biblioteca da Universidade Federal do Ceará (UFC).

"A web não está concluída, é apenas a ponta do *iceberg*. As novas mudanças irão balançar o mundo ainda mais."

(Tim Berners-Lee)

RESUMO

Um *Enterprise Knowledge Graph* (EKG) serve como uma base robusta para gerenciamento de conhecimento, integração de dados e análises avançadas em organizações. Este trabalho tem o foco na Visão Semântica da Camada de Integração Semântica de EKGs. A Visão Semântica é o resultado da integração semântica das diversas fontes de dados de uma organização. A principal contribuição deste trabalho é um *framework* para a construção e exploração da Visão Semântica. O *framework* proposto fornece uma abordagem robusta para lidar eficazmente com a maioria dos desafios de integração semântica de dados. Primeiramente, uma arquitetura é proposta para organizar dados e metadados da Visão Semântica. Em seguida, cada nível da arquitetura da Visão Semântica é detalhada, apresentando as definições de seus elementos, onde foi utilizado um estudo de caso real para demonstrar como o grafo de dados é gerado com base na especificação da Visão Semântica. Tem-se ainda como contribuição deste *framework* o desenvolvimento de uma interface gráfica interativa, projetada para explorar os recursos na Visão Semântica em seus diversos contextos. Uma outra contribuição deste trabalho envolve a construção de um vocabulário dedicado que descreve os metadados da Visão Semântica e como esse vocabulário pode auxiliar construtores e consumidores de EKGs.

Palavras-chave: integração semântica; visão semântica; *Enterprise Knowledge Graph*.

ABSTRACT

An Enterprise Knowledge Graph (EKG) serves as a robust foundation for knowledge management, data integration, and advanced analytics in organizations. This work focuses on the Semantic View of the Semantic Integration Layer of EKGs. The Semantic View is the result of the semantic integration of various data sources in the organization. The main contribution of this work is a framework for the construction and exploration of the Semantic View. The proposed framework provides a robust approach to effectively deal with most semantic data integration challenges. Firstly, an architecture to organize Semantic View data and metadata is proposed. Next, each level of the Semantic View architecture is detailed, presenting the definitions of its elements, where a real case study was used to demonstrate how the data graph is generated based on the Semantic View specification. The contribution of this framework is also the development of an interactive graphical interface, designed to explore the resources in the Semantic View in its different contexts. Another contribution of this work involves the construction of a dedicated vocabulary that describes the Semantic View metadata and how this vocabulary can help builders and consumers of EKGs.

Keywords: semantic integration; semantic view; Enterprise Knowledge Graph.

LISTA DE FIGURAS

Figura 1 – Arquitetura do EKG. 17
Figura 2 – Exemplo de um grafo RDF 24
Figura 3 – Visão geral de um EKG 28
Figura 4 – Exemplo de Visualização “bola de pelo” 30
Figura 5 – Arquitetura dos Grafos de Conhecimento da Visão Semântica. 36
Figura 6 – Etapas de construção da Visão Semântica. 37
Figura 7 – Fragmento das ontologias O_{RFB} e O_{CAD}. 40
Figura 8 – Fragmento da ontologia da Visão Semântica $O_{SEFAZMA}$. 42
Figura 9 – Vocabulário comum para as propriedades compartilhadas. 42
Figura 10 – Estudo de Caso. 45
Figura 11 – Vocabulário dos metadados da Visão Semântica. 53
Figura 12 – Consulta 1 ao KG_Meta_VIS. 56
Figura 13 – Consulta 2 ao KG_Meta_VIS. 56
Figura 14 – Arquitetura da ferramenta *EKG Context Explorer*. 59
Figura 15 – Tela seleção de classes. 59
Figura 16 – Recursos da classe de generalização *sfz:Empresa*. 60
Figura 17 – Recursos da classe específica *sfz:Empresa_CAD*. 60
Figura 18 – Tela de exploração de recurso. Contexto Visão Semântica Exportada. 61
Figura 19 – Tela de exploração de recurso. Contexto Visão de Unificação. 62
Figura 20 – Tela de exploração de recurso. Contexto Visão de Fusão. 63
Figura 21 – Ontologia TLO_SEFAZMA. 64
Figura 22 – Exemplo de grafo de *timeline*. 65
Figura 23 – Tela de *timeline* de recurso. Contexto Visão de Unificação 65

LISTA DE QUADROS

Quadro 1 – Características da Web de Documentos e Web Semântica. 23
Quadro 2 – Resultado da consulta SPARQL de exemplo. 25
Quadro 3 – Resumo dos Trabalhos Relacionados. 34
Quadro 4 – Assertivas de Correspondência de Classes para O_{CAD} e O_{RFB}. 41
Quadro 5 – Assertivas de Correspondência de Propriedades para O_{CAD} e O_{RFB}. 41

LISTA DE CÓDIGOS-FONTE

Código-fonte 1 – Exemplo de consulta SPARQL simples 25
Código-fonte 2 – Consulta SPARQL da Visão de Unificação de um recurso. 48

LISTA DE ABREVIATURAS E SIGLAS

ACC	Assertivas de Correspondência de Classes
ACP	Assertivas de Correspondência de Propriedades
AFD	Assertiva de Fusão de Dados
API	*Application Programming Interface*
CAD	Cadastro de Contribuintes do Estado do Maranhão
CE	*Cluster* de Equivalência
CIS	Camada de Integração Semântica
CNPJ	Cadastro Nacional de Pessoa Jurídica
EKG	*Enterprise Knowledge Graph*
ERP	*Enterprise Resource Planning*
GC	Grafo de Conhecimento
HTTP	*Hypertext Transfer Protocol*
IA	Inteligência Artificial
IRI	*International Resource Identifier*
KG	*Knowledge Graph*
OWL	*Web Ontology Language*
RDF	*Resource Description Framework*
RDFS	*RDF Schema*
RFB	Receita Federal do Brasil
SEFAZ-MA	Secretaria de Fazenda do Estado do Maranhão
SPARQL	*SPARQL Protocol and RDF Query Language*
SUS	Sistema Único de Saúde
URI	*Universal Resource Identifier*
W3C	*World Wide Web Consortium*

LISTA DE SÍMBOLOS

G	Classe de Generalização
σ	Estado das fontes de dados
t	Instante
Ψ	Função de resolução de conflitos

SUMÁRIO

1 INTRODUÇÃO 16
2 FUNDAMENTAÇÃO TEÓRICA 22
2.1 Tecnologias da Web Semântica 22
2.2 Ontologia 25
2.3 *Enterprise Knowledge Graph* 27
2.4 Camada de Integração Semântica 28
2.5 Sistemas de Exploração de Grafos de Conhecimento 29
3 TRABALHOS RELACIONADOS 31
3.1 Integração Semântica em Larga Escala 31
3.2 Exploração de Grafo de Conhecimento 32
4 *FRAMEWORK* PROPOSTO 35
4.1 Arquitetura dos Grafos de Conhecimento da Visão Semântica 35
4.2 Processo Incremental para a Construção dos Grafos de Dados e Metadados da Visão Semântica 37
4.2.1 Estudo de Caso 38
4.2.2 Modelagem da Ontologia da Visão Semântica do EKG 38
4.2.3 Visões Semânticas Exportadas 43
4.2.4 Visões de Ligação e Visões de Unificação 45
4.2.4.1 Visões de Ligação *45*
4.2.4.2 Visão de Unificação *47*
4.2.5 Visões de Fusão 49
5 VOCABULÁRIO DOS METADADOS DA VISÃO SEMÂNTICA 52
5.1 Estrutura do Vocabulário dos Metadados da Visão Semântica 52
5.2 Uso do Vocabulário VSKG 55
6 FERRAMENTA *EKG CONTEXT EXPLORER* 58
6.1 Arquitetura da ferramenta *EKG Context Explorer* 58
6.2 Navegação baseada em contexto 59
6.2.1 Seleção de Classe 59
6.2.2 Exploração dos Recursos 61
6.2.3 TimeLine dos Recursos 63

7 CONCLUSÕES E TRABALHOS FUTUROS 68
REFERÊNCIAS 70

1 INTRODUÇÃO

Organizações pelo mundo vêm tentando compreender de maneira mais precisa e com alta qualidade os dados gerados em seus diversos sistemas. O cruzamento desses dados com dados externos é uma estratégia substancial para obter o conhecimento necessário visando melhorar a tomada de decisão e/ou a execução de tarefas pertinentes a seus negócios (NATH *et al.*, 2020). Em geral, esse tipo de cruzamento requer utilizar recursos computacionais que sejam capazes, indispensavelmente, de realizar um eficaz gerenciamento de dados e conhecimento.

Gerenciar o conhecimento em *Big Data*, por exemplo, tornou-se um problema crítico devido ao volume, velocidade e variedade dos dados gerados na atual era da informação. Por outro lado, possuir essa capacidade de gestão em diversos segmentos da sociedade é essencial para manter a competitividade e promover a inovação em qualquer ambiente corporativo, seja ele de grande, médio ou pequeno porte.

Segundo Lopes (2020), a competitividade e a inovação dão origem à criação de oportunidades de negócios. A contextualização e o compartilhamento de informações organizacionais são recursos valiosos que devem ser amplamente considerados ao planejar uma infraestrutura tecnológica com o intuito de ditar uma nova ordem no gerenciamento de dados e conhecimento para apoiar efetivamente a tomada de decisão.

Entretanto, as abordagens e ferramentas computacionais tradicionais empregadas nas infraestruturas tecnológicas atuais das organizações não oferecem modelos formais eficientes para solucionar os desafios da gestão do conhecimento, nem são adequadas para criar sistemas que gerenciam a abundância de informações provenientes de diversas fontes de dados. Essa lacuna pode ser superada por meio da incorporação de *Enterprise Knowledge Graph* (PAN *et al.*, 2017).

Enterprise Knowledge Graph (EKG) é um novo paradigma para representação do conhecimento cuja estrutura é baseada em grafo, conectando entidades e seus relacionamentos para declarar fatos da realidade com toda semântica envolvida (BONATTI *et al.*, 2019).

Na última década, pesquisas tanto no meio acadêmico quanto no ambiente empresarial têm aplicado EKG para criar soluções mais abrangentes, proporcionando um espaço de dados completo para usuários e aplicações (GALKIN *et al.*, 2017; DIBOWSKI; SCHMID, 2021).

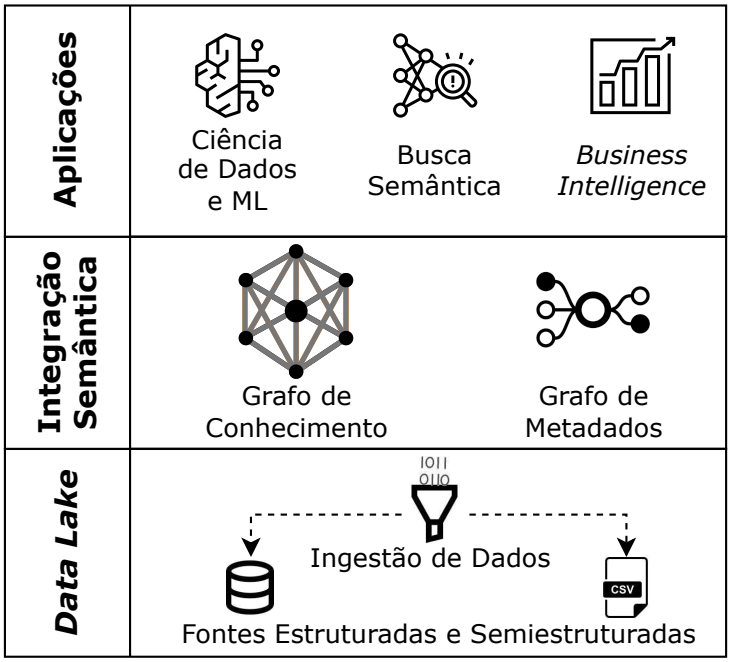
No ambiente corporativo, os EKGs estabeleceram uma posição sólida, agindo como uma peça central no gerenciamento de dados organizacionais. Atuam tanto como o principal repositório de dados de toda uma organização quanto como o *hub* para a integração de diversas

fontes de dados legados e descoberta de conhecimento (GRAINGER *et al.*, 2016).

Os EKGs são construídos usando tecnologias da Web Semântica para descrever e estruturar sintática e semanticamente o conhecimento do domínio de uma organização para suportar o raciocínio automático e inferências. Adicionalmente, a estrutura dos EKGs permite desenvolver aplicações inteligentes a partir do conhecimento e dos dados e metadados de alta qualidade representados. Essa estrutura robusta serve como alicerce para a gestão e integração semântica de dados, bem como para análises avançadas nas organizações (GRAINGER *et al.*, 2016).

A arquitetura geral de um EKG, conforme proposta em (GALKIN *et al.*, 2017) e ilustrada na Figura 1, compreende três camadas distintas:

Figura 1 – Arquitetura do EKG.



Fonte: elaborada pelo autor.

Camada do *Data Lake*. Esta camada serve como um sistema de armazenamento centralizado, projetado para acomodar uma ampla variedade de tipos de dados não-estruturados, semiestruturados e estruturados. Agrega dados de diversas fontes da organização e também pode incorporar fontes de dados externas (SAWADOGO *et al.*, 2019).

Camada de Integração Semântica. Esta camada é responsável por integrar semanticamente os dados provenientes de diversas fontes de dados, com o propósito de consolidar uma única visão coerente e significativa. Neste contexto, a visão resultante é denominada de **"Visão de Integração Semântica"** ou simplesmente **"Visão Semântica"**. Na Visão Semântica, os dados

são harmonizados e apresentados de forma que os usuários possam entender e explorar de maneira eficaz.

Camada de Aplicações. Esta camada é responsável pela implementação de aplicações ou serviços específicos que aproveitam os dados transformados e integrados, assim como os metadados, derivados da Camada de Integração Semântica. As aplicações nesta camada podem utilizar, ainda, tanto o modelo de dados integrado quanto o modelo de metadados para extrair *insights* e impulsionar processos de negócios.

A arquitetura de três camadas do EKG estabelece uma estrutura para gerenciar a aquisição, integração e utilização de dados em um ecossistema empresarial. Essa estrutura promove a interoperabilidade, facilita a tomada de decisão baseada em dados e aumenta a capacidade da organização de obter valor dos seus ativos de dados.

A relevância dos EKGs é evidenciada por uma ampla gama de aplicações e por sua adoção por organizações de diversos setores. Tais aplicações variam as funcionalidades, desde a procura e recuperação de informações até a tomada de decisão, abrangendo setores como as instituições financeiras, energia, telecomunicações e o governo, a exemplo da Bankinter, Repsol, Telefonica (LOPES, 2020) e Secretaria de Fazenda do Estado do Maranhão (SEFAZ-MA).

Em sistemas de Inteligência Artificial (IA) como *chatbots* e assistentes virtuais, os EKGs possibilitam a entrega de respostas mais contextuais e significativas, tornando esses sistemas mais eficazes na interação com os usuários e na compreensão de suas necessidades (JIOMEKONG; ASONG, 2022). Além disso, um EKG é uma ferramenta poderosa para enfrentar o desafio de conectar fontes de dados internas e externas, estruturadas, semiestruturadas e não-estruturadas. Destaca-se, ainda, os seguintes benefícios de um EKG:

- Estabelece um *framework* robusto para reconciliação de discrepâncias semânticas por meio de ontologias.
- Torna os dados de uma organização mais acessíveis à cientistas de dados e analistas de negócios, proporcionando um acesso unificado e transparente ao conhecimento de diversas fontes.
- Proporciona flexibilidade para incorporar novas fontes de dados.
- Consegue realizar consultas mais flexíveis (SPARQL, busca facetada, pergunta e resposta) respondendo a perguntas que não foram antecipadas durante a modelagem inicial.
- Permite a descoberta de novos conhecimentos (inferências).
- Suporta o desenvolvimento de aplicações inteligentes baseadas em ontologias (consulta e

busca semântica, mineração de dados semânticos, *chatbot* semântico).

– Auxilia na construção de Grafos de Conhecimento Especializado por meio da Visão Semântica.

Nesse contexto, um EKG permite que aplicações acessem informações dessas fontes, tratando de forma transparente desafios relacionados à heterogeneidade e ao acesso a recursos de diferentes origens (ARRUDA *et al.*, 2020; CRUZ, 2021).

O foco deste trabalho é a construção e exploração da Visão Semântica da Camada de Integração Semântica do EKG. Integrar dados de fontes diferentes e, ao mesmo tempo, manter a consistência e a precisão do grafo de conhecimento da Visão Semântica pode ser uma tarefa complexa, envolvendo vários desafios, incluindo:

Interoperabilidade Semântica. Dados de fontes distintas podem usar diferentes terminologias e esquemas para representar conceitos semelhantes. Alcançar a interoperabilidade semântica requer mapear precisamente e reconciliar essas diferenças para garantir que dados de diferentes fontes possam ser efetivamente integrados e consumidos.

Escalabilidade. À medida que o volume de dados cresce em uma empresa, o grafo de conhecimento deve conseguir escalar eficientemente para acomodar quantidades crescentes de dados enquanto mantém o desempenho.

Manutenção. O grafo de conhecimento precisa ser atualizado regularmente para refletir as mudanças nas fontes de dados e nos requisitos de negócios. Garantir que o grafo de conhecimento permaneça preciso, atualizado e relevante ao longo do tempo requer minuciosos processos de manutenção.

Desempenho de consulta. Projetando consultas eficientes, mecanismos para recuperar informações do grafo de conhecimento são essenciais para fornecer acesso oportuno e responsivo aos usuários.

Qualidade dos dados. Manter a qualidade dos dados em todo o processo de integração é essencial. Os dados podem conter inconsistências, erros, duplicatas e valores ausentes, que precisam ser resolvidos para garantir precisão e informações confiáveis no grafo de conhecimento.

Linhagem de dados. Em um ambiente heterogêneo com múltiplas fontes de dados e processos de integração, o rastreamento da linhagem de dados (*tracking Data Lineage*, em inglês) torna-se complexo. Dados podem sofrer inúmeras transformações e fluir através de vários sistemas antes

de serem integrados no grafo de conhecimento.

Gerenciamento de metadados. Gerenciar metadados de uma Visão Semântica é essencial para compreender, organizar e controlar os dados subjacentes. Os metadados fornecem informações contextuais sobre os dados em um grafo de conhecimento da Visão Semântica, incluindo sua estrutura, semântica, relacionamentos e uso. Quando criados e manuseados corretamente, os metadados apropriados são fundamentais para facilitar o gerenciamento e uso da Visão Semântica. Além disso, representam uma etapa importante na criação de dados FAIR[1] (*Findable*, *Accessible*, *Interoperable* and *Reusable*).

Neste trabalho, nos inspiramos no *framework* proposto por Vidal *et al.* 2015 para construir a **Visão Semântica** de EKGs. No contexto deste *framework*, a ontologia de domínio é um elemento chave da Visão Semântica e desempenha um papel crucial na integração semântica dos dados. Essa ontologia fornece um vocabulário comum e uma estrutura de conhecimento que promove a compreensão e a interpretação consistente dos dados integrados.

Ao estabelecer um vocabulário compartilhado, a ontologia ajuda a padronizar os termos e conceitos utilizados em diversas fontes de dados. Isso reduz ambiguidades e inconsistências na interpretação dos dados, facilitando a comunicação entre diferentes sistemas e partes interessadas. Além disso, a ontologia de domínio permite o mapeamento e a relação dos dados de diferentes fontes para um conjunto comum de conceitos. Essa capacidade de conectar e relacionar informações de maneira significativa amplia as possibilidades de descoberta de *insights* e conhecimentos ocultos nos dados.

Por meio da ontologia, os usuários podem realizar consultas semânticas mais sofisticadas, considerando o contexto e as relações definidas na estrutura de conhecimento. Isso facilita a exploração e análise dos dados de forma mais profunda e precisa.

Esta dissertação traz as seguintes contribuições principais:

1. Uma adaptação do *framework* proposto em (VIDAL *et al.*, 2015) para especificar a Visão de Integração Semântica de Dados de um EKG. Esta adaptação oferece uma estrutura sólida e confiável para lidar com os desafios específicos da integração semântica de dados, promovendo uma interpretação consistente e uma utilização eficaz dos dados da Visão Semântica de um EKG.
2. Implementação de uma interface gráfica interativa projetada para suportar a navegação em vários níveis da Visão Semântica. Em outras palavras, essa interface aprimora a análise,

[1] https://www.go-fair.org/fair-principles/

permitindo a exploração dos recursos na Visão Semântica em três contextos distintos: Visão Semântica da Fonte de Dados, Visão de Unificação e Visão de Fusão. Ao aproveitar esta interface, os usuários podem rastrear efetivamente a linhagem de dados de um recurso específico através de seu fluxo de transformações antes de sua integração no grafo de conhecimento da Visão Semântica. Esse recurso é inestimável para compreender as origens, transformações e integrações dos dados no grafo de conhecimento.

3. Desenvolvimento de um vocabulário para representar os metadados da Visão Semântica. Estes metadados desempenham um papel crucial no fornecimento de informações sobre os dados na Visão Semântica, permitindo sua descoberta, usabilidade e governança. Os metadados da Visão Semântica são armazenados em um grafo de conhecimento específico, mas devem estar intrinsecamente ligados ao grafo de dados da Visão Semântica, oferecendo uma perspectiva panorâmica de todo o grafo.
4. Aplicação do *framework* proposto na construção da Visão Semântica da camada de integração semântica do EKG da SEFAZ-MA. Os resultados desta aplicação são vistos no estudo de caso desenvolvido, apresentado na Seção 4.2.1.

Os capítulos subsequentes estruturam este trabalho da seguinte forma. O Capítulo 2 introduz as tecnologias utilizadas e explora os fundamentos teóricos para auxiliar na compreensão deste trabalho. No Capítulo 3, são apresentados trabalhos relacionados que têm objetivos similares aos desta pesquisa. Priorizaram-se trabalhos com tecnologias semelhantes que envolvem a construção, considerando algum processo de integração semântica de dados, e estudos que usaram ferramentas para exploração de EKG. O Capítulo 4 detalha o *framework* para especificação e construção da Visão Semântica. O Capítulo 5 descreve a construção do vocabulário de metadados da Visão Semântica e o Capítulo 6 apresenta a interface gráfica projetada para explorar os recursos na Visão Semântica. Finalmente, o Capítulo 7 é dedicado à conclusão e às perspectivas de futuras pesquisas.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo introduz as principais tecnologias e conceitos utilizados no desenvolvimento deste estudo. O objetivo é familiarizar o leitor com o tema adotado, proporcionando uma compreensão mais clara do trabalho.

2.1 Tecnologias da Web Semântica

Linked Data, RDF e SPARQL são elementos essenciais da Web Semântica. Esses tópicos serão abordados mais adiante. Antes, esta seção faz uma introdução sucinta sobre a Web Semântica. Pensada e criada pelo renomado cientista britânico Tim Berners-Lee, a **Web Semântica** é uma tecnologia para o compartilhamento global de dados, assim como a Web Clássica é para o compartilhamento de documentos HTML (BERNERS-LEE *et al.*, 2006).

No Quadro 1 é exibido um breve resumo das principais diferenças que caracterizam a Web Clássica, conhecida como Web de Documentos, e a Web Semântica. Analogamente, a Web Semântica atua como uma espécie de banco de dados global, enquanto a Web de Documentos funciona como um sistema global de arquivos. A Web Semântica tem como objeto principal as "coisas" (recursos de dados, entidades) e a Web de Documentos é orientada a documentos. Os *links* construídos na Web de Documentos são para conectar páginas HTML sendo acessadas por navegadores *web*, já os *links* na Web Semântica são para ligar "recursos" (entidades e propriedades) sendo acessados por navegadores semânticos.

As características "tipo de objeto primário" e "*links*", do Quadro 1, levam o grau de estruturação dirigida a objetos na Web de Documentos a um nível relativamente baixo, enquanto na Web Semântica esse nível é alto. Além disso, a semântica dos conteúdos na Web Semântica é completamente explícita. Na Web de Documentos, a semântica é implícita. Por fim, enquanto a Web de Documentos foi projetada principalmente para o consumo humano, a Web Semântica, por outro lado, tem seu foco voltado às máquinas, sem prejuízo do consumo humano.

Nesse contexto, a Web Semântica cria inúmeras oportunidades para o desenvolvimento de novos tipos de aplicações e ferramentas. Nessa perspectiva, Tim Berners-Lee propôs um modelo de camadas com recursos tecnológicos e linguagens chamado de "Pirâmide da Web Semântica" com o propósito de tornar a implementação da Web Semântica mais factível. Convenientemente, esse modelo evoluiu rapidamente, graças aos padrões criados pelo *World Wide Web Consortium* (W3C), sendo conhecido como "Pilha de Tecnologias da Web Semântica" (ISOTANI;

BITTENCOURT, 2015).

Quadro 1 – Características da Web de Documentos e Web Semântica.

	Web de Documentos	**Web Semântica**
Análogo a	Um sistema global de arquivos	Um banco de dados global
Objetos primários	Documentos	Recursos
Links entre	Documentos ou parte deles	Recursos (incluindo documentos)
Grau de estrutura em objetos	Relativamente baixo	Alto
Semântica de conteúdos e *links*	Implícito	Explícito
Projetado para	Consumo humano	Máquina primeiro, humanos depois

Fonte: adaptado de (BIZER *et al.*, 2008).

A iniciativa ***Linked Data*** faz parte do ferramental da Web Semântica e traz consigo novas possibilidades de construir a próxima geração de aplicações de dados inteligentes (ISOTANI; BITTENCOURT, 2015). O conceito *Linked Data* é definido como um conjunto de boas práticas para conectar e publicar conjuntos de dados de forma estruturada na Web, legíveis por máquina e preferencialmente com significados bem definidos (BIZER *et al.*, 2011). Ao exportar um conjunto de dados, antes isolado, como grafos conectados com outros conjuntos de dados, o *Linked Data* promove a criação do espaço de dados global conhecido como **Web de Dados**. O uso de *Linked Data* em projetos acadêmicos e empresariais é considerado um sucesso graças a adoção de padrões de infraestrutura da Web como o *Universal Resource Identifier* (URI), o *Hypertext Transfer Protocol* (HTTP) e padrões de estruturação semântica de dados como *Resource Description Framework* (RDF), o *RDF Schema* (RDFS) e a *Web Ontology Language* (OWL) (W3C, 2014), cumprindo as seguintes regras:

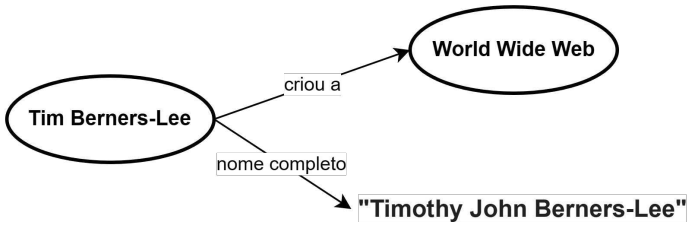
- Usar URI como nome para recursos.
- Usar URI sobre HTTP para que pessoas procurem esses nomes.
- Ao procurar por um URI, fornecer informações úteis, em RDF.
- Incluir *links* para outros URIs para que pessoas descubram mais recursos.

O ***Resource Description Framework* (RDF)** é um *framework* para representar informações na Web de forma estruturada usando um modelo de dados baseado em grafo (CYGANIAK *et al.*, 2014). Sua estrutura base é uma tripla do tipo ⟨`<sujeito>, <predicado>, <objeto>`⟩ que forma uma declaração para representar um fato cujo relacionamento entre o

<sujeito> e o <objeto> é definido por intermédio do <predicado>. O <predicado> é uma relação binária que conecta um recurso com outro recurso ou com um dado (valor literal). O agrupamento de uma ou mais triplas em um arquivo é denominado de "grafo RDF" (W3C, 2014; CYGANIAK *et al.*, 2014).

Em RDF, as declarações dos fatos são feitas sobre os recursos ou entidades. No âmbito da representação semântica, um recurso pode ser qualquer objeto da realidade como uma pessoa, uma empresa, um vídeo, uma imagem, um evento, etc. A Figura 2 apresenta um exemplo de um pequeno grafo RDF com duas triplas. Neste grafo, o sujeito <Tim Berners-Lee> tem as propriedades <criou a> e <nome completo> que se conectam, respectivamente, com o objeto <World Wide Web>, que também é um recurso, e o objeto <"Timothy John Berners-Lee">, que nesse caso é um valor literal do tipo *string*.

Figura 2 – Exemplo de um grafo RDF



Fonte: elaborada pelo autor.

Os grafos RDF publicados podem ser disponibilizados em memória, em arquivos de texto ou em um *RDF Triplestore*. *RDF Triplestore* ou simplesmente *triplestore* é um tipo de bancos de dados para o armazenamento e acesso a grafos RDF (LAUFER, 2015). A natureza dinâmica e flexível de um *triplestore* permite vincular diversos dados e indexá-los para consultas semânticas. Entre os diversos *triplestores* citados na literatura, o Virtuoso[1] e o Fuseki[2] são fortemente recomendados em aplicações para a Web Semântica, segundo (CAVALCANTE, 2017). No entanto, novos *triplestores* como o RDFOX[3], o GraphDB[4], o Stardog[5] e o Neptune[6] estão ganhando espaço por apresentar funcionalidades mais interessantes como serviço em nuvem, escalabilidade, visualização gráfica, virtualização, inferência e *Application Programming Interface* (API) nativa.

[1] https://virtuoso.openlinksw.com/
[2] https://jena.apache.org/documentation/fuseki2/
[3] https://www.oxfordsemantic.tech/rdfox
[4] https://graphdb.ontotext.com/
[5] https://www.stardog.com/platform/
[6] https://aws.amazon.com/pt/neptune/

Após a publicação, os grafos RDF estão prontos para serem consumidos, *SPARQL Protocol and RDF Query Language* (SPARQL) é a linguagem e protocolo para consultas semânticas recomendada pelo W3C (PRUD'HOMMEAUX *et al.*, 2013). O SPARQL permite a recuperação e manipulação de dados em RDF por meio da realização de consultas simples e complexas, envolvendo filtros, agregações e operadores lógicos. Devido aos seus recursos, o SPARQL é considerado uma tecnologia-chave da Web Semântica.

Para entender como o SPARQL opera, a Código-fonte 1 apresenta uma consulta SPARQL que visa responder a seguinte questão: "Quem criou a World Wide Web?". Essa consulta é executada sobre o grafo RDF de exemplo mostrado na Figura 2, e o resultado é exibido no Quadro 2.

A sintaxe aplicada no SPARQL tem várias semelhanças com a linguagem de consulta estruturada SQL (*Structure Query Language*, em inglês). Essa semelhança pode ser observada pelas palavras reservadas "SELECT" e "WHERE". As váriáveis em SPARQL são definidas com o sinal de interrogação no início da declaração, como em "*?sujeito*".

Código-fonte 1 – Exemplo de consulta SPARQL simples

```
1 PREFIX : <http://www.arida.ufc.br/VEKG#>
2 SELECT * WHERE {
3    ?sujeito :criou_a :World_Wide_Web.
4 }
```

Quadro 2 – Resultado da consulta SPARQL de exemplo.

?sujeito
<Tim Berners-Lee>

Fonte: elaborado pelo autor.

2.2 Ontologia

Ontologia é um termo de origem grega que tem um sentido especial na organização da informação. Na Filosofia, Ontologia é uma disciplina que visa estudar o mundo como ele é; estudar o ser, a realidade (MORAIS; AMBRÓSIO, 2007). Na Ciência da Computação, ontologia é usada como um artefato para a Representação do Conhecimento, um subcampo da Inteligência

Artificial (ALMEIDA, 2014).

Gruber (1993) define ontologia como a especificação de uma conceitualização; uma descrição de conceitos e relacionamentos, semelhante a uma especificação formal de um programa. Borst (1997) interpreta ontologia como uma especificação formal e explícita de uma conceitualização compartilhada. Unindo essas duas definições, ontologia converge em uma representação formal e inequívoca do conhecimento através do conjunto de conceitos, suas relações e propriedades. A representação formal significa que a ontologia é legível tanto por homem quanto por computador e inequívoca porque traduz que os conceitos não são ambíguos.

O fato das ontologias como artefato serem interpretadas por computador permite um leque de possibilidades de uso em várias áreas da Ciência da Computação, com destaque para: gestão do conhecimento, comércio eletrônico, sistemas de recomendação, processamento de linguagem natural, recuperação da informação na Web e Web Semântica.

Para Almeida (2020), a ontologia no campo da Ciência da Computação pode ser aplicada como base teórica e como artefato de *software*. Como teoria, a ontologia é usada para entender um domínio e abstraí-lo para modelos em sistemas de lógica formal. Como *software*, a ontologia é adotada para criar um vocabulário para representação em sistemas e inferência, por exemplo, as representações em OWL.

De acordo com Almeida (2020), "..., interoperabilidade pode ser definida como a capacidade de sistemas computacionais em trocar dados sem intervenção humana".

Algumas tecnologias como banco de dados relacional, sistemas *Enterprise Resource Planning* (ERP), *Data Warehouse* e *Data Lake* foram desenvolvidas para promover a interoperabilidade e integração entre sistemas. Porém, essas tecnologias não apresentam soluções satisfatórias de interoperabilidade e integração, principalmente quando há o envolvimento de dados semânticos. Os principais problemas destas tecnologia incluem:

- realizar consultas em mais de um banco de dados relacional;
- centralizar os esquemas de dados de vários sistemas em um único esquema ERP, podendo acumular centenas de tabelas;
- alto custo de *Data Warehouse* para dados agregados somente.

Diferentemente das tecnologias supracitadas, as ontologias são uma das tecnologias mais promissoras para lidar com interoperabilidade e integração. Enquanto os sistemas de banco de dados atuam com cláusula de "mudo-fechado", as ontologias trabalham com cláusula de "mundo-aberto". Cláusula de "mundo-fechado" significa que se um conjunto de dados não está

representado no banco de dados, é considerado que esse conjunto de fato não existe. Já na cláusula de "mundo-aberto", se um conjunto de dados não está representado na ontologia, ele pode de fato não existir, mas também pode existir no mundo real. E, se pode existir, ele pode ser inferido a partir da ontologia. Dessa forma, as ontologias são tecnicamente mais simples, comparadas às outras tecnologias, porém, robustas e capazes de representar e responder consultas de uma forma que outras tecnologias não podem (ALMEIDA, 2020).

2.3 *Enterprise Knowledge Graph*

Enterprise Knowledge Graph (EKG) é uma solução dinâmica e escalável para lidar com dados em larga escala (PAN *et al.*, 2017). Para tanto, um EKG atua como um *hub* para dados, metadados e conteúdo, oferecendo informações consistentes e sem ambiguidades por meio da representação do conhecimento (GRAINGER *et al.*, 2016).

Um EKG representa e gerencia o conhecimento de uma organização na forma de entidades, atributos e relacionamentos, semanticamente conceitualizados (GALKIN *et al.*, 2016). Essa estrutura permite integrar, organizar e analisar dados provenientes de diversas fontes, como documentos, bancos de dados, sistemas de informação, redes sociais, etc.

Ao criar uma representação semântica, um EKG produz um nível mais elevado de abstração e generalização que supera a barreira física ou formato dos dados. Em outras palavras, ele cria uma maneira de acessar continuamente os dados de uma organização, e também dados de terceiros ou globais. Esse acesso é uniforme, significativo e amigável para os seres humanos.

Dessa forma, os EKGs são uma solução para a qual muitos estão convergindo. São um paradigma adequado ao ecossistema global de dados do qual toda organização, grande ou pequena, faz parte, que pode ser usado para diversos fins, como:

- Lidar com a fragmentação dos dados espalhados em diferentes sistemas e formatos.
- Melhorar a busca e a recuperação de informações relevantes para os usuários da organização, por meio de consultas semânticas.
- Apoiar a tomada de decisão, por meio de inferências e análises avançadas.
- Fomentar a inovação, por meio da descoberta de novos padrões, conexões e *insights*.
- Facilitar a colaboração e o compartilhamento de conhecimento entre os membros da organização, evitando divergências de conceitos.

No geral, os EKGs fornecem o melhor quadro para integrar, organizar e aproveitar o conhecimento em uma organização, ao combinar:

- **Semântica formal.** Os padrões W3C possuem o formalismo que permite que humanos e computadores interpretem esquemas e dados de maneira inequívoca.
- **Desempenho.** Os EKGs permitem o gerenciamento eficiente de grafos com bilhões triplas.
- **Interoperabilidade.** As especificações para padronização e serialização de dados (RDF) e acesso (protocolo SPARQL), facilitam a integração, publicação e consumo dos dados.

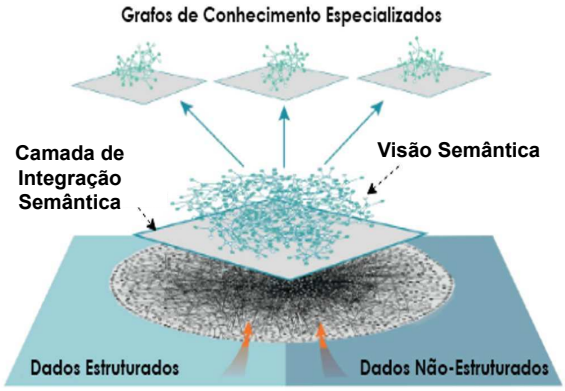
2.4 Camada de Integração Semântica

A Camada de Integração Semântica (CIS) de um EKG tem a função de integrar semanticamente informações das várias fontes de dados de maneira coerente e significativa. O produto dessa integração é a **Visão Semântica**, na qual os dados são harmonizados e dispostos de forma que os usuários possam compreender e explorá-los eficientemente.

Um ponto único de acesso aos dados é fornecido pela Camada de Integração Semântica, permitindo que consultas sejam feitas com base nos termos da ontologia de domínio. Isso significa que o usuário não precisa ter conhecimento das fontes de dados ou das relações entre elas. A ontologia do domínio serve como uma camada de abstração, fornecendo uma visão unificada das fontes de dados. Assim, usuários e aplicações podem acessar uma representação semântica dos dados que faz sentido para eles, de maneira transparente, sem a necessidade de entender os detalhes técnicos subjacentes às fontes originais (ROLIM *et al.*, 2021).

Um dos principais usos da Visão Semântica de um EKG é apoiar a criação de Grafos de Conhecimento Especializado (*Mashups* de Dados) para aplicações de análise de dados corporativos. Uma abordagem para construir um *Mashup* de Dados é utilizar as especificações (metadados) da Visão Semântica. A Figura 3 mostra o arcabouço de um EKG e da CIS.

Figura 3 – Visão geral de um EKG



Fonte: adaptada de Favio Vázquez (2018).

A princípio, após a publicação da Visão Semântica, seus dados e metadados devem estar disponíveis para apoiar diversas aplicações. Usar os metadados da Visão Semântica na construção de *Mashup* de Dados tem suas vantagens. O reaproveitamento dos mapeamentos semânticos das fontes de dados para a ontologia de domínio, por exemplo, é uma das mais significativas. Pois, reescrever mapeamentos semânticos manualmente pode ser, além de custoso, impreciso, até mesmo para pessoas com experiência. Faz sentido, portanto, que reaproveitar, integral ou parcialmente, mapeamentos já validados e em produção é benéfico para construir *Mashup* de Dados. Outra vantagem é ter metadados de qualidade que ajudam a avaliar a eficácia do *Mashup* de Dados.

Sabendo que um *Mashup* de Dados será utilizado por aplicações que necessitam de dados integrados e sem inconsistências, no sentido de ter dados com qualidade, a Visão Semântica deve conter informações de qualidade sobre seus artefatos (ARRUDA *et al.*, 2020).

2.5 Sistemas de Exploração de Grafos de Conhecimento

Essencialmente, a exploração e a visualização de dados têm como propósito facilitar a interpretação e manipulação das informações, permitindo a extração e a inferência de conhecimento. Em um contexto de *Big Data*, onde a análise manual dos dados é desafiadora, os sistemas e ferramentas de exploração e visualização desempenham um papel essencial ao viabilizar a compreensão e a interpretação mais eficiente dos dados (BIKAKIS; SELLIS, 2016). No âmbito dos Grafos de Conhecimento (GC), EKGs e ontologias como artefato, Desimoni e Po (2020) declaram que os sistemas de exploração devem disponibilizar funcionalidades como:

- Permitir uma visão panorâmica e específica sob demanda sobre os dados.
- Implementar pesquisa exploratória.
- Lidar com grandes conjuntos de dados.
- Destacar a evolução ao longo do tempo de um conjunto de dados.

Por mais de uma década, diversos sistemas, ferramentas e bibliotecas de exploração de GC têm sido desenvolvidos (BERNERS-LEE *et al.*, 2006). Bikakis e Sellis (2016) categorizam tais sistemas da seguinte maneira:

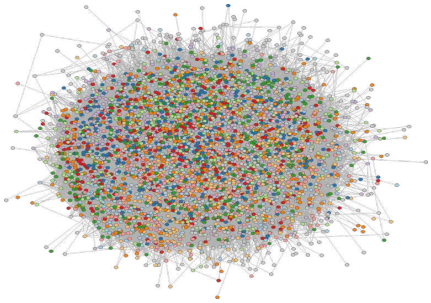
- **Navegadores e sistemas exploratórios.** Este tipo de sistema permite a navegação por textos clicáveis; e a representação de recursos e suas propriedades usam principalmente visualizações tabulares.
- **Sistemas de visualização genéricos.** São tipos de sistemas que dispõem de inúmeras

estruturas de visualização genéricas, que oferecem uma gama de tipos e operações de visualização.

- **Sistemas de visualização específicos de domínio, vocabulário e dispositivos.** Estes sistemas são especializados em atender às demandas de visualização para tipos específicos de dados e domínios. Por exemplo, existem sistemas que processam dados geoespaciais, proporcionando visualizações detalhadas e interativas de mapas, enquanto outros se dedicam a dados estatísticos, permitindo análises e compreensão das informações numéricas.
- **Sistemas de visualização baseados em grafos.** São sistemas que se utilizam de representações gráficas, como diagramas, para exibir e visualizar as relações entre nós (elementos individuais) e arestas (conexões) em um contexto de dados ou informações.
- **Sistemas de visualização de ontologias.** Na maioria destes sistemas, as ontologias são visualizadas seguindo o paradigma "nó-aresta". Por outro lado, alguns utilizam uma abordagem de contenção geométrica, representando a hierarquia de classes como um conjunto de círculos concêntricos.
- **Bibliotecas de Visualização.** Estas são bibliotecas que viabilizam a integração de visualizações em páginas HTML. Exemplos delas são o D3.js[7] e o Observable Plot[8].

Embora essas categorias de ferramentas e bibliotecas de visualização sejam amplamente usadas, o estudo de Li *et al.* (2023) aponta desafios relacionados às visualizações excessivamente complexas para os usuários finais. Os diagramas do tipo "nó-aresta", por exemplo, se tornam visualizações poluídas e difíceis de interpretação, na escala de milhares, milhões ou bilhões de nós. Esse tipo de visualização pode fácil e rapidamente se transformar em uma "bola de pelo"[9,10] (Figura 4) e ser ineficaz para fornecer *insights* de grandes GC.

Figura 4 – Exemplo de Visualização "bola de pelo"



Fonte: https://www.visual-computing.org/2016/04/18/untangling-networks/

[7] d3js.org
[8] observablehq.com
[9] https://cambridge-intelligence.com/how-to-fix-hairballs/
[10] Conexões densas que não podem ser visualizadas de forma útil.

3 TRABALHOS RELACIONADOS

Esta seção apresenta estudos que tiveram como objeto a construção de grafos de conhecimento semânticos, com foco especial em algum processo de integração semântica de dados. Além disso, expõe trabalhos que abordaram a exploração de grafos de conhecimento semânticos.

3.1 Integração Semântica em Larga Escala

A integração semântica de dados em larga escala é um desafio na Ciência da Computação devido à necessidade de combinar diversas fontes de dados que podem ter significados e estruturas diferentes. A complexidade dessa integração semântica envolve:

(i) lidar com um grande volume de dados mantendo um desempenho viável;
(ii) a heterogeneidade dos dados, que podem ser estruturados, semi-estruturados ou não-estruturados;
(iii) a qualidade dos dados, que podem conter erros, inconsistências ou ambiguidades e;
(iv) a interoperabilidade dos sistemas, que devem poder comunicar e trocar dados de forma eficiente e segura.

Para superar esses desafios, a integração semântica utiliza técnicas como ontologias, mapeamentos, alinhamentos, anotações e inferências, que permitem representar e harmonizar os significados dos dados de forma explícita e formal.

Em seu trabalho, Nadal (2019) propôs um processo de integração regido por uma camada semântica, implementada por meio de um repositório de metadados. O autor apresentou, ainda, uma arquitetura de referência como um modelo para implantar um conjunto de sistemas, sendo seu núcleo o repositório de metadados em conjunto com um modelo de metadados baseado em grafos para gerenciamento de metadados. Contudo, a abordagem do autor limitou-se, no que tange à representação dos elementos que constituem a construção de um grafo de conhecimento, às visões de ligação e os mapeamentos, demandando de adequação por parte de quem reusar à sua solução proposta.

A construção de um *Enterprise Knowledge Graph* (EKG), realizando a integração semântica de duas bases de dados da área da saúde no Brasil, foi o cerne do trabalho realizado em (CRUZ *et al.*, 2019). No entanto, esse trabalho não apresentou um vocabulário para especificar uma Camada de Integração Semântica, nem outros metadados do processo de criação do EKG.

O estudo realizado em (ROLIM *et al.*, 2020), por sua vez, usou um vocabulário para instanciar os elementos que especificam apenas as visões exportadas, as visões de ligação semântica e as publicações dessas visões. Esse estudo também propôs a construção de um EKG para integrar dados de saúde pública no Brasil. O enfoque adotado baseou-se na construção incremental, usando ontologias e um processo de Integração Semântica. Essa abordagem criou uma camada semântica sobre os dados do Sistema Único de Saúde (SUS) do Brasil visando tornar mais eficiente a análise e a exploração desses dados.

(SEQUEDA; LASSILA, 2021) propuseram um guia prático para construir *Enterprise Knowledge Graphs* a partir de bancos de dados relacionais corporativos. Eles apresentaram uma estrutura centrada em padrões de mapeamento para conectar esses bancos de dados com um modelo conceitual do domínio. O guia destaca o papel dos EKGs na criação de sistemas inteligentes que integram dados e conhecimento em larga escala. A metodologia de Sequeda e Lassila (2021) para construir o EKG é baseada em questões de negócio e na técnica "*pay-as-you-go*". No entanto, não abordaram a questão de *links* semânticos ou problemas de conflitos de dados.

Ainda como trabalho relacionado a processo ou especificação de integração semântica, Mohammadi (2022) propôs métodos computacionais para gerar metadados descritivos de GCS. Esses métodos combinaram padrões baseados em conteúdo e *logs* de consultas SPARQL realizadas por usuários, visando tornar os metadados mais informativos e úteis. Porém, não apresentou um vocabulário dedicado para descrever os metadados.

Por último, Azizi (2023) construiu um grafo de conhecimento por meio de um processo de integração de dados visando fornecer uma base de conhecimento para análise integrada dos dados advindos de fontes heterogêneas. Sua solução incluiu um esquema unificado, que define as relações entre todos os elementos no sistema de integração de dados $DIS = \langle G, S, M, F \rangle$. Contudo, a solução apresentada por Azizi não forneceu uma especificação para as visões de ligação e de fusão.

3.2 Exploração de Grafo de Conhecimento

A exploração de Grafos de Conhecimento (GC) é uma técnica para extrair informações relevantes de grandes conjuntos de dados estruturados. No entanto, alguns problemas e desafios podem surgir, tais como:

(i) A qualidade e a confiabilidade dos dados: nem todos os GCs são criados com os mesmos

critérios e padrões de qualidade. Além disso, alguns GCs podem ser incompletos ou desatualizados;

(ii) A complexidade e a escalabilidade: explorar GC requer algoritmos eficientes, capazes de lidar com a abundância e variedade de dados. No entanto, esses algoritmos podem ser computacionalmente custosos e exigir recursos significativos de tempo e memória.

Com relação às abordagens de exploração de GCS, o trabalho de Haase *et al.* (2019) introduziu a *metaphactory*, uma plataforma que visa facilitar a gestão de grafos de conhecimento. Sua arquitetura é fundamentada em padrões abertos, o que a torna adequada para ser implementada em diversas áreas, se enquadrando na categoria de sistemas de exploração genéricos. A *metaphactory* disponibiliza interfaces personalizáveis para a criação rápida de aplicativos para uma variedade de cenários de uso. A interface de usuário da *metaphactory* é baseada em componentes web e orientada a recursos, ou seja, ela gera uma visualização HTML com base em um URI associado a um recurso no grafo de conhecimento.

O estudo realizado em (SOUZA *et al.*, 2022) desenvolveu a ferramenta "TKGEvol-Viewer" para visualizar a evolução de GCs temporais. Essa ferramenta permite a análise visual dos grafos com base em métricas codificadas nas estruturas, usando um componente modal e filtrando informações de análises. A exploração é realizada ao selecionar "questões-guia", pré-definidas, e um período em anos disponibilizados em formulários. Sua visualização é do tipo "nó-aresta" e encaixa-se nas categorias: sistemas de visualização baseados em grafos e sistemas de visualização específicos de domínio, vocabulário e dispositivos.

Sellami e Zarour (2022) trataram da exploração de grandes quantidades de dados heterogêneos na Web e redes sociais para criação de conhecimento. Os autores desenvolveram um modelo chamado KGMap para unir fontes de dados diversas em um grafo de conhecimento. Eles também introduziram o KeyFSI, uma interface de usuário de navegação facetada e responsiva, baseado em palavras-chave, projetada para facilitar a exploração e visualização de dados incorporados por trás do grafo de conhecimento.

Por fim, Avila e Vidal (2023) apresentaram a ferramenta LiRB, uma interface Web leve e interativa, no estilo das tradicionais páginas da Web, que permite uma exploração sobre GCS. LiRB é transparente com relação às consultas SPARQL promovendo uma menor curva de aprendizagem aos usuários, além de exibir os eventos dos recursos do GCS com uma visualização do tipo linha do tempo. Contudo, LiRB não exibe informações de proveniência dos recursos, tão pouco das propriedades.

O resumo exibo no Quadro 3 mostra que os trabalhos relacionados desta seção não apresentaram um vocabulário que especifica toda a Visão Semântica. Além disso, eles também não apresentaram uma interface gráfica interativa para explorar da Visão Semântica em diversos níveis e contextos, e/ou rastrear a linhagem dos dados (*tracking Data Lineage*).

Quadro 3 – Resumo dos Trabalhos Relacionados.

	Vocabulário de Metadados da Visão Semântica	Processo de Integração Semântica	Interface de explorar GC	Exploração baseada em contexto da Visão Semântica	*Tracking Data Lineage*
NADAL, 2019	✓	✓	-	-	-
CRUZ *et al.*, 2019	-	✓	✓	-	-
ROLIM *et al.*, 2020	✓	✓	-	-	-
SEQUEDA; LASSILA, 2O21	-	✓	-	-	-
MOHAMMADI, 2022	-	✓	✓	-	-
AZIZI, 2023	-	✓	-	-	-
HAASE *et al.*, 2019	-	-	✓	-	-
SOUZA *et al.*, 2022	-	-	✓	-	-
SELLAMI; ZAROUR, 2022	-	✓	✓	-	-
AVILA; VIDAL, 2023	-	-	✓	-	-
Nossa Proposta	✓	✓	✓	✓	✓

Fonte: elaborado pelo autor.

4 *FRAMEWORK* PROPOSTO

Este capítulo introduz o *framework* proposto para especificar a Visão de Integração Semântica de Dados de um EKG (**Visão Semântica**). Inspirado no trabalho apresentado em (VIDAL *et al.*, 2015), este *framework* oferece uma estrutura sólida e confiável para lidar com os desafios específicos de integração semântica de dados, promovendo uma interpretação consistente e uma utilização eficaz dos dados por meio da Visão Semântica do EKG.

O *framework* desenvolvido por Vidal *et al.* (2015) é amplamente utilizado e referenciado por outros autores, como Lopes *et al.* (2016), Lopes *et al.* (2017), Cruz *et al.* (2019) e Arruda *et al.* (2020). Relevante, ele fornece uma abordagem robusta para tratar eficazmente a maioria dos desafios de integrar dados.

A seguir, a Seção 4.1 apresenta uma nova arquitetura destinada a estabelecer a Visão Semântica e a Seção 4.2 aborda o processo de construção dessa visão e os principais desafios que impulsionam a necessidade do *framework* proposto.

4.1 Arquitetura dos Grafos de Conhecimento da Visão Semântica

A arquitetura do *framework* proposto, exibida na Figura 5, estabelece uma estrutura para gerenciar a integração semântica de dados. Nessa arquitetura, a Visão Semântica é composta por dois grafos de conhecimento: um de dados e outro de metadados.

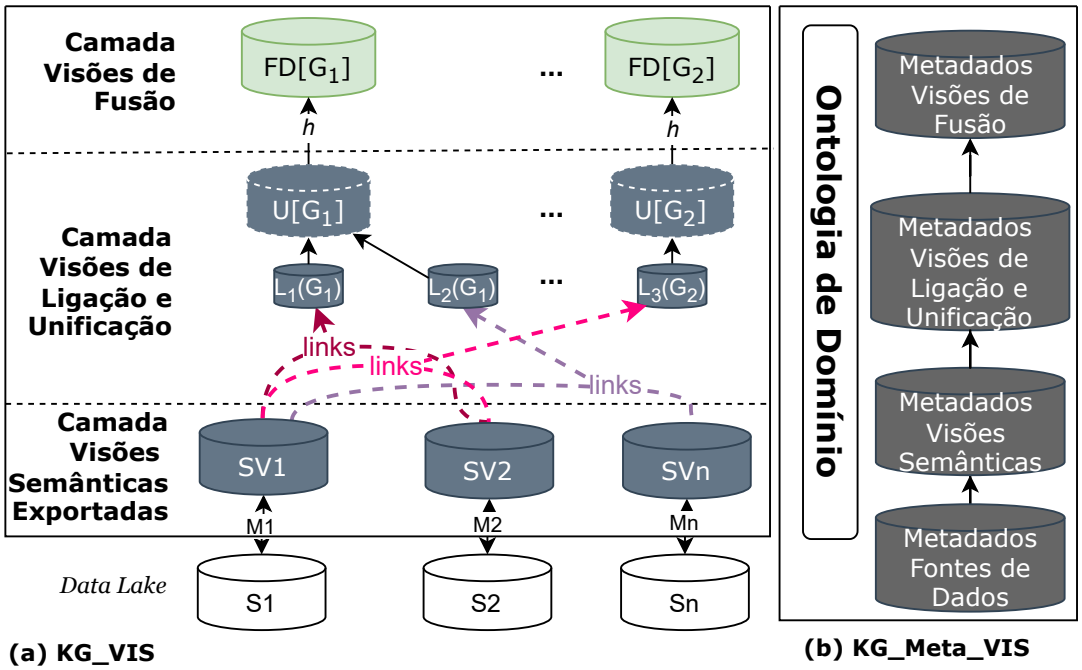
O grafo de dados da Visão Semântica é montado a partir dos grafos de dados das visões das três camadas. Os grafos de dados dessas visões são gerados automaticamente com base nas próprias especificações. Esse processo de construção automatizado garante consistência e coerência na representação de dados da Visão Semântica. As visões das camadas de nível superior utilizam as especificações fornecidas pelas visões subjacentes, estabelecendo uma estrutura hierárquica para definir a Visão Semântica.

Os metadados da Visão Semântica são armazenados em um grafo de conhecimento distinto, mas devem estar intrinsecamente vinculados ao grafo de dados, oferecendo uma perspectiva panorâmica da Visão Semântica.

Na arquitetura proposta, o grafo de conhecimento (*Knowledge Graph* (KG), em inglês) de dados **KG_VIS** e o grafo de metadados **KG_Meta_VIS** são discutidos a seguir.

O **KG_VIS**, na Figura 5(a), representa o grafo de conhecimento de dados da Visão Semântica, derivado da integração de dados de diversas fontes em um determinado momento. O

Figura 5 – Arquitetura dos Grafos de Conhecimento da Visão Semântica.



Fonte: elaborada pelo autor.

KG_VIS é estruturado em três camadas de visões:

- **Camada das Visões Semânticas Exportadas.** No *Data Lake*, cada fonte de dados exporta uma visão RDF construída por meio de mapeamento sistemático dos dados de origem para um vocabulário comum estabelecido por uma ontologia de domínio.
- **Camada das Visões de Ligação e Unificação.** As visões de ligação *"sameAs"* envolvem especificar relacionamentos (*links*) entre entidades que são equivalentes ou idênticas em diferentes conjuntos de dados. Essa integração ajuda a criação de uma representação unificada e abrangente de entidades, agregando atributos, relacionamentos e metadados de várias Visões Semânticas Exportadas.
- **Camada das Visões de Fusão.** O objetivo das Visões de Fusão é resolver conflitos que surjam nas Visões de Unificação quando diferentes fontes têm informações conflitantes sobre a mesma entidade ou relacionamento.

O *framework* proposto adota uma abordagem híbrida para a construção do KG_VIS. Neste modelo híbrido, certas visões do KG_VIS são geradas dinamicamente *on-the-fly* (visões virtuais), fornecendo acesso em tempo real aos dados, enquanto outras visões são pré-computadas e armazenadas de maneira persistente (visões materializadas), oferecendo vantagens como latência de consulta reduzida e tempos de resposta aprimorados. Esta arquitetura híbrida visa encontrar um equilíbrio entre o imediatismo das visões virtuais e a eficiência das visões materializadas, garantindo um equilíbrio ideal entre a utilização de recursos e o desempenho da consulta.

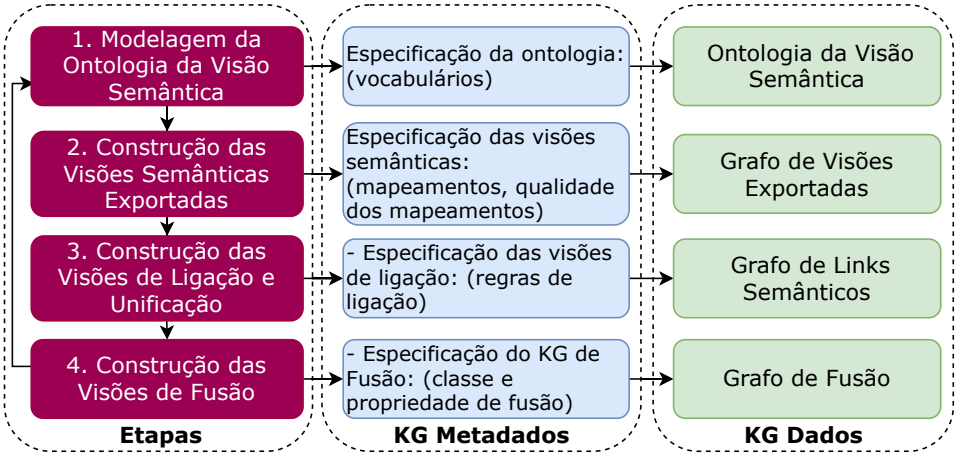
O **KG_Meta_VIS**, representado na Figura 5(b), serve como repositório de metadados que definem as visões do KG_VIS. Desempenha um papel crucial no fornecimento de informações sobre os dados no KG_VIS, garantindo sua descoberta, usabilidade, linhagem de dados e governança. O KG_Meta_VIS é organizado em quatro níveis de metadados e os metadados do nível superior dependem ou utilizam os metadados da camada imediatamente abaixo. Um componente-chave do KG_Meta_VIS é a Ontologia de Domínio, por atuar no estabelecimento de um vocabulário comum para transformação de dados nas fontes de dados. Além dos metadados para as visões no KG_VIS, o KG_Meta_VIS também abrange metadados relacionados às fontes de dados.

É importante notar que o KG_VIS deve estar intrinsecamente ligado ao KG_Meta_VIS, oferecendo uma perspectiva panorâmica de toda a Visão Semântica, delineando os ativos de dados do KG_VIS. Que essa abordagem interconectada garante uma compreensão abrangente da Visão Semântica e aprimora o gerenciamento e o uso do KG_VIS e KG_Meta_VIS. E, que a arquitetura proposta para a Visão Semântica foi projetada para atender diversas necessidades, como: (i) explorar recursos em contextos específicos; (ii) facilitar a manutenção da Visão Semântica; (iii) auxiliar na criação de Visões de Fusão de qualidade e (iv) obter *Data Lineage*.

4.2 Processo Incremental para a Construção dos Grafos de Dados e Metadados da Visão Semântica

No *framework* proposto neste trabalho, o processo de especificação e construção da Visão Semântica de um *Enterprise Knowledge Graph* (EKG) segue o método "*pay-as-you-go*" de (SEQUEDA *et al.*, 2019) executando quatro etapas principais mostradas na Figura 6.

Figura 6 – Etapas de construção da Visão Semântica.



Fonte: elaborada pelo autor.

Cada uma das etapas acima é discutida detalhadamente nesta seção. A base para estas discussões é lançada na Seção 4.2.1, onde é apresentado um estudo de caso real com dados de empresas e contribuintes da Unidade Federativa Maranhão/Brasil. Este estudo de caso serve como exemplo prático para demonstrar como essas etapas são aplicadas em um cenário real durante a construção da Visão Semântica de um EKG.

4.2.1 Estudo de Caso

O estudo de caso apresentado nesta seção foi desenvolvido na Secretaria de Fazenda do Estado do Maranhão/Brasil (SEFAZ-MA) que construiu um EKG onde a sua camada semântica integrou onze fontes de dados. Para os exemplos do restante desta dissertação, consideraremos apenas as duas fontes de dados descritas a seguir.

(i) Cadastro de Contribuintes do Estado do Maranhão (CAD): **CAD** é uma fonte de dados que contém informações sobre empresas e pessoas físicas, bem como dados de estabelecimentos e de contribuintes do Governo do Maranhão/Brasil.

(ii) Receita Federal do Brasil (RFB): **RFB** é uma fonte de dados responsável por fornecer informações confiáveis sobre empresas, estabelecimentos, sociedades, sócios e dados gerais sobre todos os tipos de pessoa jurídica por meio do Cadastro Nacional de Pessoa Jurídica (CNPJ). As empresas não cadastradas no CNPJ da RFB estarão exercendo suas atividades ilegalmente.

Para o estudo de caso, consideramos a integração de dados de empresas das fontes de dados CAD e RFB.

4.2.2 Modelagem da Ontologia da Visão Semântica do EKG

No *framework* proposto, a construção da ontologia da Visão Semântica do EKG é empreendida em uma abordagem iterativa e *bottom-up*, envolvendo quatro passos principais, apresentada a seguir:

Passo 1. Modelagem das Ontologias das Fontes de Dados:

O processo começa modelando, individualmente, as ontologias das fontes de dados. Esta etapa envolve uma análise e a representação das estruturas, entidades, relacionamentos e atributos inerentes a cada fonte de dados. Ao adotar uma perspectiva ascendente, constrói-se uma compreensão fundamental das diversas estruturas ontológicas presentes nos dados.

Passo 2. Comparação das Ontologias das Fontes de Dados:

Após a modelagem de cada fonte de dados, é realizada uma análise comparativa para identificar pontos em comum e as diferenças entre as ontologias de diversas fontes de dados. Esta etapa envolve um exame sistemático das sobreposições e distinções conceituais, estabelecendo as bases para constituir conexões e alinhamentos entre conjuntos de dados. Esta abordagem comparativa garante uma compreensão da natureza heterogênea dos dados.

O resultado deste passo é um conjunto de Assertivas de Correspondência de Classes (ACC), juntamente com um conjunto de Assertivas de Correspondência de Propriedades (ACP). As assertivas de correspondência de classes especificam relacionamentos de equivalência semântica entre classes de diferentes fontes de dados. Dizemos que duas classes são semanticamente equivalentes quando se referem ao mesmo conceito do mundo real. Já as assertivas de correspondência de propriedades especificam relacionamentos de equivalência semântica entre as propriedades de classes semanticamente equivalentes. O problema de geração das assertivas de correspondência está fora do escopo deste trabalho.

Passo 3. Integração das Classes Semanticamente Equivalentes:

No passo de integração das classes semanticamente equivalentes, com base nas ACC, as classes das Ontologias das Fontes de Dados são agrupadas em *clusters* de equivalência (CE) semântica. Essa abordagem visa melhorar a compreensão e a organização das classes relacionadas, por meio de grupos que compartilham significados semânticos congêneres. Como veremos, isso facilita a modelagem da ontologia da Visão Semântica, crucial para o entendimento e a representação dos conceitos no sistema de dados.

Quando as classes de um *Cluster* de Equivalência (CE) têm o mesmo nome, é essencial resolver essa ambiguidade utilizando um sufixo, como o nome da fonte de dados. Isso é necessário para preservar a integridade das classes das ontologias das fontes de dados na ontologia da Visão Semântica. Além disso, quando um CE contém mais de uma classe, é necessário criar uma **classe de generalização** para representar um conceito mais amplo e geral em relação a todas as outras classes do *cluster*. Isso contribui para uma estrutura mais organizada e compreensível da ontologia da Visão Semântica, que abrange todas as classes das ontologias das fontes de dados, juntamente com as classes de generalização.

Essencialmente, a classe de generalização foi projetada para consolidar e unificar as propriedades inerentes às suas subclasses. Ao obter a união das propriedades dessas subclasses, a classe de generalização torna-se uma entidade abrangente que incorpora as propriedades

compartilhadas dos conceitos mais especializados na hierarquia de classes.

Passo 4. Integração das Propriedades Semanticamente Equivalentes:

Neste passo, com base nas ACP, é realizada a fusão das propriedades semanticamente equivalentes das classes de generalização. Esse processo envolve a definição de um vocabulário comum para as propriedades das classes de generalização, visando alinhar as propriedades compartilhadas e criar uma representação abrangente que vai além das fontes de dados individuais.

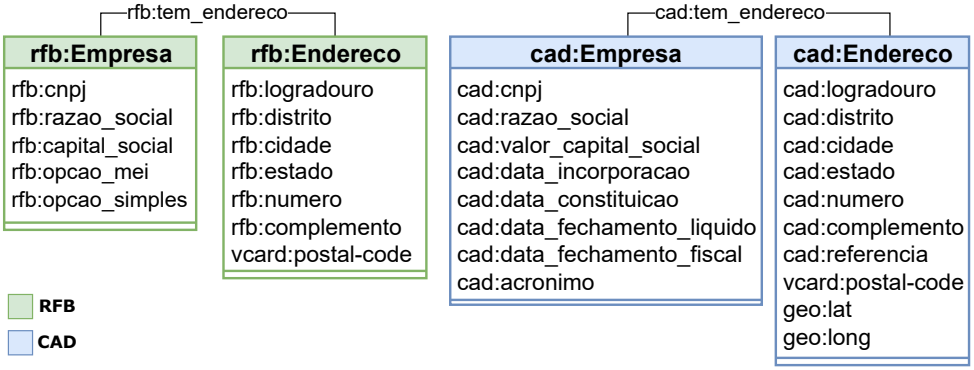
A definição desse vocabulário comum permite que as propriedades sejam reutilizadas pelas subclasses, garantindo consistência e padronização na representação dos dados. Dessa forma, a classe de generalização consolida e unifica as propriedades inerentes às suas subclasses, proporcionando uma estrutura coesa e integrada para a modelagem dos dados no sistema.

Essa abordagem facilita a interoperabilidade e a compreensão dos dados, ao mesmo tempo em que promove a consistência semântica e a reutilização de conceitos, contribuindo para uma representação mais completa e unificada das informações.

A seguir, detalhamos cada um dos passos na construção de uma parte da ontologia da Visão Semântica do EKG da SEFAZ-MA, denominada $O_{SEFAZMA}$:

Passo 1. Considerando o estudo de caso, inicia-se com a modelagem da ontologia das fontes de dados CAD e RFB, a partir da análise dos seus esquema e entidades. A Figura 7 apresenta o diagrama de classes e o vocabulário de um pequeno fragmento das ontologias O_{CAD} e O_{RFB}.

Figura 7 – Fragmento das ontologias O_{RFB} e O_{CAD}.



Fonte: elaborada pelo autor.

Passo 2. Na comparação das ontologias O_{CAD} e O_{RFB}, foram identificadas as assertivas de correspondência de classes mostradas no Quadro 4, e as assertivas de correspondência de propriedades exibidas no Quadro 5.

Quadro 4 – Assertivas de Correspondência de Classes para O_{CAD} e O_{RFB}.

Identificador da Assertiva	Assertiva
ACC1	cad:Empresa ≡ rfb:Empresa
ACC2	cad:Endereco ≡ rfb:Endereco

Fonte: elaborado pelo autor.

Quadro 5 – Assertivas de Correspondência de Propriedades para O_{CAD} e O_{RFB}.

Identificador da Assertiva	Assertiva
ACP1	rfb:cnpj ≡ cad:cnpj
ACP2	rfb:razao_social ≡ cad:razao_social
ACP3	rfb:capital_social ≡ cad:valor_capital_social
ACP4	rfb:logradouro ≡ cad:logradouro
ACP5	rfb:bairro ≡ cad:bairro
ACP6	rfb:cidade ≡ cad:cidade
ACP7	rfb:estado ≡ cad:estado
ACP8	rfb:numero ≡ cad:numero
ACP9	vcard:postal-code ≡ vcard:postal-code

Fonte: elaborado pelo autor.

Passo 3. Com base nas assertivas de correspondência de classes *ACC*1 e *ACC*2, no Quadro 4, são identificados dois *clusters* de classes semanticamente equivalentes:

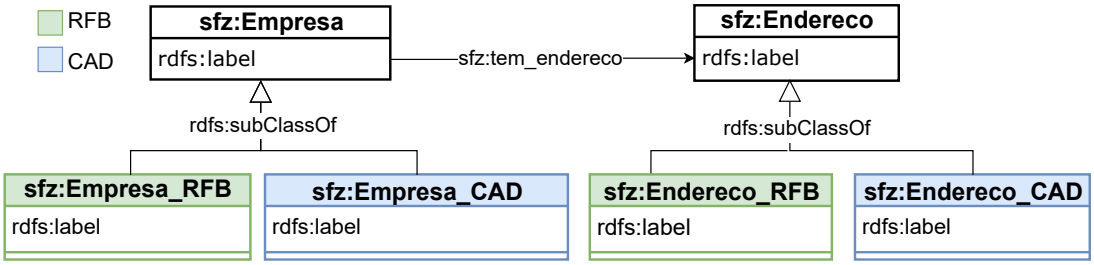
(i) CE_1: ⟨rfb:Empresa, cad:Empresa⟩ e;

(ii) CE_2: ⟨rfb:Endereco, cad:Endereco⟩.

Para o *cluster* CE_1 é criada a classe de generalização *sfz:Empresa*, e para o *cluster* CE_2 é criada a classe de generalização *sfz:Endereco*.

Após a definição das classes de generalização, é criada a hierarquia de classes, por meio da propriedade *rdfs:subClassOf*, realizando a linhagem entre os conceitos. Além disso, é efetuada a fusão das relações da O_{CAD} e O_{RFB} para geração da ontologia da visão semântica $O_{SEFAZMA}$, exibida na Figura 8. Seguindo as diretrizes do *framework* proposto, o alinhamento das classes semânticas *rfb:Empresa* e *cad:Empresa* se dá com a mudança nas suas terminologias, usando o seguinte padrão: "{prefixo-da-classe-generalização}:{nome-da-classe-generalização}_{nome-da-fonte}". Dessa forma, *rfb:Empresa* muda para *sfz:Empresa_RFB* e *cad:Empresa* passa a ser *sfz:Empresa_CAD*. Idem para as classes *rfb:Endereco* e *cad:Endereco*.

Figura 8 – Fragmento da ontologia da Visão Semântica $O_{SEFAZMA}$.



Fonte: elaborada pelo autor.

Passo 4. Neste passo deve ser definido um vocabulário comum para as propriedades das classes de generalização, alinhando propriedades semanticamente equivalentes. Com base nas ACP das classes do *cluster* CE_1: ⟨ACP1, ..., ACP3⟩, é realizada a fusão das propriedades semanticamente equivalentes. Da mesma forma, com base nas ACP das classes do *cluster* CE_2: ⟨ACP4, ..., ACP9⟩, é realizada a fusão das propriedades semanticamente equivalentes. A Figura 9 mostra as propriedades das classes de generalização *sfz:Empresa* e *sfz:Endereco* após a fusão.

Figura 9 – Vocabulário comum para as propriedades compartilhadas.

Propriedades da classe Empresa	Propriedades da classe Endereço
rdfs:label	rdfs:label
sfz:cnpj	sfz:logradouro
sfz:razao_social	sfz:distrito
sfz:valor_capital_social	sfz:cidade
sfz:opcao_mei	sfz:estado
sfz:opcao_simples	sfz:numero
sfz:data_incorporacao	vcard:postal-code
sfz:data_constituicao	sfz:complemento
sfz:data_fechamento_liquido	sfz:referencia
sfz:data_fechamento_fiscal	geo:lat
sfz:acronimo	geo:long

Fonte: elaborada pelo autor.

A ontologia da Visão Semântica foi construída em OWL no formato ***Turtle*** (uma serialização RDF) com o auxílio do editor Protègè[1], uma ferramenta amplamente usada para construir ontologias. A implementação da ontologia da Visão Semântica como artefato foi realizada utilizando os vocabulários das ontologias das fontes de dados e de ontologias existentes encontradas em meio a buscas feitas no portal *Linked Open Vocabularies*[2], seguindo os padrões RDF, RDF(S) e *Linked Data*.

[1] https://protege.stanford.edu/
[2] https://lov.linkeddata.es/dataset/lov/vocabs

4.2.3 Visões Semânticas Exportadas

Após concluir a modelagem da ontologia da Visão Semântica do EKG, o próximo passo envolve a construção das Visões Semânticas Exportadas das fonte de dados do *Data Lake*. Uma **Visão Semântica Exportada (VSE)** é alcançada mapeando sistematicamente os dados originais de uma fonte para o vocabulário da ontologia da Visão Semântica. Esse mapeamento estratégico é uma solução para o desafio da heterogeneidade de vocabulário que surge entre diversas fontes de dados. Ao alinhar os dados de origem com o vocabulário padronizado da ontologia da Visão Semântica, consegue-se uma representação harmonizada e unificada. Isso não só atenua a questão das terminologias díspares, mas também facilita uma integração coesa de dados de diferentes fontes, promovendo uma estrutura de conhecimento mais abrangente e interoperável.

A seguir, a Definição 4.2.1 formaliza os elementos que constituem a especificação de uma Visão Semântica Exportada. A Definição 4.2.2 estrutura e formaliza o grafo de dados inferido por uma determinada especificação de Visão Semântica Exportada, considerando o estado atual da fonte de dados.

Definição 4.2.1 (Visão Semântica Exportada) *No framework proposto, a especificação de uma Visão Semântica Exportada é uma quíntupla* $\langle \mathcal{V}, S, O_D, O_\mathcal{V}, M_\mathcal{V} \rangle$*, onde:*

- $\mathcal{V}$ *é o nome da visão.*
- S *é uma fonte de dados do Data Lake que exporta a visão.*
- O_D *é a ontologia no domínio da Visão Semântica do EKG.*
- $O_\mathcal{V}$ *é a ontologia da Visão Semântica Exportada.*
- $M_\mathcal{V}$ *é um conjunto de regras de mapeamento entre* S *e* O_D*.*

Definição 4.2.2 (Grafo da Visão Semântica Exportada) *Considere:*

- $\langle \mathcal{V}, S, O_D, O_\mathcal{V}, M_\mathcal{V} \rangle$*, a especificação da visão semântica exportada* $\mathcal{V}$*.*
- $S(t)$*, o estado da fonte* S *em um instante* t*.*

O grafo de dados de uma visão semântica exportada $\mathcal{V}$ em t, denotado $\mathcal{V}(t)$, é definido como:

$$\mathcal{V}(t) = \{(s,p,o) \mid (s,p,o) \text{ é uma tripla em } M_\mathcal{V}(S(t))\} \tag{4.1}$$

Conforme a Definição 4.2.2, o grafo de dados de uma Visão Semântica Exportada é o resultado da aplicação das Regras de Transformação contidas em $M_{\mathcal{V}}$ sobre o estado da fonte de dados $\boldsymbol{S}$ em um determinado momento $\boldsymbol{t}$. Esse grafo pode ser virtual ou materializado.

No enfoque materializado, os dados extraídos das fontes são transformados em grafo RDF conforme os mapeamentos e posteriormente armazenados em um *triplestore*. A materialização oferece vantagens como melhores tempos de resposta às consultas, já que são processadas diretamente no grafo materializado. Mas, há a necessidade de atualizar o grafo quando ocorrem alterações na fonte de dados.

No enfoque virtual, o grafo é uma visão do resultado da reescrita de consultas feitas em SPARQL no *endpoint* de acesso às fontes originais. **Grafos virtuais** são gerados dinamicamente *on-the-fly*, proporcionando acesso aos dados em tempo real. Isso garante dados atualizados, mas pode afetar o desempenho das consultas em fontes com grande volume de dados (CALVANESE *et al.*, 2017).

O *framework* proposto adota uma abordagem híbrida, combinando estrategicamente grafos virtuais e materializados.

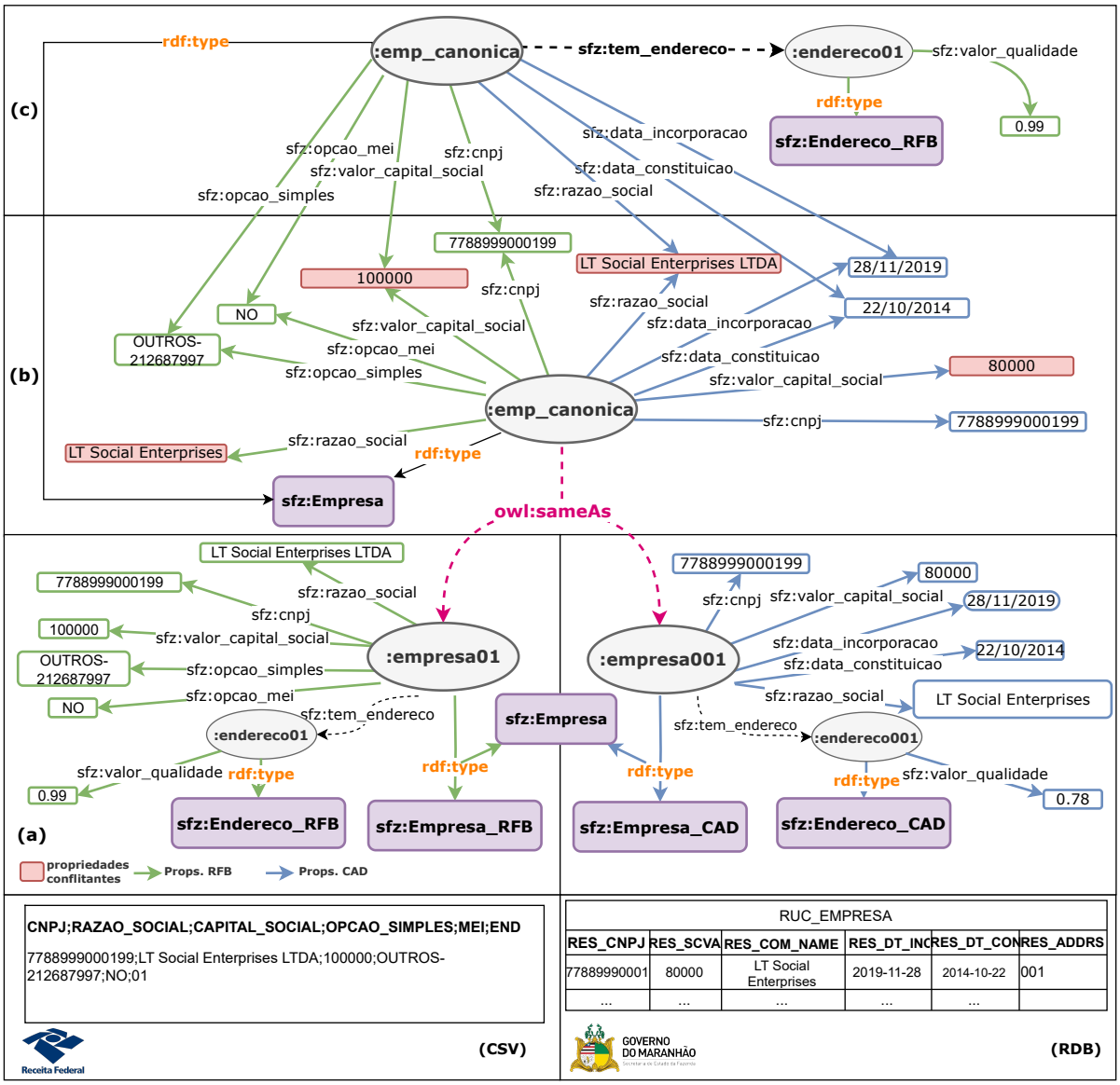
Considerando as fontes de dados CAD e RFB e a $O_{SEFAZMA}$ (Figura 8), as especificações das Visões Semânticas Exportadas podem ser definidas assim: ($\boldsymbol{V_{RFB}}$, RFB, $O_{SEFAZMA}$, O_{RFB}, M_{RFB}) e ($\boldsymbol{V_{CAD}}$, CAD, $O_{SEFAZMA}$, O_{CAD}, M_{CAD}). A materialização dessas Visões Semântica Exportadas, a partir da sua especificação, requer a transformação das terminologias da fonte de dados no vocabulário da $O_{SEFAZMA}$, conforme especificado pelas regras de transformação dos mapeamentos M_{RFB} e M_{CAD}. Essas regras de transformação incluem a construção do *International Resource Identifier* (IRI) que visa guardar o contexto da proveniência dos recursos adotando o seguinte padrão: <{URLBase}/resource/{fonte-de-dados}/{nome-da-classe}/{chave-única}>. Um exemplo desse padrão de IRI é o seguinte: **<http://www.exemplo.com.br/**resource**/CAD/Empresa/8394>**.

As implementações dos mapeamentos foram realizadas com a linguagem de transformação de dados *RDF Mapping Language*[3] (RML), uma extensão da amplamente utilizada linguagem *RDB to RDF Mapping Language*[4] (R2RML).

Referindo-se ao nosso estudo de caso, a Figura 10(a) mostra as visões semânticas exportadas das fontes de dados RFB e CAD ($\boldsymbol{V_{RFB}}$ e $\boldsymbol{V_{CAD}}$) representadas pelas instâncias *:empresa01* e *:empresa001*, respectivamente.

[3] https://rml.io/
[4] https://www.w3.org/TR/r2rml/

Figura 10 – Estudo de Caso.



Fonte: elaborada pelo autor.

4.2.4 Visões de Ligação e Visões de Unificação

No restante do trabalho, considera-se σ o atual estado das fontes de dados em um determinado instante t.

4.2.4.1 Visões de Ligação

A Visão de Ligação "*sameAs*" é crucial para alinhar e conectar fontes de dados distintas. Envolve a especificação de relacionamentos entre entidades equivalentes ou idênticas em diferentes conjuntos de dados. Esta etapa ajuda a unificar dados de várias fontes, garantindo que o grafo de conhecimento possa integrar perfeitamente todas as informações.

Como vantagem, as visões de ligação possibilitam a descoberta de novas informações e fatos entre fontes de dados diferentes, por meio do enriquecimento e agrupamento de dados seguindo os princípios *Linked Data* e a ontologia da Visão Semântica.

A seguir, a Definição 4.2.3 especifica a visão de ligação "*sameAs*", enquanto a Definição 4.2.4 formaliza e estrutura do grafo de dados inferido por uma determinada especificação de visão de ligação, considerando os estados atuais das visões semânticas exportadas das fontes de dados.

Definição 4.2.3 (Visão de Ligação) *A especificação de uma Visão de Ligação (links "sameAs") é uma tupla* $\langle \mathcal{L}, G, T, W, \mu \rangle$*, onde:*

- $\mathcal{L}$ *é o nome da visão de ligação.*
- G *é uma classe de generalização.*
- T *e* W *são classes de diferentes visões semânticas exportadas, onde* T *e* W *são subclasses de* G*.*
- μ *é uma "função de correspondência".*

A função de correspondência é projetada para comparar o estado de duas instâncias, e_1 e e_2, das classes T e W, respectivamente, e retornar "true" caso e_1 e e_2 satisfaçam a condição de "*match*" de μ; caso contrário, retorna "falso".

Definição 4.2.4 (Grafo de Dados da Visão de Ligação) *Sejam:*

- $\langle \mathcal{L}, G, T, W, \mu \rangle$*, uma especificação de uma Visão de Ligação* $\mathcal{L}$*.*
- $T(t)$ *e* $W(t)$*, os estados de* T *e* W *em* t*.*

O grafo de dados da visão $\mathcal{L}$ no intante t, denotado $\mathcal{L}(t)$, é definido como:

$$\mathcal{L}(t) = \{(r_1, sameAs, r_2) \mid r_1 \in T(t), r_2 \in W(t), e\ \mu(r_1, r_2) = true\} \tag{4.2}$$

Considerando as visões semânticas exportadas das fontes de dados V_{CAD} e V_{RFB}, uma Visão de Ligação poderia ser especificada combinando instâncias da classe de generalização *sfz:Empresa*. Como exemplo, considere a especificação de visão de ligação ($\mathcal{L}_1$, *sfz:Empresa*, V_{CAD},V_{RFB}, *exact_match*()) e use *sfz:cnpj* para definir que as instâncias *:empresa01* e *:empresa001*, entradas da função *exact_match()*, são idênticas, por meio da propriedade *owl:sameAs*. Referindo-se ao nosso estudo de caso, a Figura 10(b) mostra o *link* "*sameAs*" gerado automaticamente usando $\mathcal{L}_1$.

Nossa abordagem sugere duas formas de construir as visões de ligação. A primeira delas é pelo auxílio de ferramentas apropriadas como o Silk[5], principalmente quando não há chave única de identificação dos recursos. O Silk, por exemplo, pode combinar várias métricas de similaridade para descoberta de *links* semânticos. A outra forma é por meio de mapeamentos RML onde a consulta SQL definida engloba as regras de equivalência entre duas visões semânticas exportadas distintas.

4.2.4.2 Visão de Unificação

As Visões de Ligação desempenham um papel crucial na geração das Visões de Unificação. As Visões de Unificação reúnem recursos e informações de diversas fontes vinculadas pelo relacionamento "*sameAs*". Isso facilita a visualização e consulta de recursos em um contexto unificado, permitindo uma compreensão e análise mais abrangentes de dados interligados.

A seguir, a Definição 4.2.5 estabelece a especificação de uma Visão de Unificação, enquanto a Definição 4.2.6 fornece uma definição formal do grafo de dados inferido por uma especificação de visão de unificação, considerando os estados atuais das visões semânticas exportadas e das visões de ligação.

Definição 4.2.5 (Visão de Unificação) *A especificação de uma Visão de Unificação é uma tripla* $\langle \mathcal{U}, \mathbf{G}, \boldsymbol{\eta} \rangle$*, onde:*

- $\mathcal{U}$ *é o nome da Visão de Unificação.*
- $\mathbf{G}$ *é uma classe de generalização da ontologia da Visão Semântica.*
- $\boldsymbol{\eta}$ *é uma função de normalização, que mapeia todos os IRIs das instâncias de* $\mathbf{G}$*, que estão relacionados por links "sameAs", para um IRI alvo canônico.*

A função de normalização deve satisfazer o seguinte axioma (usando uma notação infixa para *sameAs*):

$$\forall x_1 \forall x_2 \ (x_1 \ sameAs \ x_2 \Leftrightarrow \boldsymbol{\eta}(x_1) = \boldsymbol{\eta}(x_2))$$

Portanto, os IRIs x_1 e x_2 são unificados ao mesmo IRI canônico se forem declarados equivalentes por meio de uma instrução "*sameAs*" no formato "x_1 *sameAs* x_2". A função de unificação particiona os IRIs dos recursos de visões semânticas exportadas em um conjunto de classes de equivalência. Observe que um IRI x pertence a apenas uma classe de equivalência.

[5] http://silkframework.org/

Definição 4.2.6 (Grafo de Dados da Visão de Unificação) *Sejam:*

- $\langle \mathcal{U}, G, \eta \rangle$, *a especificação de uma visão de unificação para a classe de generalização* G.
- $G(t)$, *o estado de* G *em* t.

O grafo de dados da visão de unificação de G no estado t, denotado $\mathcal{U}[G(t)]$, é definido como:

$$\mathcal{U}[G(t)] = \{(s,p,o) \mid (r,p,o) \in G(t) \text{ e } \eta(r) = s, o \text{ IRI canônico para } r\} \tag{4.3}$$

No grafo de dados de uma visão de unificação, todos os IRIs na mesma classe de equivalência são homogeneizados agrupando todas as propriedades desses IRIs em seu IRI alvo canônico. Nesse sentido, deve ser definida uma Visão de Unificação para cada **classe de generalização** da Visão Semântica. Para um IRI canônico c, a visão de unificação de c em t, denotado $\mathcal{U}(c(t))$, é definida por:

$$\mathcal{U}(c(t)) = \{(c,p,o) \mid (r,p,o) \in \mathcal{U}[G(t)] \text{ onde } \eta(r) = c\} \tag{4.4}$$

O grafo de dados das visões de unificação é virtual e a visão de unificação de um recurso, identificado por uma $\{uri\}$, é definida pela consulta SPARQL do Código-fonte 2.

Código-fonte 2 – Consulta SPARQL da Visão de Unificação de um recurso.

```
1  PREFIX owl: <htt://www.w3.org/2002/07/owl#>
2  SELECT ?object WHERE {
3     { <{uri}> ?props ?object . }
4     UNION {
5        {
6           <{uri}> owl:sameAs+ ?same.
7           ?same ?props ?object.
8     } UNION {
9           ?same owl:sameAs+ <{uri}>.
10       ?same ?props ?object.
11    }
12    } FILTER(?props != owl:sameAs)
13 } ORDER BY ?props
```

A consulta SPARQL da Visão de Unificação de um recurso seleciona todas as propriedades do recurso {uri} usando o padrão de tripla definido na linha 3.

Note que o recurso {uri} de uma fonte A pode ter um *link* "*sameAs*" com um recurso de uma fonte B. O recurso da fonte B pode ter um *link* "*sameAs*" com um recurso da fonte C. Logo, dado a inferência por transitividade, o {uri} deve ter um *link* com o recurso da fonte C. Além disso, pode existir um *link* "*sameAs*" de uma fonte D para o recurso {uri}. Nesses casos, a consulta que gera a Visão de Unificação de um recurso deve ser capaz de resolver a transitividade e o relacionamento com o caminho inverso da propriedade *owl:sameAs*.

Na consulta SPARQL da Visão de Unificação de um recurso, a transitividade e a união com as propriedades dos recursos que são objetos na tripla ({uri} *owl:sameAs*+ ?same .) são obtidas nas linhas 4-7. O *link* "*sameAs*" com o caminho inverso ou com a propriedade de entrada para o {uri} e a união com as propriedades dos recursos que são sujeitos da tripla (?same *owl:sameAs*+ {uri} .) são alcançados nas linhas 8-10.

4.2.5 Visões de Fusão

O objetivo de uma Visão de Fusão é resolver conflitos que possam surgir quando diferentes fontes dão informações divergentes sobre a mesma entidade ou relacionamento, visando melhorar a qualidade, a precisão e a confiabilidade das informações do grafo de dados **KG_VIS**.

A partir da Visão de Unificação de um recurso é possível detectar informações conflitantes de uma entidade ou relacionamento. Por exemplo, pode-se observar na Figura 10(b), o recurso *:emp_canonica*, que representa uma visão de unificação de um recurso e apresenta conflitos de informações em duas propriedades: *sfz:razao_social* e *sfz:valor_capital_social*.

A solução para esse tipo de inconsistência requer implementar mecanismos de resolução de conflitos para identificar e resolver discrepâncias. Isso pode envolver a atribuição de pontuações de confiança às fontes de dados, a consideração de aspectos temporais ou o emprego de mecanismos de votação para determinar a informação mais confiável.

No *framework* proposto, o usuário é livre para definir como resolver o problema de valores de atributos contraditórios ao combinar múltiplas representações do mesmo objeto em uma única representação (IRI canônico). Isso é especificado com a ajuda de "assertivas de fusão de dados de propriedades".

Definição 4.2.7 (Assertiva de Fusão de Dados) *Uma Assertiva de Fusão de Dados (AFD) para uma propriedade P, no contexto da classe de generalização G, é uma quádrupla* $\langle \mathcal{A}, \boldsymbol{G}, \boldsymbol{P}, \boldsymbol{\Psi} \rangle$*, onde:*

- $\mathcal{A}$ *é o nome da assertiva de fusão de dados.*
- G *é uma classe de generalização da ontologia da Visão Semântica.*
- P *é uma propriedade de* G.
- Ψ *é uma função de resolução de conflitos.*

No contexto de uma classe de generalização G, é essencial estabelecer uma AFD para cada propriedade de G onde os potenciais valores conflitantes foram identificados.

Para a classe de generalização *sfz:Empresa* do nosso estudo de caso, foram definidas as seguintes AFDs:

(i) $\langle \mathcal{A}_1$, `sfz:Empresa`, **`sfz:razao_social`**, `KeepSingleValueByReputation()`$\rangle$: resolve conflitos da propriedade razão social das instâncias de *sfz:Empresa*.

(ii) $\langle \mathcal{A}_2$, `sfz:Empresa`, **`sfz:valor_capital_social`**, `KeepSingleValueByReputation()`$\rangle$: escolhe o melhor valor para a propriedade capital social das instâncias de *sfz:Empresa*.

Todas as AFDs usam a função de resolução de conflito "`KeepSingleValueByReputation()`". Essa função é projetada para selecionar o valor da fonte de dados que é considerada a mais confiável ou de maior qualidade em relação às outras fontes disponíveis. Isso implica que cada fonte de dados tem associada a ela um índice de reputação ou qualidade, usado para avaliar sua confiabilidade. Geralmente, esse índice é calculado com base em vários fatores, como precisão histórica, confiabilidade da fonte, ou validade dos dados.

Uma função de resolução de conflitos de uma propriedade P, no contexto da classe de generalização G, aceita como entrada um conjunto de valores de P, associados com um IRI canônico em G, e produz um único valor, conforme definido a seguir.

Definição 4.2.8 (Resolução de conflitos baseada em AFD) *Sejam:*

- $\langle \mathcal{U}, G, \eta \rangle$*, a especificação de uma visão de unificação para a classe de generalização G.*
- $\langle \mathcal{A}, G, P, \Psi \rangle$*, uma assertiva de fusão de dados para a propriedade P de G.*
- $\mathcal{U}[G(t)]$*, o estado da visão de unificação de* G *em* t.
- c *um IRI canônico em* $\mathcal{U}[G(t)]$.
- $V = \{v_i \mid (c, P, v_i) \in \mathcal{U}[G(t)]$ *para* $i = (1, \ldots, n)$*, (*V *contém todos os valores da propriedade* P *associados com o IRI canônico* c*.* $\}$.

O resultado da fusão de dados da propriedade P, com base na AFD $\mathcal{A}$, para o IRI canônico c em $\mathcal{U}[G(t)]$, denotado $\mathcal{A}(c(t))$, é definida como:

$$\mathcal{A}(c(t)) = \{(c, P, v) \mid v = \Psi(V)\}$$

Para demonstrar a resolução de conflito baseada em AFD, vejamos o exemplo da AFD $\langle \mathcal{A}_1$, `sfz:Empresa, sfz:razao_social, KeepSingleValueByReputation()`$\rangle$ do nosso estudo de caso. Seguindo a Definição 4.2.8, temos:

- $\mathbf{V} = \langle$"LT Social Enterprises LTDA", "LT Social Enterprises" $\rangle$;
- $\boldsymbol{c} =$ *:emp_canonica* e;
- `KeepSingleValueByReputation()` retorna "LT Social Enterprises LTDA".

Assim,

$$\mathcal{A}_1(\textbf{:emp\_canonica}(t)) = \langle(\textbf{:emp\_canonica}, \textit{sfz:razao\_social}, \text{“}LTSocialEnterprisesLTDA''\rangle)$$

Para computar a visão de fusão de dados para um IRI canônico, no contexto de uma classe de generalização $\boldsymbol{G}$, é preciso resolver os conflitos conforme especificado por todas AFDs de $\boldsymbol{G}$, como definido a seguir.

Definição 4.2.9 (Visão de Fusão de um IRI canônico) *A especificação de uma Visão de Fusão é uma n-tupla* $\langle \mathcal{F}, G, \mathcal{A}_1, \ldots, \mathcal{A}_m \rangle$*, onde:*

- $\mathcal{F}$ *é o nome da Visão de Fusão.*
- $\boldsymbol{G}$ *é uma classe de generalização.*
- $\langle \mathcal{A}_1, \ldots, \mathcal{A}_m \rangle$ *todas as AFDs para a classe* $\boldsymbol{G}$*.*

Sejam:

- $\mathcal{U}[\boldsymbol{G}(t)]$*, o estado da visão de unificação de* $\boldsymbol{G}$ *em* t*.*
- $\boldsymbol{c}(t)$*, o estado de um IRI canônico* $\boldsymbol{c}$ *em* $\mathcal{U}[\boldsymbol{G}(t)]$*.*

O grafo da visão de fusão de dados de $\boldsymbol{c}(t)$*, denotado* $\mathcal{F}(\boldsymbol{c}(t))$*, é definida por:*

$$\mathcal{F}(c(t)) = \bigcup_{i=1}^{n} \mathcal{A}_i(c(t))$$

Deve ser definida uma Visão de Fusão para cada **classe de generalização** da Visão Semântica. Além disso, dever ser definida uma AFD para cada propriedade da classe de generalização que exista potencial de inconsistência.

O resultado da implementação da Visão de Fusão para um IRI canônico baseado na Definição 4.2.9 pode ser visto no recurso higienizado *:emp_canonica* exibido na Figura 9(c). O capítulo seguinte detalha a criação do vocabulário de metadados da Visão Semântica do EKG.

5 VOCABULÁRIO DOS METADADOS DA VISÃO SEMÂNTICA

A partir dos componentes, das especificações e da construção da **Visão Semântica** apresentados anteriormente, foi possível definir um vocabulário dedicado para representar os metadados do **KG_Meta_VIS**. As seções 5.1 e 5.2 detalham a estrutura e o uso desse vocabulário.

5.1 Estrutura do Vocabulário dos Metadados da Visão Semântica

O vocabulário dos metadados da Visão Semântica (VSKG) fornece padrões semânticos que visa servir de ponte entre os construtores e consumidores de EKGs. Além disso, pode ser usado em muitas situações, desde a descoberta de dados até a catalogação e arquivamento de conjuntos de dados, auxiliando os usuários a encontrar os dados certos para suas tarefas.

Baseado nas especificações da Visão Semântica, o vocabulário **VSKG** é implementado como um artefato OWL e é estruturado em cinco camadas, conforme ilustrado na Figura 11. Para representá-lo, foram reusados vocabulários bem conhecidos na comunidade da Web Semântica, incluindo:

- *Data Catalog Vocabulary* (DCAT)[1] (prefixo *dcat:*). DCAT é um vocabulário RDF desenvolvido para facilitar a interoperabilidade entre catálogos de dados publicados na Web. O DCAT permite que o construtor do EKG descreva os conjuntos e serviços de dados em um catálogo usando um modelo padrão e um vocabulário que facilita o consumo e a agregação de metadados do *Data Lake* (DIBOWSKI *et al.*, 2020).
- *Data Reference Model*[2] (prefixo *drm:*). Um metamodelo para dados governamentais, publicado pelo Governo dos Estados Unidos.
- *RDB to RDF Mapping Language Schema*[3] (prefixo *rr:*). Um vocabulário que pode ser usado para especificar um mapeamento de dados relacional para RDF, publicado pelo W3C. Esse vocabulário é composto por 15 classes e 25 propriedades.
- *Vocabulary of Interlinked Data*[4] (prefixo *void:*). Esse vocabulário é um esquema RDF para expressar metadados sobre *datasets* ou grafos RDF inteconectados, também publicado pelo W3C.
- *Data Quality Vocabulary*[5] (prefixo *dqv:*). O Vocabulário de Qualidade de Dados (DQV) é

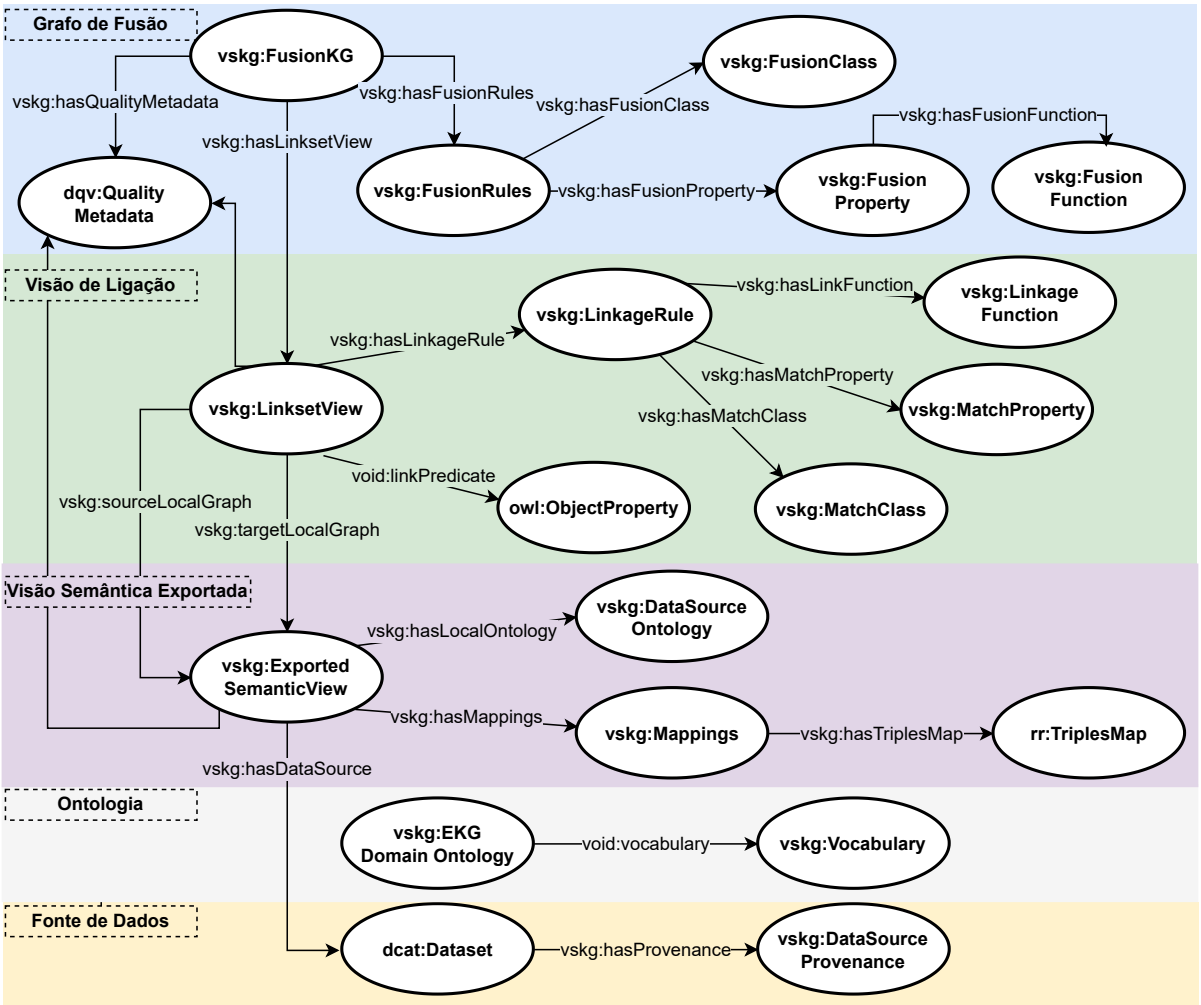
[1] https://www.w3.org/TR/vocab-dcat-3/
[2] https://lov.linkeddata.es/dataset/lov/vocabs/drm
[3] https://www.w3.org/ns/r2rml#
[4] https://www.w3.org/TR/void/
[5] https://www.w3.org/TR/vocab-dqv/

visto como uma extensão do DCAT para cobrir a qualidade dos dados.

– *Vocabulary Of Attribution and Governance* (VOAG)[6] (prefixo *voag:*). VOAG destina-se a especificar licenciamento, atribuição, proveniência e governança de uma ontologia.
– *MOD: Metadata for Ontology Description and Publication*[7] (prefixo *mod:*). Publicado pelo Instituto Indiano de Estatística, MOD é um vocabulário de metadados para descrever e publicar ontologias.

Além dos prefixos listados acima, foi adotado o prefixo ***"vskg:"*** (*Vocabulary of Semantic Knowledge Graph*) para representar novas classes, propriedade e relacionamentos. As camadas do vocabulário de metadados da Visão Semântica são descritas a seguir:

Figura 11 – Vocabulário dos metadados da Visão Semântica.



Fonte: elaborada pelo autor.

• **Camada do VSKG para Fonte de Dados:** contém as classes e propriedades que ajudam a descrever os metadados das fontes de dados e representar a proveniência de tais fontes. Os metadados

[6] https://lov.linkeddata.es/dataset/lov/vocabs/voag
[7] https://www.isibang.ac.in/ns/mod/index.html

capturados nessa camada registram informações essenciais, como a origem dos dados, frequência de atualização, esquema, autoria, dicionário de dados e até *scripts* SQL relacionados, são eles: ⟨*dcat:Dataset,vskg:DataSourceProvenance, vskg:urlDataDictionary,vskg:urlDataSchemaDiagram, pav:providedBy, voag:frequencyOfChange, vskg:urlScript, foaf:page, vskg:hasProvenance*⟩.

• **Camada do VSKG para Ontologia:** inclui a classe de metadados que especifica a ontologia da Visão Semântica do EKG e a classe que instancia os vocabulários importados utilizados, bem como propriedades de relacionamentos, no seguinte conjunto: ⟨*vskg:Vocabulary, vskg:hasDomainOntology, dcterms:format, dcterms:language, vann:preferredNamespacePrefix, vskg:uriOntology, dcterms:hasVersion, void:vocabulary, vskg:EKGDomainOntology*⟩.

• **Camada do VSKG para Visão Semântica Exportada:** possui classes de metadados essenciais que instanciam as visões semânticas exportadas pelas fontes; que especificam as ontologias das fontes exportadas; e, que descrevem os mapeamentos que transformam essas fontes. Classes e propriedades desta camada incluem: ⟨*vskg:ExportedSemanticView, vskg:DataSourceOntology, vskg:Mappings, vskg:hasDataSourceOntology, vskg:hasMappings, vskg:hasTriplesMap*⟩.

• **Camada do VSKG para Visão de Ligação:** os vocábulos desta camada descrevem as visões de ligação semântica; as regras de correspondência, assim como as Visões Semânticas Exportadas das fontes de dados (origem e alvo) que têm em comum uma classe de generalização semântica. Vejamos os termos desta camada: ⟨*vskg:LinksetView, vskg:LinkageRules, vskg:sourceSemanticView, vskg:targetSemanticView, void:linkPredicate, vskg:hasLinkageRule, vskg:hasMatchProperty, vskg:hasMatchClass, vskg:hasLinkFunction*⟩.

• **Camada do VSKG para Grafo de Fusão:** os termos aqui são usados para declarar grafos de fusão (Visão de Fusão) e as regras de fusão do grafo. Entre as regras de fusão estão a definição da classe e propriedades de fusão, além da função de fusão. Entre os termos desta camada, estão: ⟨*vskg:FusionKG, vskg:FusionRules, vskg:FusionClass, vskg:FusionProperty, vskg:FusionFunction, vskg:hasLinksetView, vskg:hasFusionRules, vskg:hasFusionProperty, vskg:hasFusionClass, vskg:hasFusionFunction*⟩.

5.2 Uso do Vocabulário VSKG

Resumidamente, o vocabulário de metadados VSKG é fundamental para a compreensão, gerenciamento e aproveitamento eficaz do conhecimento contido na **Visão Semântica** e, segundo Bird *et al.* (2016), metadados podem ser usados ativamente, como, no processo de construção, manutenção e consumo da **Visão Semântica**.

Na inserção de uma nova fonte de dados, por exemplo, será possível realizar ações como:

(i) sugerir mapeamentos para a construção da visão semântica exportada da nova fonte, baseado nos metadados dos mapeamentos prévios e na hierarquia de classes de generalização;

(ii) identificar quais as visões de ligação devem ser definidas conforme os metadados de equivalência de classes e sugerir a função de correspondência a ser utilizada.

O vocabulário VSKG é um artefato importante que permite que os ativos do EKG sejam encontrados, acessados, transitados e reutilizados por humanos e, principalmente, por máquinas. Com relação à descoberta de dados na **Visão Semântica**, o VSKG pode ser utilizado para responder Questão de Competência como:

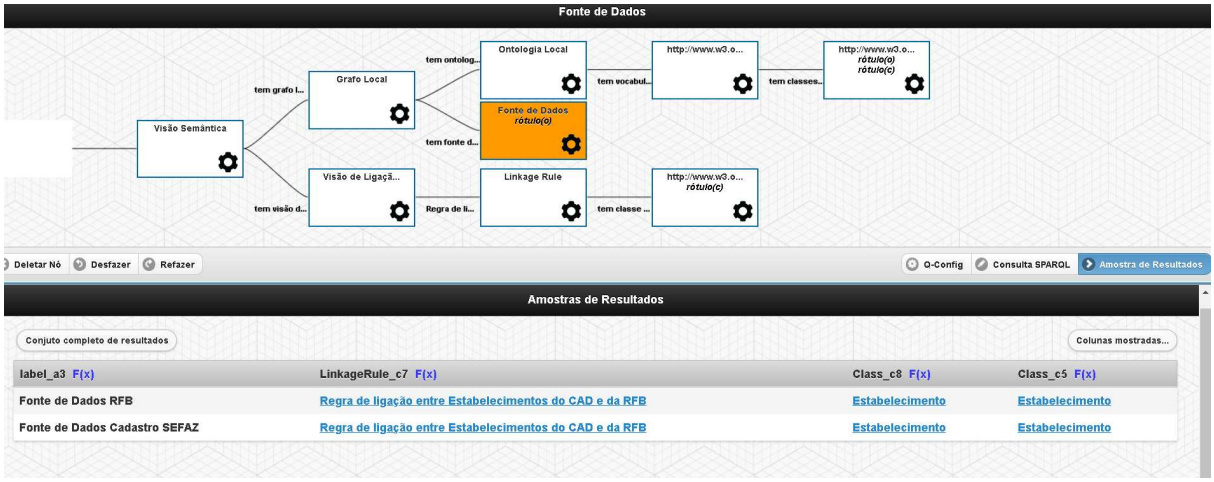
– "Quais fontes de dados compõem a camada de dados do EKG?"
– "Qual a frequência de atualização da fonte de dados $\boldsymbol{F}$?"
– "Quais Visões Semânticas Exportadas pertencem ao *cluster* da classe de generalização $\boldsymbol{X}$?"
– "Quais os mapeamentos que aplicam regras de transformação para a classe $\boldsymbol{Y}$?"
– "Qual a função de fusão usada para resolver conflitos da propriedade $\boldsymbol{Z}$ da classe de generalização $\boldsymbol{X}$?"
– "Qual o padrão de IRI usado no mapeamento da classe de Visão Semântica Exportada $\boldsymbol{W}$?"

As Figuras 12 e 13 mostram consultas semânticas ao **KG_Meta_VIS** usando o vocabulário de metadados VSKG. Essas consultas foram construídas com o auxílio da ferramenta gráfica visual OptiqueVQS (SOYLU *et al.*, 2018). Ferramentas desse tipo auxiliam na elaboração de consultas SPARQL complexas.

A primeira consulta (Figura 12), visa responder a seguinte questão de competência: "Quais fontes de dados tem *link* semântico cuja classe é equivalente com a classe de generalização `Estabelecimento`?". A segunda consulta (Figura 13), por sua vez, busca responder a seguinte questão de competência: "Quais mapeamentos possuem tabelas que correspondem com a classe

de generalização `Estabelecimento`?”.

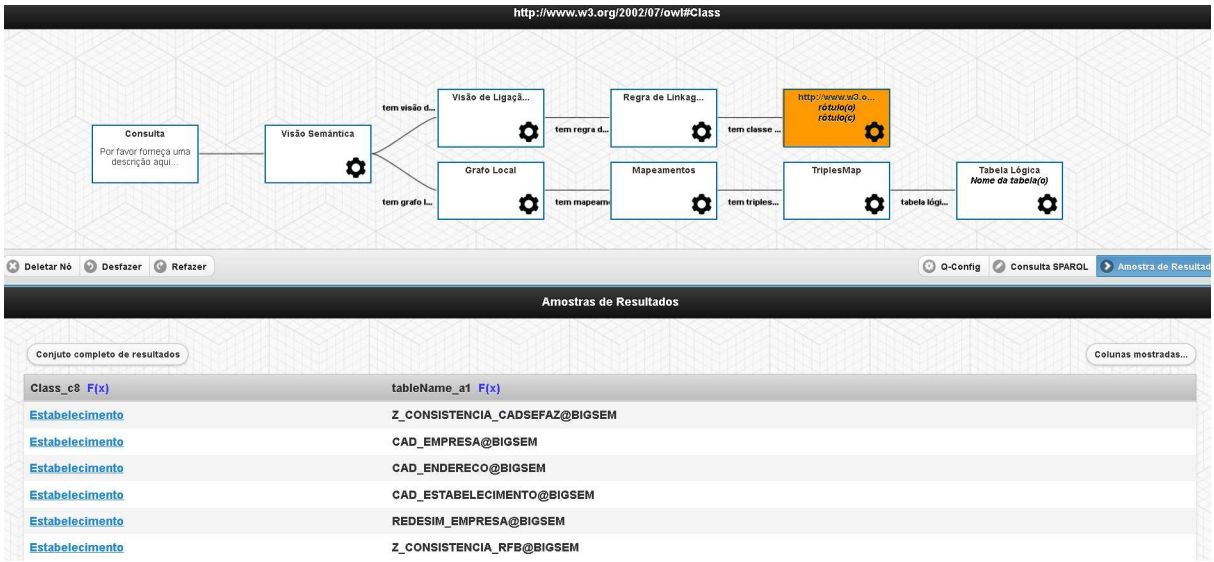
Figura 12 – Consulta 1 ao KG_Meta_VIS.



Fonte: Ferramenta OptiqueVQS.

A Consulta 1 (Figura 12) verifica quais recursos que são do tipo Visão Semântica Exportada (*vskg:ExportedSemanticView*), exportados por fonte de dados (*dcat:Dataset*) cuja ontologia da fonte de dados (*vskg:DataSourceOntology*) tem a classe Estabelecimento (*sfz:Estabelecimento*). Paralelamente, averigua quais regras de ligação (*vskg:LinkageRules*) da Visão de Ligação (*vskg:LinksetView*) tem a classe de equivalência (*vskg:hasMatchClass*) Estabelecimento (*sfz:Estabelecimento*). O conhecimento no **KG_Meta_VIS**, diz que a “Fonte de Dados RFB” e a “Fonte de Dados Cadastro SEFAZ” têm em comum a classe Estabelecimento.

Figura 13 – Consulta 2 ao KG_Meta_VIS.



Fonte: Ferramenta OptiqueVQS.

A segunda consulta (Figura 13) afere os mapeamentos (*vskg:Mappings*) que geram as Visões Semânticas Exportadas cujas tabelas das fontes de dados (*rml:LogicalSource*) são mapeadas para a classe Estabelecimento (*sfz:Estabelecimento*).

O vocabulário VSKG foi criado para lidar com os metadados da **Visão Semântica** apropriadamente, a fim de facilitar seu gerenciamento e uso. Além disso, outras aplicações podem ser desenvolvidas na Camada de Aplicações do EKG (Figura 1) e consumir o vocabulário VSKG para automatizar algum processo de construção ou análises.

6 FERRAMENTA *EKG CONTEXT EXPLORER*

Ciente dos desafios de exploração de grafos de conhecimento descritos na Seção 2, este estudo implementou a ***EKG Context Explorer***, uma interface gráfica interativa que possibilita a exploração baseada em contexto do grafo de conhecimento da **Visão Semântica** de EKGs.

A exploração baseada em contexto considera, por exemplo, o ambiente ou a situação específica em que o conhecimento existe e como ele foi construído. Isso pode incluir fatores como estrutura e organização do grafo, o local de origem dos dados, níveis de visão, rastreamento de transformações ou qualquer outra informação que afete o significado ou a relevância dos dados. Além disso, a ferramenta *EKG Context Explorer* apresenta as seguintes vantagens:

- É uma aplicação que necessita apenas de um navegador *web* conectado a uma rede, reduzindo significativamente qualquer complexidade associada à instalação de *software(s)*.
- Propicia baixa curva de aprendizado, porque apresenta uma interface intuitiva baseada em textos, semelhante às tradicionais páginas HTML.
- É baseada em consultas SPARQL, rápidas e otimizadas, ideais para GC virtuais.
- Pode ser usada em qualquer domínio do conhecimento.

6.1 Arquitetura da ferramenta *EKG Context Explorer*

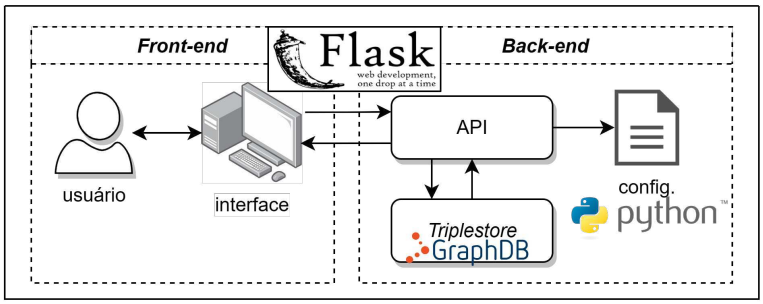
A ferramenta *EKG Context Explorer* foi desenvolvida com a linguagem Python, utilizando o *framework* Flask[1] para prover um serviço *web* leve. A *EKG Context Explorer* pode ser hospedada em um servidor com o interpretador Python e acessada por qualquer dispositivo via navegador *web* conectado à rede. Atua como uma interface sobre *endpoints* SPARQL, realizando consultas em SPARQL para acessar os dados e metadados.

A Figura 14 exibe a arquitetura base da ferramenta *EKG Context Explorer*, dividia conceitualmente em duas partes: *back-end* e *front-end*. O *back-end* é responsável por gerenciar as requisições, transacionar os dados, produzir e retornar algum resultado. Conta com uma *API RESTful*, um banco de dados de armazenamento e manipulação de triplas (*triplestore*) e um arquivo de configuração.

Para o *triplestore*, a *EKG Context Explorer* utiliza a versão gratuita do GraphDB[2] da Ontotext. O arquivo de configuração denominado “config.py” é essencial para o perfeito funcionamento da ferramenta. Esse arquivo armazena informações que possibilitam a utilização

[1] https://flask.palletsprojects.com/en/3.0.x/
[2] https://graphdb.ontotext.com/documentation/10.4/

Figura 14 – Arquitetura da ferramenta *EKG Context Explorer*.



Fonte: elaborada pelo autor.

da ferramenta genericamente, ou seja, o uso é independente do domínio do Grafo de Conhecimento. Além disso, o arquivo "config.py" comporta as definições de preferências do usuário. O *front-end*, no que lhe diz respeito, tem a incumbência de apresentar as informações da **Visão Semântica**, bem como proporcionar interação com uma boa experiência de usuário.

6.2 Navegação baseada em contexto

Com a ferramenta *EKG Context Explorer*, primeiro, é selecionada uma classe da ontologia da Visão Semântica. Então, os recursos da classe selecionada podem ser explorados em diferentes contextos.

6.2.1 Seleção de Classe

A exploração da **Visão Semântica** começa na tela de seleção de classes da ferramenta (Figura 15), exibindo as classes da ontologia da Visão Semântica, rótulos, descrições e o vocabulário (URI) que as representam. A tela de seleção de classes é dividida em duas partes: as **Classes de Generalização** (C1) e as **Classes das Visões Semânticas Exportadas** (C2).

Figura 15 – Tela seleção de classes.

Fonte: Ferramenta *EKG Context Explorer*.

O usuário pode iniciar a navegação selecionando uma classe de generalização ou de visão semântica exportada de uma fonte de dados específica. No caso de selecionar uma de classe generalização C, a tela de exploração inicia exibindo a lista de recursos que são instâncias da classe C. Neste caso, por C ser de generalização, os recursos podem ser provenientes de diversas fontes, como visto na (Figura 16). Portanto, é mostrado o "nome" da fonte de dados ao lado de cada recurso. Dessa forma, o usuário adquire o conhecimento apropriado para selecionar um recurso desejado no contexto de uma fonte específica.

Figura 16 – Recursos da classe de generalização *sfz:Empresa*.

Fonte: Ferramenta *EKG Context Explorer*.

Caso o usuário escolha uma classe C de uma visão semântica exportada V, a tela de exploração inicia com uma lista dos recursos que são instâncias de C (Figura 17). Note que C é uma classe da visão semântica exportada da fonte V. Assim, o nome da visão semântica exportada V deve estar explícito com o recurso.

Figura 17 – Recursos da classe específica *sfz:Empresa_CAD*.

Fonte: Ferramenta *EKG Context Explorer*.

6.2.2 Exploração dos Recursos

Dado que o usuário escolhe um recurso $\boldsymbol{R}$, de uma classe de generalização ou visão semântica exportada da fonte de dados, a exploração se inicia no contexto da visão semântica da fonte indicada ao lado do recurso $\boldsymbol{R}$. A tela de exploração de recursos mostra as informações no contexto da visão semântica do recurso $\boldsymbol{R}$.

A Figura 18, por exemplo, exibe a Visão Semântica Exportada da empresa "EMPRESA FICTÍCIA EXEMPLO", no contexto da origem dos dados da fonte CAD. O elemento HTML retangular vermelho com bordas arredondadas apresenta o contexto atual de exploração, que no exemplo da imagem é "contexto: Cadastro_SEFAZ-MA".

O botão "*Visual Graph*" redireciona para uma página do *triplestore* GraphDB que exibe de forma visual gráfica o grafo do recurso selecionado apresentado nós (círculos) e arestas (setas direcionadas).

Figura 18 – Tela de exploração de recurso. Contexto Visão Semântica Exportada.

Fonte: Ferramenta *EKG Context Explorer*.

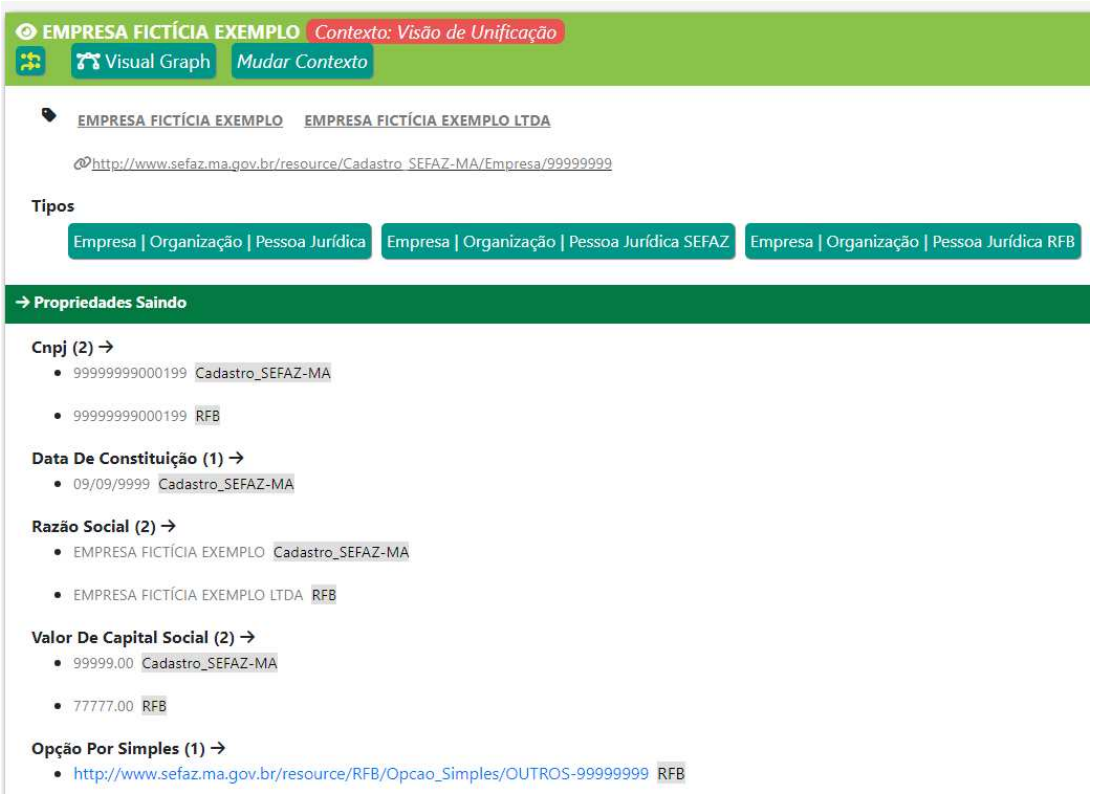
Com a *EKG Context Explore*, existe a possibilidade de acessar as informações de um recurso a partir de outro contexto utilizando o menu de navegação de contexto exibido na Figura 18(M). Esse menu é acionado ao clicar na opção "Mudar Contexto" e apresenta opções de contextos que dependem da existência de ligações "*sameAs*" com recursos de outras visões semânticas exportadas. Por exemplo, a empresa "EMPRESA FICTÍCIA EXEMPLO" na Figura 18 tem quatro opções de contextos diferentes:

– A Visão Semântica Exportada pela RFB.

– A Visão Semântica Exportada pela CAD.
– A Visão de Unificação.
– A Visão de Fusão.

A Figura 19 mostra a Visão de Unificação da empresa “EMPRESA FICTÍCIA EXEMPLO”. No segundo nível da arquitetura da **Visão Semântica**, essa visão apresenta informações no contexto da RFB e CAD.

Figura 19 – Tela de exploração de recurso. Contexto Visão de Unificação.



Fonte: Ferramenta *EKG Context Explorer*.

Nesse contexto de Visão de Unificação, todas as classes que essa empresa pertence são exibidas na área dos “Tipos” do recurso. Além disso, algumas propriedades podem se apresentar mais de uma vez com valores ou objetos distintos, como a propriedade “Razão Social”. Isso facilita a observação das divergências.

A Figura 20 mostra a tela de exploração de recurso na Visão de Fusão da empresa “EMPRESA FICTÍCIA EXEMPLO”, onde apresenta somente as triplas resultantes das assertivas de fusão de dados da empresa no contexto das fontes RFB e CAD.

Figura 20 – Tela de exploração de recurso. Contexto Visão de Fusão.

Fonte: Ferramenta *EKG Context Explorer*.

6.2.3 TimeLine dos Recursos

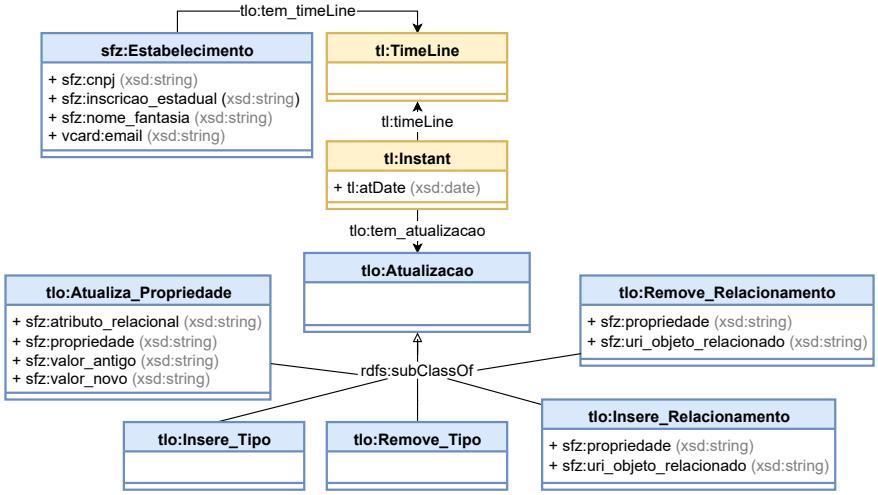
A ferramenta ***EKG Context Explorer*** dispõe, ainda, da opção de visualizar a linha do tempo dos recursos (*TimeLine* dos Recursos). As *TimeLines* dos Recursos são geradas a partir de tabelas de histórico. Essas tabelas de histórico são processadas a cada nova carga de uma fonte de dados que faz parte do *Data Lake* do EKG. Os detalhes sobre os processamentos dos históricos não fazem parte do escopo deste trabalho e, consequentemente, não são aqui detalhados.

As *TimeLines* dos Recursos também são estruturadas em grafos de conhecimento semânticos. Suas triplas, que descrevem as mudanças nos recursos ao longo do tempo, são construídas seguindo os conceitos e relacionamentos definidos pela ontologia **TLO_SEFAZMA**, exibida na Figura 21. Essa ontologia reutiliza o vocabulário da ontologia *The TimeLine Ontology*[3] cujo prefixo é **"tl:"** e possue termos específicos para representar as mudanças sofridas por um recurso, adotando, para isso, o prefixo **"tlo:"**.

Na ontologia **TLO_SEFAZMA**, todo recurso do tipo *sfz:Estabelecimento* tem uma linha do tempo *tl:TimeLine*. A classe *tl:Instant* registra a data que uma atualização foi efetuada e pode conter n eventos de atualização. Para o conceito de atualização (*tlo:Atualizacao*), existem cinco subtipos de classes mais específicas que auxiliam na identificação dos tipos de eventos acometidos ao recurso, explicadas a seguir:

[3] https://motools.sourceforge.net/timeline/timeline.html

Figura 21 – Ontologia TLO_SEFAZMA.

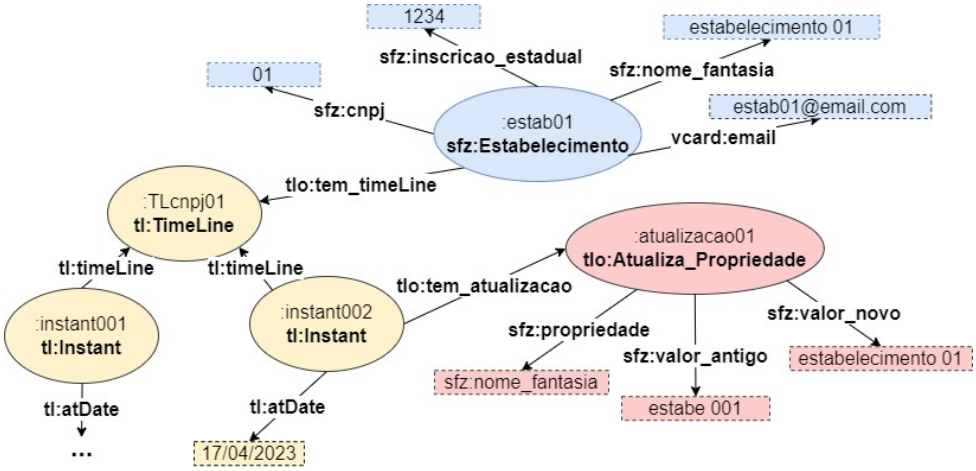


Fonte: elaborada pelo autor.

1. ***tlo:Atualiza_Propriedade*** é a classe da ontologia **TLO_SEFAZMA** que representa a atualização de uma propriedade de dados registrando o atributo de um documento ou a coluna de uma tabela relacional que foi atualizada, exibindo o valor antigo e o valor atual. Informa, também, a propriedade da ontologia da Visão Semântica que correspondente à propriedade de dados modificada.
2. ***tlo:Insere_Tipo*** representa, na linha do tempo, a classe de atualização de recurso que indica quando um novo recurso é adicionado na fonte de dados, e posteriormente materializado no grafo. Tem-se, portanto, a inserção de um novo nó no grafo da **Visão Semântica**, instanciado com o tipo desta classe.
3. ***tlo:Remove_Tipo*** classifica uma atualização ou mudança de um recurso quando ele deixa de existir em uma fonte de dados, por motivo qualquer que seja. Além disso, esse evento registra na linha do tempo uma nova classe para o recurso removido, dependendo do tipo. Para os estabelecimentos e empresas, por exemplo, usam-se as classes *tlo:Estabelecimento_Removido* e *tlo:Empresa_Removida*. Note o seguinte padrão para esse tipo de classe: <tlo:{classe-de-generalização}_Removida(o)>.
4. ***tlo:Insere_Relacionamento*** é a classe de atualização designada para representar as mudanças nos relacionamentos de um recurso no grafo de *TimeLines*, quando um recurso estabelece um novo relacionamento com outro recurso. A linha do tempo, nesse caso, mantém o registro da propriedade do recurso, do tipo *owl:ObjectProperty*, que sofreu a atualização e da IRI do objeto relacionado.
5. ***tlo:Remove_Relacionamento*** é a subclasse de *tlo:Atualização* utilizada para conservar na linha do tempo a informação sobre um relacionamento que deixou de existir entre dois

recursos. Por exemplo, quando um sócio deixa uma sociedade ou quando um estabelecimento não pertence mais a uma empresa.

Figura 22 – Exemplo de grafo de *timeline*.



Fonte: elaborada pelos autores.

A Figura 22 exibe um exemplo de triplas de *timeline* sob a óptica da ontologia **TLO_SEFAZMA**. A atualização de propriedade de dados *:atualizacao01* associada ao instante *:instante02* na data (*tl:atDate*) "17/04/2023", registra os valores, antigo e novo, da propriedade *sfz:nome_fantasia*. Os instantes de tempos *:instant001* e *:instant002* pertencem à linha do tempo *:TLcnpj01* do estabelecimento *:estab01*. A atualização no *:instant01* registra o conhecimento de que o nome fantasia do *:estab01* passou de "estabe 001" para "estabelecimento 01".

A tela de exibição de *timeline* dos recursos (Figura 23) apresenta os eventos sofridos por um recurso selecionado, agrupados cronologicamente. A linha do tempo é organizada em instantes *tl:Instant*, onde cada instante possui uma data e sua lista de atualizações. Para atualizações de atributos, os valores antigos e novos são exibidos como na Figura 23(1), assim como as atualizações de relacionamentos são mostradas como na Figura 23(2).

Figura 23 – Tela de *timeline* de recurso. Contexto Visão de Unificação

Fonte: Ferramenta *EKG Context Explorer*.

Se o contexto da Visão de Unificação estiver ativado, a tela de *timeline* dos recursos exibirá as atualizações sofridas por todas as diferentes versões do recurso selecionado. Isso é visto na Figura 23 onde há, na mesma tela, instante da visão semântica exportada da fonte de dados do CAD, representada por um componente HTML retangular azulado, cuja proveniência é vista pelo conteúdo "Cadastro_SEFAZ-MA" e instante da visão semântica exportada da fonte de dados da RFB, esse representado por outro componente HTML retangular rosado e proveniência "RFB".

No contexto da construção **Visão Semântica**, as *timelines* podem ser usadas para rastrear a linhagem de dados. A linhagem de dados ou *Data Lineage* em inglês, na sua forma mais geral, descreve de onde vieram os dados, como foram derivados e como foram atualizados ao longo do tempo (IKEDA; WIDOM, 2009). Nas diferentes abordagens existentes, *Data Lineage* é rastreada no nível das fontes de dados e metadados, registrando informações sobre autoria, versionamento e derivações das fontes de dados na totalidade. Entretanto, no *framework* proposto neste trabalho, *Data Lineage* é rastreada no nível dos recursos das instâncias, permitindo-nos rastrear a proveniência e o histórico de atualizações de propriedades para cada instância individualmente.

Este artifício permite visualizar a fonte de dados de cada atualização sofrida por um recurso, bem como o estado de um recurso (suas propriedades) em cada etapa da construção da **Visão Semântica**.

Como exemplo da aplicação da visualização da linha do tempo no rastreamento da *Data Lineage*, ao considerar o estudo de caso na Figura 10, visualizar a empresa em questão no nível Visão de Fusão mostraria que o valor consolidado da propriedade *sfz:razao_social* seria "LT Social Enterprises LTDA", sendo este valor proveniente da visão semântica exportada pela fonte de dados da RFB. Se o usuário optar por verificar os *timelines* dos recursos no nível da fonte de dados, a visualização mostrará que na fonte de dados CAD, esse recurso tem o valor "LT Social Enterprises". Tal valor conflita com o valor "LT Empreendimentos Sociais LTDA" apresentado na fonte de dados da RFB, que foi escolhido como valor higienizado. Dessa forma, a visualização da linha do tempo é um recurso importante na compreensão do processo de construção, higienização e fusão dos dados.

Uma observação importante ao considerar o uso da ferramenta *EKG Context Explorer* está relacionada ao acoplamento ou a implantação em um ambiente ou infraestrutura tecnológica de gerenciamento de dados e conhecimento para recuperação integrada de informações e análises.

Dois pontos são ressaltados.

De um lado, se uma organização deseja utilizar a ferramenta *EKG Context Explorer* com um Grafo de Conhecimento (GC) personalizado, que não foi construído com nosso *framework*, isso não será possível. A ferramenta foi projetada seguindo a arquitetura desenvolvida exclusivamente para a navegação em vários níveis de visões baseadas em contextos. Portanto, se o grafo de conhecimento não aderir a essa arquitetura, a integração com o *EKG Context Explorer* é inviável.

Por outro lado, a ferramenta é facilmente acoplada a um GC que foi construído seguindo nosso *framework*, bastando alterar as informações pertinentes no arquivo de configuração.

Em resumo, certifique-se de que o GC esteja alinhado com a arquitetura proposta nesta dissertação para aproveitar ao máximo a ferramenta *EKG Context Explorer* e obter resultados satisfatórios.

7 CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho, foi apresentado um *framework* para construção e exploração da **Visão Semântica** de *Enterprise Knowledge Graphs* (EKGs) usando tecnologias de Web Semântica. Cada nível da arquitetura do Grafo de Conhecimento da Visão Semântica foi detalhado, apresentando uma especificação de seus elementos, onde foi utilizado um estudo de caso real para demonstrar como o grafo de dados é integrado com base na especificação da Visão Semântica.

Este trabalho também implementou um vocabulário dos metadados da Visão Semântica. O uso desse vocabulário para descobrir informações mostrou-se relevante ao responder questões de competência, além propor suporte na construção incremental da Visão Semântica.

Também foi demonstrada a exploração da Visão Semântica nos diferentes níveis e contextos utilizando a ferramenta *EKG Context Explorer*. A ferramenta *EKG Context Explorer* facilitou a interpretação e manipulação das informações, permitindo a identificação da origem dos recursos e propriedades de diversas fontes de dados, graças a arquitetura de três níveis proposta para o grafo de conhecimento da Visão Semântica. Funções, componentes HTML e consultas SPARQL foram implementadas para navegar nos contextos da arquitetura e proporcionar uma melhor experiência para o usuário.

Outros enfoques e arquiteturas de construção da Visão Semântica criam a ontologia de maneira diferente, bem mais genérica, dificultando a navegação em contexto. Usando as fontes de dados do nosso estudo de caso como exemplo, os dados de empresas da RFB e CAD seriam mapeados diretamente em instâncias da classe Empresa, sem hierarquia de classes ou *cluster* de equivalência. Além disso, as ligações semânticas seriam criadas genericamente, entre instâncias de Empresa. Nesse caso, seria perdida a linhagem dos dados e o contexto da origem das informações, impossibilitando a navegação em contexto e o entendimento do processo de integração dos dados.

No nosso *framework*, por outro lado, a arquitetura do GC da Visão Semântica se baseia em níveis de visões e na ontologia da Visão Semântica que é construída com hierarquia de classes que se equivalem semanticamente. Nossas regras de ligações semânticas são definidas no contexto das Visão Semântica Exportada. As regras de fusão são definidas no contexto das propriedades divergentes em um estado da Visão de Unificação de uma classe de generalização.

A combinação de todos esses elementos do *framework* apresentado e nossa arquitetura da Visão Semântica torna possível: (i) suportar um processo *pay-as-you-go* de garantindo uma integração de dados eficiente e flexível; (ii) *tracking Data Lineage*; (iii) permitir a construção

de *Mashups* de dados de alta qualidade e; (i) possibilitar a exploração dos recurso em diferentes contextos.

A capacidade de explorar recursos em diversos contextos de um EKG se mostrou muito importante para a análise organizacional, e isso se tornou possível através da nossa arquitetura projetada para a Visão Semântica. Importante salientar que a Visão Semântica construída com o *framework* atende os requisitos de interoperabilidade semântica, escalabilidade, desempenho de consulta, e linhagem de dados.

Para trabalhos futuros propõem-se estender o *framework* para:

- Permitir a atribuição de métricas de qualidade às visões dos três níveis da Visão Semântica.
- Facilitar a construção de grafos de conhecimento especializados de alta qualidade.
- Apoiar a evolução e manutenção incremental.
- Implementação do grafo de conhecimento no enfoque hibrido.

A arquitetura da Visão Semântica garante dados e metadados, essenciais para a construção de aplicações diversas de análise, treinamento de modelos e outros.

Nesse sentido, como pesquisas futuras, sugerimos uma investigação sobre a construção e a manutenção de *Mashup* de Dados aproveitando o Grafo de Conhecimento de Metadados (Mohammadi, 2022) da **Visão Semântica**. Os *Mashup* de Dados devem fornecer informações valiosas para melhorar a flexibilidade, a precisão e a eficiência das operações realizadas no mundo corporativo. A perspectiva é que o Grafo de Conhecimento de Metadados possa ajudar a resolver desafios críticos de gerenciamento de dados, construção, manutenção e reutilização do conhecimento especializado.

REFERÊNCIAS

ALMEIDA, M. B. Uma abordagem integrada sobre ontologias: Ciência da informação, ciência da computação e filosofia. **Perspectivas em Ciência da Informação**, v. 19, n. 3, p. 242–258, 2014.

ALMEIDA, M. B. **Ontologia em Ciência da Informação. Teoria e Método**. Curitiba: Editora CRV, 2020.

ARRUDA, N.; VENCESLAU, A.; CRUZ, M. M. Lima da; VIDAL, V. M.; PEQUENO, V. M. Publishing and consuming semantic views for construction of knowledge graphs. *In:* **22nd International Conference on Enterprise Information Systems (ICEIS)**. [*S. l.*]: SCITEPRESS Digital Library, 2020. p. 197–204.

AVILA, C.; VIDAL, V. Lirb: Um navegador leve baseado em texto para knowledge graphs rdf. *In:* **Anais Estendidos do XXXVIII Simpósio Brasileiro de Bancos de Dados**. Belo Horizonte: SBC, 2023. p. 102–107.

AZIZI, S. **Documenting Data Integration Using Knowledge Graphs**. Dissertação (Mestrado em Ciência da Computação) – Gottfried Wilhelm Leibniz Universität, Hannover, 2023.

BERNERS-LEE, T.; CHEN, Y.; CHILTON, L.; CONNOLLY, D.; DHANARAJ, R.; HOLLENBACH, J.; LERER, A.; SHEETS, D. Tabulator: Exploring and analyzing linked data on the semantic web. *In:* **Proceedings of the 3rd international semantic web user interaction workshop**. Athens: [*S. n.*], 2006. p. 159.

BIKAKIS, N.; SELLIS, T. Exploration and visualization in the web of big linked data: A survey of the state of the art. **6th International Workshop on Linked Web Data Management (LWDM 2016)**, 2016.

BIRD, C.; COLES, S.; GARRELFS, I.; GRIFFIN, T.; HAGDORN, M.; KLYNE, G.; MINETER, M.; WILLOUGHBY, C. Using metadata actively. **International Journal of Digital Curation**, v. 11, 2016.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data: Principles and state of the art. *In:* CITESEER. **World wide web conference**. [*S. l.*], 2008. v. 1, p. 40.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. **Linked data:** the story so far. *In:* **Semantic services, interoperability and web applications: emerging concepts**. [*S. l.*]: IGI Global, 2011. p. 205–227.

BONATTI, P. A.; DECKER, S.; POLLERES, A.; PRESUTTI, V. Knowledge graphs: New directions for knowledge representation on the semantic web (dagstuhl seminar 18371). *In:* **Dagstuhl reports**. [*S. l.*]: Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

BORST, W. N. **Construction of engineering ontologies for knowledge sharing and reuse**. Tese (Doutorado em Ciência da Computação) – University of Twente, Netherlands, 1997.

CALVANESE, D.; COGREL, B.; KOMLA-EBRI, S.; KONTCHAKOV, R.; LANTI, D.; REZK, M.; RODRIGUEZ-MURO, M.; XIAO, G. Ontop: Answering sparql queries over relational databases. **Semantic Web**, IOS Press, [*S. l.*], v. 8, n. 3, p. 471–487, 2017.

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 do Ceará, Departamento de Telemática, PPGCC, Fortaleza, 2017.

CRUZ, M. M. L. da. **Uma abordagem para construção de mashups de dados especificados como uma visão sobre um EKG**. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal do Ceará, Fortaleza, 2021.

CRUZ, M. M. L. da; AVILA, C. V. S.; VIDAL, V. M. P.; JUNIOR, N. M. A. Semanticsus: Um portal semântico baseado em ontologias e dados interligados para acesso, integraçao e visualizaçao de dados do sus. *In:* **Anais Estendidos do XIX Simpósio Brasileiro de Computação Aplicada à Saúde**. Niterói: SBC, 2019. p. 13–18.

CYGANIAK, R.; WOOD, D.; LANTHALER, M.; KLYNE, G.; CARROLL, J. J.; MCBRIDE, B. Rdf 1.1 concepts and abstract syntax. **W3C recommendation**, World Wide Web Consortium Cambridge, MA, USA, v. 25, n. 02, p. 1–22, 2014.

DESIMONI, F.; PO, L. Empirical evaluation of linked data visualization tools. **Future Generation Computer Systems**, Elsevier, v. 112, p. 258–282, 2020.

DIBOWSKI, H.; SCHMID, S. Using knowledge graphs to manage a data lake. *In:* **INFORMATIK 2020**. Karlsruhe: Gesellschaft für Informatik, Bonn, 2021.

DIBOWSKI, H.; SCHMID, S.; SVETASHOVA, Y.; HENSON, C.; TRAN, T. Using semantic technologies to manage a data lake: Data catalog, provenance and access control. *In:* **SSWS@ ISWC**. Athen: [*S. n.*], 2020. p. 65–80.

GALKIN, M.; AUER, S.; SCERRI, S. Enterprise knowledge graphs: a backbone of linked enterprise data. *In:* IEEE. **2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)**. Omaha, 2016. p. 497–502.

GALKIN, M.; AUER, S.; VIDAL, M.-E.; SCERRI, S. Enterprise knowledge graphs: A semantic approach for knowledge management in the next generation of enterprise information systems. *In:* **Proceedings of the 19th International Conference on Enterprise Information Systems - Volume 2: ICEIS**. [*S. l.*]: SciTePress, 2017. v. 2, p. 88–98.

GRAINGER, T.; ALJADDA, K.; KORAYEM, M.; SMITH, A. The semantic knowledge graph: A compact, auto-generated model for real-time traversal and ranking of any relationship within a domain. *In:* **2016 ieee international conference on data science and advanced analytics (DSAA)**. Montreal: IEEE, 2016. p. 420–429.

GRUBER, T. **What is an Ontology**. 1993. Disponível em: http://www-ksl.stanford.edu/kst/what-is-an-ontology.html. Acesso em: 05 fev. 2024.

HAASE, P.; HERZIG, D. M.; KOZLOV, A.; NIKOLOV, A.; TRAME, J. metaphactory: A platform for knowledge graph management. **Semantic Web**, IOS Press, v. 10, n. 6, p. 1109–1125, 2019.

IKEDA, R.; WIDOM, J. Data lineage: A survey. **Stanford University Publications**, Citeseer, v. 8090, n. 918, p. 1, 2009.

ISOTANI, S.; BITTENCOURT, I. I. **Dados abertos conectados: em busca da web do conhecimento**. São Paulo: Novatec Editora, 2015.

JIOMEKONG, A. A.; ASONG, F. M. D. Designing, implementing and deploying an enterprise knowledge graph from a to z. *In:* **Proceedings of the Federated Africa and Middle East Conference on Software Engineering**. New York, NY, USA: Association for Computing Machinery, 2022. (FAMECSE '22), p. 87–88. ISBN 9781450396639.

LAUFER, C. **Guia de Web Semântica**. São Paulo: Novatec Editora, 2015.

LI, H.; APPLEBY, G.; BRUMAR, C. D.; CHANG, R.; SUH, A. Knowledge graphs in practice: Characterizing their users, challenges, and visualization opportunities. **IEEE Transactions on Visualization and Computer Graphics**, v. 30, n. 1, p. 584–594, 2023.

LOPES, D. C. F. **Grafos de conhecimento: perspectivas e desafios para a organização e representação do conhecimento**. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal de São Carlos, São Carlos, 2020.

LOPES, G.; OLIVEIRA, M.; VIDAL, V. *et al.* Lais: Towards to a linked data framework to support decision-making on healthcare. *In:* **5th International Workshop on ADVANCEs in ICT INfrastructures and Services**. Evry: ADVANCE, 2017.

LOPES, G.; VIDAL, V.; OLIVEIRA, M. A framework for creation of linked data mashups: A case study on healthcare. *In:* **Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web**. New York, NY, USA: Association for Computing Machinery, 2016. (Webmedia '16), p. 327–330.

MOHAMMADI, M. (semi-) automatic construction of knowledge graph metadata. *In:* **The Semantic Web: ESWC 2022 Satellite Events Proceedings**. Hersonissos: Springer, 2022. p. 171–178.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. **Ontologias:** conceitos, usos, tipos, metodologias, ferramentas e linguagens. **Relatório Técnico–RT-INF-001/07**, 2007.

NADAL, S. **Metadata-driven data integration**. Tese (Doutorado em Ciência da Computação) – Universitat Politècnica de Catalunya (UPC), 2019.

NATH, R. P. D.; HOSE, K.; PEDERSEN, T. B.; ROMERO, O.; BHATTACHARJEE, A. Setlbi: An integrated platform for semantic business intelligence. *In:* **Companion Proceedings of the Web Conference 2020**. New York, NY, USA: Association for Computing Machinery, 2020. (WWW '20), p. 167–171. ISBN 9781450370240.

PAN, J. Z.; VETERE, G.; GOMEZ-PEREZ, J. M.; WU, H. **Exploiting linked data and knowledge graphs in large organisations**. Heidelberg: Springer, 2017.

PRUD'HOMMEAUX, E.; HARRIS, S.; SEABORNE, A. **SPARQL 1.1 Query Language**. **W3C Recommendation**, 2013. Disponível em: https://www.w3.org/TR/sparql11-query/. Acesso em: 05 out. 2023.

ROLIM, T.; AVILA, C.; ARRUDA, N.; SILVA, J.; MAIA, J.; OLIVEIRA, M.; ANDRADE, L.; VIDAL, V. Um enfoque incremental para construçao do grafo de conhecimento do sus. *In:* SBC. **Anais do XX Simposio Brasileiro de Computação Aplicada à Saúde**. Salvador, 2020. p. 72–83.

ROLIM, T. V.; AVILA, C. V. S.; JUNIOR, N. M. A.; COSTA, F. J.; MARIANO, R. G.; CALIXTO, T.; VIDAL, V. M. P. Kg-e: Um grafo de conhecimento semântico baseado na integraçao de dados de empresas e sancionados. *In:* SBC. **Anais do IX Workshop de Computação Aplicada em Governo Eletrônico**. [*S. l.*], 2021. p. 155–166.

SAWADOGO, P.; KIBATA, T.; DARMONT, J. Metadata management for textual documents in data lakes. *In:* INSTICC. **Proceedings of the 21st International Conference on Enterprise Information Systems - Volume 1: ICEIS**. [*S. l.*]: SciTePress, 2019. p. 72–83.

SELLAMI, S.; ZAROUR, N. E. Keyword-based faceted search interface for knowledge graph construction and exploration. **International Journal of Web Information Systems**, v. 18, n. 5/6, p. 453–486, 2022.

SEQUEDA, J.; LASSILA, O. **Designing and Building Enterprise Knowledge Graphs**. [*S. l.*]: Morgan & Claypool Publishers, 2021.

SEQUEDA, J. F.; BRIGGS, W. J.; MIRANKER, D. P.; HEIDEMAN, W. P. A pay-as-you-go methodology to design and build enterprise knowledge graphs from relational databases. *In:* **The Semantic Web–ISWC 2019: 18th International Semantic Web Conference**. Auckland, New Zealand: Springer, 2019. p. 526–545.

SOUZA, E. M. F. de; ROSSANEZ, A.; REIS, J. C. dos; TORRES, R. da S. Visualização interativa da evolução de grafos de conhecimento. *In:* SBC. **Anais do XXXVII Simpósio Brasileiro de Bancos de Dados**. [*S. l.*], 2022. p. 343–354.

SOYLU, A.; KHARLAMOV, E.; ZHELEZNYAKOV, D.; JIMENEZ-RUIZ, E.; GIESE, M.; SKJÆVELAND, M. G.; HOVLAND, D.; SCHLATTE, R.; BRANDT, S.; LIE, H. *et al.* Optiquevqs: A visual query system over ontologies for industry. **Semantic Web**, IOS Press, v. 9, n. 5, p. 627–660, 2018.

VIDAL, V. M.; 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:* SPRINGER. **International Conference on Advanced Information Systems Engineering**. [*S. l.*], 2015. p. 214–229.

W3C. **RDF 1.1 Concepts and Abstract Syntax**. 2014. Disponível em: https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225. Acesso em: 05 out. 2023.