UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE COMPUTAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM EM CIÊNCIA DA COMPUTAÇÃO

TULIO VIDAL ROLIM

SISIFO: UMA ABORDAGEM SEMÂNTICA PARA CONSTRUÇÃO DE ENTERPRISE KNOWLEDGE GRAPHS

FORTALEZA

2020

TULIO VIDAL ROLIM

SISIFO: UMA ABORDAGEM SEMÂNTICA PARA CONSTRUÇÃO DE ENTERPRISE KNOWLEDGE GRAPHS

Dissertação apresentada ao Curso de do Programa de Pós-Graduação em em Ciência da Computação do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Banco de Dados.

Orientadora: Profa. Dra. Vânia Maria Ponte Vidal.

FORTALEZA

2020

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)

R654s Rolim, Tulio Vidal.
Sisifo: Uma Abordagem Semântica para Construção de Enterprise Knowledge Graphs / Tulio Vidal Rolim. – 2020.
93 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Ciência da Computação, Fortaleza, 2020.
Orientação: Profa. Dra. Vania Maria Ponte Vidal.

1. Grafo de Conhecimento Corporativo. 2. Grafo de Conhecimento. 3. Integração Semântica. 4. Ontologia. I. Título.

CDD 005

TULIO VIDAL ROLIM

SISIFO: UMA ABORDAGEM SEMÂNTICA PARA CONSTRUÇÃO DE ENTERPRISE KNOWLEDGE GRAPHS

Dissertação apresentada ao Curso de do Programa de Pós-Graduação em em Ciência da Computação do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Ciência da Computação. Área de Concentração: Banco de Dados.

Aprovada em: 10/03/2020.

BANCA EXAMINADORA

Profa. Dra. Vânia Maria Ponte Vidal (Orientadora)
Universidade Federal do Ceará (UFC)

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

Prof. Dr. José Gilvan Rodrigues Maia
Universidade Federal do Ceará (UFC)

Dedico este trabalho primeiramente à Deus por ter me dado forças para suportar todas as dificuldades e seguir durante o desempenhar deste mestrado. Especialmente, dedico à meus pais Heldo de Sousa Rolim e Telma Vidal Silva Rolim por terem sido meus melhores amigos, por terem me amado, educado e sobretudo, ensinado bons princípios, possibilitando a construção do homem no qual sou hoje.

Obrigado é uma palavra muito singular para descrever a imensidão do meu amor por vocês, pois sei que nunca estivemos distantes, e como o sol e a lua, mesmo distantes um dia nos encontraremos. Tudo isso foi feito por vocês, e por vocês meu amor é eterno, *in memorian*.

Também dedico este trabalho à minha prima Jane Maria Silva Monte por ter sido minha outra mãe, estando ao meu lado e fazendo tudo que estava ao seu alcance para ver o meu bem, isso também é por você *in memorian*.

Um dia estaremos todos juntos.

AGRADECIMENTOS

Agradeço primeiramente a minha orientadora e espelho como profissional, a Profa. Dra. Vânia Maria Ponte Vidal por sua imensa contribuição neste trabalho bem como na minha formação profissional e pessoal. Me sinto honrado e agradecido por ter sido seu aluno, pois desde às aulas e orientações, a cada instante pude aprender mais com sua sabedoria, sendo a quem tenho imensa admiração, carinho e respeito.

Ao amigo e colega de mestrado Caio Viktor da Silva Ávila, por ter sido um grande companheiro e parceiro de pesquisa no qual pude compartilhar de boa parte das vivências do mestrado. Além do grande aluno e profissional exemplares, sua prestatividade e bom coração são raros, sendo alguém no qual tem minha sincera amizade e admiração, meu muito obrigado por tudo meu amigo.

Ao amigo Nickson Arrais por ter sido um dos primeiros que conheci ao chegar a UFC, tendo participado das primeiras aprendizagens e aventuras, desde conhecer a universidade, como a adequação à cidade de Fortaleza. Foi também com quem compartilhei preocupações em disciplinas e de trabalhos, meu muito obrigado. Ao amigo José Wellington Franco da Silva por ter sido a primeira pessoa a me acolher no laboratório ARiDA, parceiro no qual sempre me guiou ao bom caminho através de suas valiosas dicas. Também agradeço ao amigo Narciso Arruda Moura Júnior por ter estado sempre disponível para me ajudar com sua grande expertise e conhecimento nas tecnologias semânticas. Aos três, ainda agradeço pelas inúmeras caronas e apoio na ambientação com a cidade.

À minha família, especialmente a minha vó Francisca Vidal (A Dona Francisca) por ter sido uma outra mãe e cuidado de mim desde meu nascimento. As minhas primas Emily Victor, Jamily Melka e primo Marden Filho por terem se tornado irmãos, sempre me ajudando no que fosse preciso. Agradeço também a Marden Rômulo por ter me ajudado e estimulado a seguir nessa caminhada com seus conselhos.

Agradeço também com carinho à senhora Lourdes (Dona Lurdinha) por ter me acolhido de braços abertos em sua residência juntamente com o Messias e o Douglas durante boa parte do mestrado, sempre me tratando com cuidado e carinho.

Aos amigos Pedro Luis Saraiva Barbosa e Antonio Thalis Fonseca Lima por terem estado comigo durante o período, me tentando trazer um pouco mais de alegria e foco através de suas genuínas amizades.

Aos amigos do ARiDA e da UFC: Tiago Vinuto, Arlino Magalhães, Amanda Driely,

Matheus Mayron, Salomão Santos, Ernando Sousa, Gustavo e Francisco Carlos pelas parcerias e troca de conhecimentos.

Também a amiga de Mestrado: Vitória Regina, conterrânea no qual vivenciou e compartilhou parte da experiência e das dificuldades de um cidadão icoense na capital.

Aos profissionais da Secretaria Jonatas e Glaucia que sempre prestaram seu trabalho com excelência e humanidade, meus sinceros agradecimentos. Ao seu Reginaldo, pelo seu grande entusiasmo e simpatia.

A todos os professores por me proporcionarem o conhecimento não apenas racional, mas também na manifestação do caráter e afetividade da educação no processo de formação profissional, por tanto que se dedicaram a mim, não somente por terem me ensinado, mas por terem me feito aprender.

E a Fundação Cearense de Apoio ao Desenvolvimento (FUNCAP), na pessoa do Presidente Tarcísio Haroldo Cavalcante Pequeno pelo financiamento da pesquisa de mestrado via bolsa de estudos.

A todos os demais familiares, primos, tias, madrinha, padrinho, assim como aos amigos pessoais, meu muito obrigado a todos vocês que colaboraram para este momento. Vocês foram importantes para isso.

"A sabedoria de um homem não está em não errar, chorar, se angustiar e se fragilizar, mas em usar seu sofrimento como alicerce de sua maturidade."

(Augusto Cury)

RESUMO

Como uma tecnologia para apoiar o gerenciamento do conhecimento *Knowledge Graph (KG)s* podem ser utilizados como forma de expressar o conhecimento utilizando-se especialmente do significado dos dados e representando formalmente os relacionamentos entre os conceitos de forma independente. A aplicação de um modelo de dados utilizando-se de uma camada semântica sobre os dados integrados em um KG no âmbito corporativo produz um Grafo de Conhecimento Corporativo - *Enterprise Knowledge Graph* (EKG). Em um EKG, o conhecimento é organizado em uma rede de nós e links, contendo objetos de negócios e tópicos intimamente vinculados, classificados, semanticamente enriquecidos e conectados a dados e documentos existentes. Entretanto, construir e manter EKGs que utilizam da semântica como forma de enriquecimento no tratamento dos dados torna-se uma tarefa complexa, muito em razão da necessidade de conhecimento e esforço investido nas atividades e manuseio de ferramentas utilizadas no processo de integração semântica. Este trabalho propõe uma abordagem semântica para construção, manutenção e reúso de EKG de forma semi-automática. Para tanto, as principais contribuições deste trabalho são: i) Uma Abordagem Semântica para construção de EKGs; ii) Organizar e estruturar o conhecimento envolvido na Construção de um EKG através das etapas de: questões de competência, modelagem da ontologia de domínio, especificação (modelagem de cada ontologia exportada, mapeamentos e definições de metadados), publicação das visões e validação do EKG; iii) Apoiar, por meio de ferramentas apropriadas e orientações todo o processo de construção semântica de EKGs de modo à fornecer a possibilidade de extensão e reúso do conhecimento; iv) Uma Representação Ontológica para especificar e descrever EKGs (*EKG-Ontology*). Como validação, foram utilizados experimentos através de estudos de caso com ênfase na construção de EKGs utilizando os passos da abordagem proposta guiando-se em uma versão operacional da *EKG-Ontology* para instanciação e representação semântica do EKG no domínios fiscal (empresas, sócios e contribuintes). Após a realização dos experimentos, SISIFO apresentou bons resultados no que tange a validação por meio de estudos de caso. Assim, SISIFO demonstrou fornecer uma maneira de construir, especificar formalmente e gerenciar EKGs por meio do vocabulário descrito em *EKG-Ontology*, possibilitando que seu uso seja estendido para demonstrar outras aplicações e estudos.

Palavras-chave: grafo de conhecimento corporativo; grafo de conhecimento; integração semântica; ontologia.

ABSTRACT

As a technology to support knowledge management, Knowledge Graphs (KGs) can be used as a way to express knowledge using especially the meaning of the data and formally representing the relationships between the concepts of independently. The application of a data model using a semantic layer on the data integrated in a KG in the corporate scope produces a Enterprise Knowledge Graph (EKG). In a EKG, knowledge is organized into a network of nodes and links, containing business objects and topics closely linked, classified, semantically enriched and connected to existing data and documents. However, building and maintaining EKGs that use semantics as a way of enriching data processing becomes a complex task, largely due to the need for knowledge and effort invested in the activities and handling of tools used in the process of semantic integration. This work proposes a semantic approach for building, maintaining and reusing EKG in a semi-automatic way. For this, the main contributions of this work are: i) A Semantic Approach for building EKGs; ii) Organize and structure the knowledge involved in the Construction of an EKG through the steps of: competence issues, modeling of the domain ontology, specification (modeling of each exported ontology, mapping and metadata definitions), publication of EKG views and validation ; iii) Support, by means of appropriate tools and guidelines, the whole process of semantic construction of EKGs in order to provide the possibility of extending and reusing knowledge; iv) An Ontological Representation to specify and describe EKGs (EKG-Ontology). As validation, experiments were used through case studies with an emphasis on the construction of EKG s using the steps of the proposed approach guided by an operational version of EKG-Ontology for instantiation and semantic representation of EKG in the tax tax domain (organizations, partners and taxpayers). After carrying out the experiments, SISIFO presented good results in terms of validation through case studies. Thus, SISIFO demonstrated to provide a way to build, formally specify and manage EKGs through the vocabulary described in EKG-Ontology, allowing its use to be extended to demonstrate other applications and studies.

Keywords: enterprise knowledge graph; knowledge graph; semantic integration; ontology.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE CONSULTAS

LISTA DE ABREVIATURAS E SIGLAS

EKG	*Enterprise Knowledge Graph*
EKGO	EKG Ontology
EO	Engenharia de Ontologias
HTTP	Hypertext Transfer Protocol
IRI	Internationalized Resource Identifier
IS	Integração Semântica
KG	*Knowledge Graph*
MS	Mediador Semântico
OBDA	Ontology Based Data Access
OWL	*Ontology Web Language*
QCs	Questões de Competência
RDF	*Resource Description Framework*
RDFS	*Resource Description Framework Schema*
SPARQL	*SPARQL Protocol and RDF Query Language*
UFO	Unified Foundational Ontology
URI	Uniform Resource Identifier
VKG	Virtual Knowledge Graph

LISTA DE SÍMBOLOS

λ	Lambda
$\cup$	União
$\in$	Pertence a

SUMÁRIO

1 INTRODUÇÃO

Cada vez mais empresas e organizações vem buscando inserir a inovação por meio de mudanças no que tange a seu modelo de funcionamento e uso de tecnologias para gerenciamento dos dados. Nesse sentido, o conhecimento corporativo demanda de métodos, técnicas e ferramentas para estruturar, representar e analisar esse conhecimento de modo a fornecê-lo para todo o ecossistema da empresa (GALKIN *et al.*, 2016).

Como uma tecnologia para apoiar o gerenciamento do conhecimento, *Knowledge Graphs* (KGs) podem ser utilizados como forma de expressar e representar o conhecimento corporativo. KGs baseiam-se em regras que aplicam anotações semânticas em dados brutos ou semiestruturados, possibilitando que a máquina possa utilizar esse conhecimento para realizar inferências e responder perguntas significantes a um determinado domínio.

Em outras palavras, um KG pode ser compreendido como um conjunto de triplas, onde cada tripla representa intuitivamente uma 'afirmações', sendo que construído corretamente, essas afirmações de podem ser vistas como 'fatos'. KGs podem utilizar dos recursos da Web Semântica para representação, análise e integração do conhecimento, utilizando-se especialmente do significado dos dados e representando formalmente os relacionamentos entre os conceitos de forma independente (BONATTI *et al.*, 2019).

A combinação de Ontologias e *Linked Data* proporcionou novas oportunidades para enfrentar os desafios no desenvolvimento de um KG. O sucesso da iniciativa *Linked Data* se deve principalmente à adoção de padrões da Web conhecidos, como padrões de infraestrutura da Web (*Uniform Resource Identifier (URI)s* e *GlsHTTP*), padrões da Web Semântica (*Resource Description Framework (RDF)* e *Resource Description Framework Schema (RDFS)*) e vocabulários, que facilitam a implantação de fontes de dados ligadas.

A aplicação do modelo de dados através de uma camada semântica sobre os dados integrados em um KG produz um Grafo de Conhecimento Corporativo - *Enterprise Knowledge Graph* (EKG) (FORBES, 2018), o qual representa o domínio de conhecimento de uma organização. Segundo (JETSCHNI; MEISTER, 2017), EKGs podem prover contribuições como:

- Fornecer um contexto semântico ao conhecimento da organização bem como das atividades e fluxos relativos ao negócio;
- Explicitar o conhecimento de forma compartilhada e formalizada;
- Permitir a descoberta de relações entre os conceitos relativos à organização;
- Integrar dados advindos de fontes, formatos, vocabulários e *schemas* heterogêneos;

Em um EKG, o conhecimento é organizado em uma rede de nós e links, com isso, pessoas e máquinas podem se beneficiar de uma rede semântica de fatos sobre coisas que cresce dinamicamente, possibilitando seu uso para integração de dados, descoberta de conhecimento e análises detalhadas.

Acredita-se que um EKG é fundamental para alavancar totalmente o poder por trás das tecnologias da web semântica e fornecer o contexto, o significado e os relacionamentos que podem ser aproveitados através de um novo tipo de processo de descoberta de dados (GALKIN *et al.*, 2017).

Entretanto, construir e manter EKGs que utilizam da semântica como forma de enriquecimento no tratamento dos dados torna-se uma tarefa complexa, muito em razão da necessidade de conhecimento acerca das atividades e manuseio de ferramentas no processo de Integração Semântica (IS).

Assim, a construção de um EKG suportado pelas etapas do processo de integração semântica é composto pelos seguintes problemas:

1. Identificação de *links* entre recursos em diferentes fontes de dados;
2. Heterogeneidade das fontes de dados e vocabulários;
3. Combinação e fusão de múltiplas representações do mesmo objeto do mundo real em uma única representação e resolução inconsistências e conflitos para melhorar a qualidade dos dados;
4. Suportar cenários para construção de EKGs sob os enfoques materializado e virtual de forma híbrida;

Este trabalho propõe SISIFO, uma abordagem semântica para construção de EKGs baseada na especificação e publicação de visões exportadas e de linksets. Em conjunto, é apresentado uma representação ontológica para descrever e apoiar o EKG produzido.

Para tanto, as principais contribuições do trabalho são:

- Uma Abordagem Semântica para construção de EKGs;
- Organizar e estruturar o conhecimento envolvido na Construção de um EKG através das etapas de: questões de competência, modelagem da ontologia de domínio, especificação (modelagem de cada ontologia exportada, mapeamentos e definições de metadados), publicação das visões e validação do EKG;
- Apoiar, por meio de ferramentas apropriadas e orientações todo o processo de construção semântica de EKGs de modo à fornecer a possibilidade de extensão e reúso do

conhecimento;

– Uma Ontologia para especificação formal de EKGs;

Logo, a proposta de construção de EKGs neste trabalho tende a viabilizar o acesso aos dados através de uma visão semanticamente integrada, utilizando-se de uma padronização por meio do RDF e de uma ontologia com regras para facilitar sua validação, provendo aos usuários finais uma maior qualidade dos dados.

Ainda, a capacidade de registrar a integração de dados como elementos compreensíveis pela máquina, faz com que o EKG gerado forneça total transparência dos dados e dos artefatos que fazem parte da integração semântica, possibilitando também a incorporação de regras através de metadados que armazenem diferentes valores, recursos e definições. Para tanto, as principais contribuições do trabalho são:

– Uma Abordagem Semântica para construção de EKGs;
– Organizar e estruturar o conhecimento envolvido na Construção de um EKG através das etapas de: questões de competência, modelagem da ontologia de domínio, especificação (modelagem de cada ontologia exportada, mapeamentos e definições de metadados), publicação das visões e validação do EKG;
– Apoiar, por meio de ferramentas apropriadas e orientações todo o processo de construção semântica de EKGs de modo à fornecer a possibilidade de extensão e reúso do conhecimento;
– Uma Ontologia para especificação formal de EKGs;

Assim, a construção de um EKG suportado pelas etapas do processo de integração semântica é composto pelos seguintes problemas:

1. Identificação de *links* entre recursos em diferentes fontes de dados;
2. Heterogeneidade das fontes de dados e vocabulários;
3. Combinação e fusão de múltiplas representações do mesmo objeto do mundo real em uma única representação e resolução inconsistências e conflitos para melhorar a qualidade dos dados;
4. Suportar cenários para construção de EKGs sob os enfoques materializado e virtual de forma híbrida;

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta fundamentos de Web Semântica, *Linked Data*, *Knowledge Graphs* e *Enterprise Knowledge Graphs* importantes para compreensão do restante deste trabalho.

2.1 Web Semântica, *Linked Data* e suas Tecnologias

A Web Semântica vem caracterizando-se como uma das áreas emergentes de maior relevância em pesquisas realizadas da área de tecnologia. Por conceito, a Web Semântica pode ser vista como uma extensão da *World Wide Web* 2.0, promovendo um conjunto de elementos semânticos bem definidos que são processados e compreendidos cooperativamente por humanos e computadores (BERNERS-LEE *et al.*, 2001).

Como principal objetivo, a Web Semântica visa fornecer semântica para a web atual, possibilitando ao computador uma compreensão dos termos e conceitos presentes nos dados e em páginas web, possibilitando assim, que a máquina possa realizar tarefas que comumente são feitas de forma manual. Logo, de forma geral, a Web Semântica pode ser vista como um conjunto de tecnologias e padrões que permite ao computador interpretar a semântica das informações na web (YU, 2011).

A Web Semântica organiza-se em camadas, distribuição conhecida como *cake-layer* (BERNERS-LEE *et al.*, 2001), onde uma dessas camadas é representada pelo RDF RDF (W3C, 2004) é um arcabouço para representar informações na Web. RDF permite fazer afirmações sobre recursos. Recursos são quaisquer coisas, tanto concretas quanto abstratas.

O RDF é representado semanticamente como uma tupla-tripla (<sujeito> <predicado> <objeto>). Como codificação, o RDF utilizada URIs, recurso que garante que todos os conceitos sejam representados através de um identificador único, sendo essa uma representação compartilhada em toda a web (SHADBOLT *et al.*, 2006).

Para expressar a semântica, a tecnologia usada pela Web Semântica consiste na utilização de ontologias. Uma ontologia é basicamente compreendida como uma especificação formal e explícita de uma conceituação compartilhada (STUDER *et al.*, 1998), sendo constituída por uma taxonomia juntamente com um conjunto de regras de inferência. A taxonomia define as classes e os relacionamentos entre as instâncias, já as regras de inferência fornecem a capacidade de inferir novos fatos que à priori não eram possíveis através de axiomas.

As ontologias são construídas através de uma linguagem bem definida e projetada de

modo a possibilitar a representação dos conceitos (classes) e relações (propriedades) permitindo compatibilidade com a Web. A linguagem utilizada para criar e representar ontologias é baseada na lógica descritiva, a *Ontology Web Language* (OWL). OWL é a linguagem padrão mais utilizada por engenheiros e arquitetos de ontologia. O OWL divide-se em (*OWL Lite*, *OWL DL* e *OWL Full*), tendo como ideia central a representação eficiente de uma ontologia através do seu conjunto de regras lógicas (SHADBOLT *et al.*, 2006).

O uso das tecnologias da web semântica traz muitos benefícios para enfrentar os desafios no desenvolvimento de aplicações onde se existe a necessidade de integrar semanticamente fontes de dados heterogêneas. Os principais benefícios são os seguintes:

- Interoperabilidade: O paradigma de Dados Interligados é baseado no modelo RDF que usa URIs para identificar recursos sem ambiguidade;
- Semântica e Inferência: Os padrões de dados interligados, como RDF, RDFS e OWL definem relacionamentos semânticos de alto nível entre vários tipos de recursos. A semântica do OWL é baseada em Lógicas de Descrição e, portanto podem ser usadas para inferência e raciocínio automático através de módulos de software prontos para uso, os chamados "raciocinadores";
- Reuso do Conhecimento: A interconexão de uma variedade de fontes de dados disponíveis publicamente por meio dos padrões *Linked Data* pode facilitar significativamente sua reutilização, maior exploração e possível extensão.

Além de ontologias, a Web Semântica é constituída também pela Web de Dados, uma camada adicional relacionada com a Web clássica de documentos. Como características, a web de dados é tida como genérica, visto sua capacidade de conter qualquer tipo de dados. Possui uma enormidade de vocabulários e ontologias no qual representam os dados através de links RDF, além de possibilitar que qualquer pessoa possa publicar seus dados. Como base para publicação desses dados, surgiu o *Linked Data*.

O *Linked Data* baseia-se, principalmente em duas tecnologias: URIs) - responsáveis por endereçar entidades do mundo real na web através de um endereço único e o Hypertext Transfer Protocol (HTTP) - protocolo padrão de envio e recepção de informações na web (BIZER *et al.*, 2011).

Como consequência, sua adoção baseada nas melhores práticas para ligar dados permitiu a extensão da web de dados em um âmbito global, conectando dados dos mais variados domínios, como: saúde, educação, entretenimento e outros. O *Linked Data* possui um conjunto

de práticas para se publicar e ligar dados estruturados na Web propostas por (BERNERS-LEE, 2006) conforme exibido abaixo:

1. Atribua nomes de URIs às "coisas"em conjuntos de dados em formato RDF;
2. Defina URIs com base em HTTP que sejam compreensíveis às pessoas;
3. Forneça meios de acessar a informação utilizando os padrões (RDF, SPARQL);
4. Defina links RDF para outras fontes de dados na Web, para que sejam acessíveis na Web de Dados e para descoberta de novos fatos.

A tecnologia utilizada para realizar consultas sob a Web Semântica e *Linked Data* é o *SPARQL Protocol and RDF Query Language* (SPARQL). SPARQL baseia-se em padrões de grafo e correspondências de sub-grafos. Em uma consulta SPARQL, o bloco de construção base, a partir do qual os padrões de consulta SPARQL são construídos é conhecido como *basic graph pattern (BGP)*, um conjunto de *triple patterns* expressos como (W3C, 2006):

- *(RDF* $-$ *T*$\cup$*V) x (I*$\cup$*V) x (RDF* $-$ *T*$\cup$*V)*, onde:
  - $RDF - T$: conjunto de termos RDF;
  - I: conjunto de todas as IRIs (*Internationalized Resource Identifier (IRI)s*);
  - V: conjunto de variáveis em uma consulta;

O padrão de *matching* em grafos inclui características como elementos opcionais , filtragem, união de padrões, agrupamento assim como a possibilidade de seleção do grafo ou fonte de *matching*. O resultado de uma consulta SPARQL é um conjunto de variáveis especificados via SELECT, podendo também, ser um grafo via CONSTRUCT. SPARQL possui sintaxe similar ao SQL, possibilitando a utilização de operadores clássicos como SELECT / Projeção, DISTINCT, LIMIT, OFFSET e ORDER BY (PÉREZ *et al.*, 2009).

A utilização das tecnologias da Web Semântica e *Linked Data* visam realizar uma Integração Semântica dos Dados (IS), com ênfase na geração de valor por meio de novas relações e fatos que a priori não eram possíveis em fontes de dados não integradas.

Uma IS é o processo que utiliza uma representação conceitual dos dados e suas relações para eliminar heterogeneidades e discrepâncias nos dados. Formalmente, a integração semântica pode ser definida formalmente com base no *framework* apresentado em (VIDAL *et al.*, 2015), onde:

- O_D representa uma ontologia de domínio responsável por estabelecer e fornecer um vocabulário comum compartilhado entre as visões locais das fontes de dados;
- V representa um conjunto de visões locais $V_1, \ldots, V_n$ que especificam as fontes de dados

$S_1, \ldots, S_n$ utilizando os termos em O_D. Uma especificação da visão local V_i é uma tupla (O_{Vi}, M_{Vi}), onde:

  - O_{Vi} é a ontologia da visão local, vocabulário conjunto de O_D com ocorrências em M_{Vi}.
  - M_{Vi} é um conjunto de mapeamentos que relaciona os termos do vocabulário de O_D com os termos de *Si*;

- *L* representa um conjunto de visões de especificações de linkset $L_1, \ldots, L_m$ que especifica links *owl:sameAs* entre recursos de diferentes visões locais. Esses links são usados para relacionar instâncias de uma mesma entidade no mundo real, sendo criados com base em regras de *linkage* definidos em uma especificação de visões de ligação (CASANOVA *et al.*, 2014).

2.2 Grafos de Conhecimento (*Knowledge Graphs*)

Nos últimos anos, os Grafos de Conhecimento (*Knowledge Graphs (KG*s)) vêm se tornando um instrumento eficaz para representação do conhecimento, tendo um papel importante na organização, integração e utilização do conhecimento.

O termo *Knowledge Graph* no qual compreende-se foi apresentado pelo *Google* em 2012 (SINGHAL, 2012) como um aprimoramento semântico de sua função de pesquisa, não fazendo buscas por palavras chaves exatas, mas possibilitando a realização consultas mais complexas através de "coisas"e objetos do mundo real.

Além do Google, outras interpretações e abordagens foram estabelecidas por grandes empresas, com suas próprias características, padrão de arquitetura e tecnologias utilizadas, como o Grafo de Conhecimento Satori da Microsoft[1] e o Grafo de Entidades do Facebook [2]. Essa variedade veio por tornar difícil uma definição consensual do que seria um KG.

Nessa premissa, algumas definições podem ser encontradas, a primeira é mais simples e define que KGs são como grandes redes de entidades e seus tipos semânticos, propriedades e relações entre essas entidades (KROETSCH; WEIKUM, 2016). Já Färber *et al.* (2018) conceitua um KG como um grafo RDF composto por um conjunto de triplas RDF, onde cada tripla RDF é composta por (*s*, *p*, *o*), sendo: um

- Um sujeito $s \in \bigcup \cup B$;

[1] https://blogs.bing.com/search/2013/03/21/understand-your-world-with-bing/
[2] https://tinyurl.com/1nt2n4ez

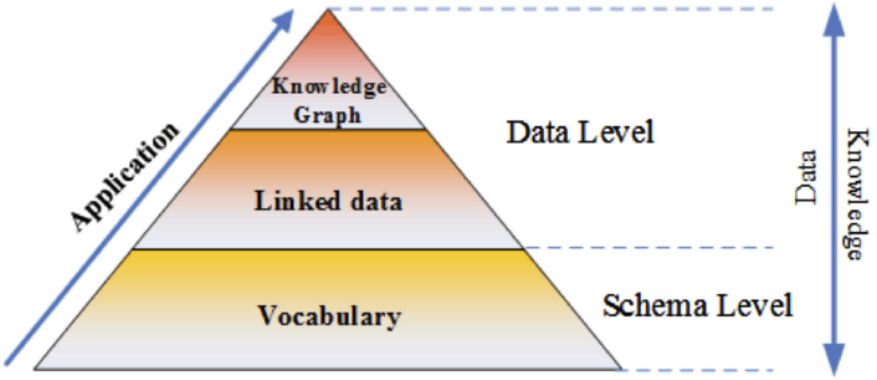
– Um predicado $p \in \bigcup$;
– Um objeto $\bigcup \cup B \cup L$ - tendo uma URI $\mu \in \bigcup$, um nó em branco $b \in B$, ou um literal $l \in L$;

Um KG é constituído de uma camada de metadados (schema) e uma camada de dados. A camada de *schema* determina a estrutura para representação semântica do conhecimento embutido na camada de dados em alto nível. Já a camada de dados contém as instâncias, seus tipos, definições e relações em baixo nível (QIAO *et al.*, 2017).

Através dessas duas principais camadas, KGs fornecem meios para análise de instâncias de entidades e suas relações, a fim de gerar novos conhecimentos, proporcionando uma melhoria às aplicações com base em sua capacidade de descoberta e inferência de novos fatos.

KGs podem ser vistos também como uma evolução do *Linked Data*, percepção justificada a partir da compreensão de que um KG passa pelas mesmas etapas que o *Linked Data* propõe para uma integração semântica de dados além de utilizar de vocabulários para ambientes de larga escala como em sistemas *big data*. Assim, um KG também é um conceito mais complexo e com uma maior capacidade de raciocínio e inferência que o de uma ontologia ou base de conhecimento pura (JIA, 2020). Jia (2020) sugere que a representação do conhecimento é evoluída a partir de três níveis, conforme exibido na Figura 1.

Figura 1 – Visão de Dados e Níveis de Conhecimento.



Fonte: (JIA, 2020).

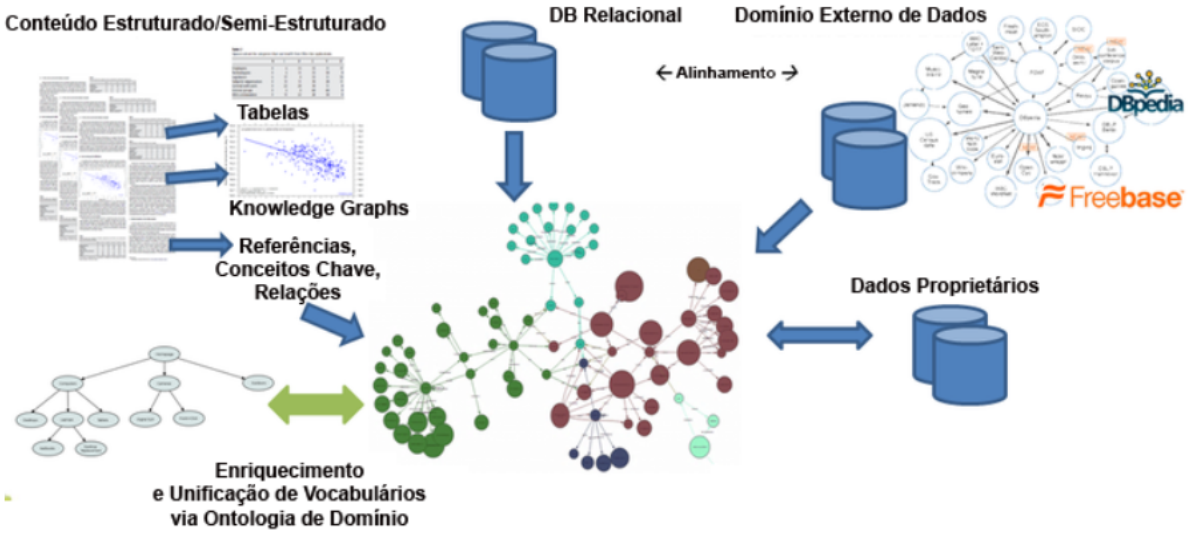
2.3 *Enterprise Knowledge Graphs*

Enterprise Knowledge Graphs (EKGs) vêm sendo usados como um mecanismo para consolidar semanticamente e integrar um grande número de fontes de dados heterogêneas em

um espaço de dados disponível (GOMEZ-PEREZ *et al.*, 2017).

EKGs surgiram com uma proposta de aplicação mais poderosa que KGs para gerenciamento de dados no domínio corporativo, tendo como principal diferencial a intenção de que todos as fontes de dados independe de tipo, formato ou *schema* estejam representados e integrados. A Figura 2 apresenta uma visão de um EKG e sua composição por múltiplas fontes heterogêneas.

Figura 2 – EKG e suas múltiplas fontes heterogêneas.



Fonte: Adaptado de (ICSC, 2014)

Um EKG consiste em uma estrutura formal e semântica representada em grafo cujo objetivo direciona-se em relacionar e enriquecer dados (DUAN; XIAO, 2019), sendo composto por fontes de dados possivelmente heterogêneas, em conjunto com uma ontologia de domínio que fornece um vocabulário que provê semântica aos conceitos de uma organização (GALKIN *et al.*, 2017).

EKGs armazenam e representam objetos de negócios semanticamente ligados, classificados, e conectados a dados e documentos existentes (POOLPARTY, 2019). Em uma compreensão mais clara, segundo (FRAUNHOFER, 2020) um EKG é especificamente composto por:

- **dados de instância** advindos de fontes de dados abertas (por exemplo, DBpedia, WikiData) e privadas (por exemplo, dados da cadeia de suprimentos, modelos de produtos);
- **dados derivados** e agregados;
- **dados de *schema*** (vocabulários, ontologias, taxonomias), categorizando entidades;
- **metadados** por exemplo, proveniência, versão, documentação, licenciamento);

- **links** entre dados internos e externos;
- **mapeamentos** para dados armazenados em outros sistemas e bancos de dados.

O principal objetivo de um EKG é fornecer uma camada de dados unificada, flexível e amigável ao ser humano, estando semanticamente conectada aos dados da empresa, para que aplicações possam ter acesso integrado às fontes de dados por meio de uma camada semântica, oportunizando a descoberta de *insights* valiosos acerca do negócio de uma organização e reunindo tecnologias semânticas e infraestruturas de dados corporativos.

Dessa maneira, um EKG pode oferecer suporte à consultas *ad-hoc* não planejadas além de possibilitar exploração de dados por usuários e agentes de negócios sem exigir um pré-processamento complicado e custoso.

Para tanto, a adoção de EKGs tornam-se uma grande tecnologia para integração de dados, permitindo a capacidade de conectar e fornecer sentido às máquinas através da semântica nos dados, sendo uma tecnologia promissora nos próximos anos.

3 TRABALHOS RELACIONADOS

A metodologia empregada baseou-se em uma Revisão Sistemática (RS). A RS tem como propósito realizar uma pesquisa com profundidade e não adota como objetivo fazer uma pesquisa em abrangência na literatura, tendo como resultados fornecer uma visão específica de determinada área pesquisada (KITCHENHAM; CHARTERS, 2007).

A RS é organizada de acordo com o Protocolo de Revisão Sistemática (PRS). Neste protocolo, são apresentados a descrição da condução do trabalho, processo avaliativo e de extração dos resultados. O PRS baseia-se na utilização de Questões de Pesquisa. Por conseguinte, são utilizadas estratégias para a realização de buscas dos artigos primários, de modo a localizar a maior quantidade de artigos. O PRS deste estudo foi organizado em: Questões de Pesquisa, Estratégia de Busca e Resultados.

3.1 Questões de Pesquisa

As questões de pesquisa são identificadas respectivamente por QP1, QP2, QP3 e QP4 e foram definidas da seguinte maneira:

- **(QP1)** Os trabalhos abrangem todos os passos do processo de Integração Semântica?;
- **(QP2)** É proposto algum modelo de construção e.g *pay-as-you-go*?;
- **(QP3)** Nos estudos, é apresentada alguma proposta ontológica para viabilizar e facilitar a construção de EKGs?;
- **(QP4)** Lida com a construção de EKGs formados por visões tanto virtuais como materializadas?;
- **(QP5)** Uma Camada ou Visão Semântica é adotada na construção dos *EKG*s?

3.2 Estratégia de Busca

A estratégia de busca a ser usada é um ponto chave para o sucesso ou fracasso de uma RS. Ela inclui a definição de: *(i)* métodos de busca, i.e., como as buscas serão realizadas; *(ii)* fontes de pesquisa, ou seja, os locais onde os estudos serão procurados e *string* de busca a ser usada; *(iii)* critérios de inclusão e exclusão; e *(iv)* análise dos trabalhos.

O método utilizado foi a busca automática dos artigos primários, de modo que a maior quantidade de artigos primários fosse encontrada. Após encontrar uma gama de estudos,

foi realizado um método para seleção dos estudos encontrados, levando-se em consideração os critérios de inclusão e exclusão. Também são definidos critérios para avaliar a qualidade dos estudos selecionados a partir de cada fonte de pesquisa.

3.2.1 Fontes de Pesquisa e String de Busca

As Fontes de Busca definidas foram: IEEE e Springer. Durante a definição da string de busca, o foco foi direcionado na identificação de termos relacionados a construção de EKGs nos estudos primários alvos da RS. Uma boa prática para formular a string de busca consiste em agrupar termos relativos a um mesmo aspecto, que podem ser considerados sinônimos, concatenando-os com o conectivo OU (OR em inglês). Posteriormente, cada grupo de termos foi concatenado com os demais por meio de conectivos E (AND em inglês). Na Figura 3 é apresentado o processo de definição da *string* de busca.

Figura 3 – Processo de Definição da String de Busca.

A *string* de busca inicial foi construída com base no PICO: *(Population (P) – Intervention(I) – Comparator(C - Opcional) – Outcome(O))*, sendo definida como P = Framework, I = Building, C = {Ø} e O = *Enterprise Knowledge Graphs OR Semantic Enterprise Knowledge Graphs*.

Como objetivo, as buscas preliminares norteiam-se nas palavras-chaves dos artigos de modo a descobrir termos que não fazem parte da *string* de busca inicial, esse processo foi repetido diversas vezes até se esgotarem os termos relevantes para inclusão na *string*. Após as buscas preliminares, foram adicionados os termos: *P = Tool OR Approach OR Methodology e I = Create*. Ficando: *Framework OR Software OR Environment OR Tool OR Approach, I*

= Building OR Create, C Ø e O = Enterprise Knowledge Graphs OR Semantic Enterprise Knowledge Graphs.

Contudo, após a busca experimental o X (quantidade) inicial de trabalhos encontrados ter se apresentado muito grande para análise, optou-se então pela definição de uma *string* que trouxesse uma quantidade Y menor de estudos que pudessem ser analisados. Assim, a *string* oficial adotada foi: *((Framework OR Environment) AND (Building) AND "Enterprise Knowledge Graph")*.

Ressalta-se que, cada fonte de busca possui características, modo de funcionamento e limitações próprias. Por exemplo, deve-se considerar se a fonte aceita termos no plural, caracteres especiais e, ainda, se permite o uso de partes do texto como resumo, palavras-chave, título ou se permite apenas a busca por texto completo etc (KITCHENHAM *et al.*, 2010). Logo, fez-se necessário a adaptação da *string* de busca para que seja executada adequadamente sobre cada fonte de busca individualmente.

3.2.2 Critérios de Inclusão e Exclusão

Os critérios de inclusão e exclusão foram definidos com base no pressuposto de que todos os trabalhos estavam disponíveis para acesso nas bases. Assim o critério de inclusão estabelecido foi: **CI1)** Estudos que abordam o processo de construção de EKGs; **CI2)** Trabalhos que contenham pelo menos dois dos termos da *string* de busca em seu título; já os de exclusão foram: **CE1)** Trabalhos não publicados nos últimos 5 anos (2015 a 2020); **CE2)** Teses, dissertações, *workshops*, slides de apresentação, dentre outros que não artigos completos ou capítulos de livro.

3.3 Análise dos Trabalhos

A realização da pesquisa ocorreu no dia 3 de outubro de 2020. A busca na base Springer retornou 18 trabalhos relacionados. Já o *IEEE* retornou 75 resultados. A junção dos resultados do Springer e do IEEE geraram uma quantidade de 93 trabalhos para serem analisados no total. Destes, 3 foram removidos por serem trabalhos duplicados, resultando em 90 artigos finais relacionados ao tema. Para cada um destes foram lidos o *abstract* e as palavras-chaves, onde em caso de dúvida, o trabalho era analisado de forma completa. O processo de análise e condução foi realizado com auxílio do *software StArt* [1].

[1] (http://lapes.dc.ufscar.br/tools/start_tool)

Após leitura e aplicação dos critérios de exclusão, e análise, 3 trabalhos foram considerados como adequados e de interesse a este estudo.

Gomez-Perez *et al.* (2017) propõe um *framework* conceitual formal para a construção de EKGs utilizando os conceitos e tecnologias do processo de integração semântica. Além disso, no trabalho, é apresentado um ciclo de vida para a construção e manutenção do EKG, abrangendo as etapas de: Especificação, Modelagem, Transformação (Levantamento de Dados), Publicação, Limpeza e Resolução de Inconsistências (Curadoria de Dados).

Por sua vez, Sequeda e Miranker (2017) propõe a *pay-as-you-go* abordagem para a construção de EKGs com base no uso do OBDA (Onlogy-Based Data Access, acesso de dados baseado em ontologia) (POGGI *et al.*, 2008) e usando questões de competência para construção e validação do EKG sob um *schema relacional*.

Fensel *et al.* (2020) propõe uma metodologia para a construção de um EKG, fornece recomendações de ferramentas para assistência manual, semiautomática e automática para apoiar as etapas do processo.

Para analisar os trabalhos, as Questões de Pesquisa (QPs) foram utilizadas como critérios para se identificar diferenciais entre os estudos na literatura e est, visto que apesar de apresentarem propostas na construção de EKGs, os trabalhos possuem limitações. A Tabela 1 apresenta uma análise dos estudos relacionados com base nas questões descritas na Seção 3.1.

Tabela 1 – Comparação de trabalhos relacionados

	(QP1)	(QP2)	(QP3)	(QP4)	(QP5)
Gomez-Perez *et al.* (2017)	X	X			X
Sequeda e Miranker (2017)		X		X	X
Fensel *et al.* (2020)	X	X			X
Este trabalho	X	X	X	X	X

4 CONSTRUÇÃO DE EKGS

Neste capítulo, a Seção 5.1 expõe motivações acerca do uso de EKGs, e por conseguinte na Seção 4.1 são apresentadas a visão da arquitetura de EKGs proposta detalhando-se seus componentes.

Por composição, um EKG deve conter um *schema*, metadados e dados de instâncias relevantes ao domínio de atuação de uma empresa. Esses dados são distribuídos e percebidos como dados relativos a todo o negócio da empresa, como: produtos, dados de clientes, estratégias além de informações relevantes externas, como dados sobre a região ou localidade no qual atua, dados sobre o ramo de negócio bem como de seus concorrentes.

Uma função importante dos EKGs é definir e manter links entre as informações da empresa e as providas em fontes de dados externas. Esse processo demanda de extração, ampliação e integração sob um mesmo vocabulário ou *schema* único, contudo, essa não é uma tarefa trivial, sendo importante o seguimento de um conjunto de passos bem orientados para construção do EKG.

O processo de construção do EKG é realizado com base na especificação e publicação das visões exportadas e de *linksets*, devendo este ser povoado com os dados obtidos a partir da publicação das visões exportadas. Existem duas formas de importar os dados das fontes de dados, o enfoque virtual e o materializado.

No enfoque materializado, os dados da visão exportada são realmente importados da fonte de dados e armazenados em um fragmento do EKG. Esse enfoque é mais adequado para extrair dados de fontes de dados não estruturadas e também de fontes que não são atualizadas com frequência. A desvantagem desse enfoque é que a visão exportada tem que ser mantida quando atualizações são efetuadas nas fontes de dados.

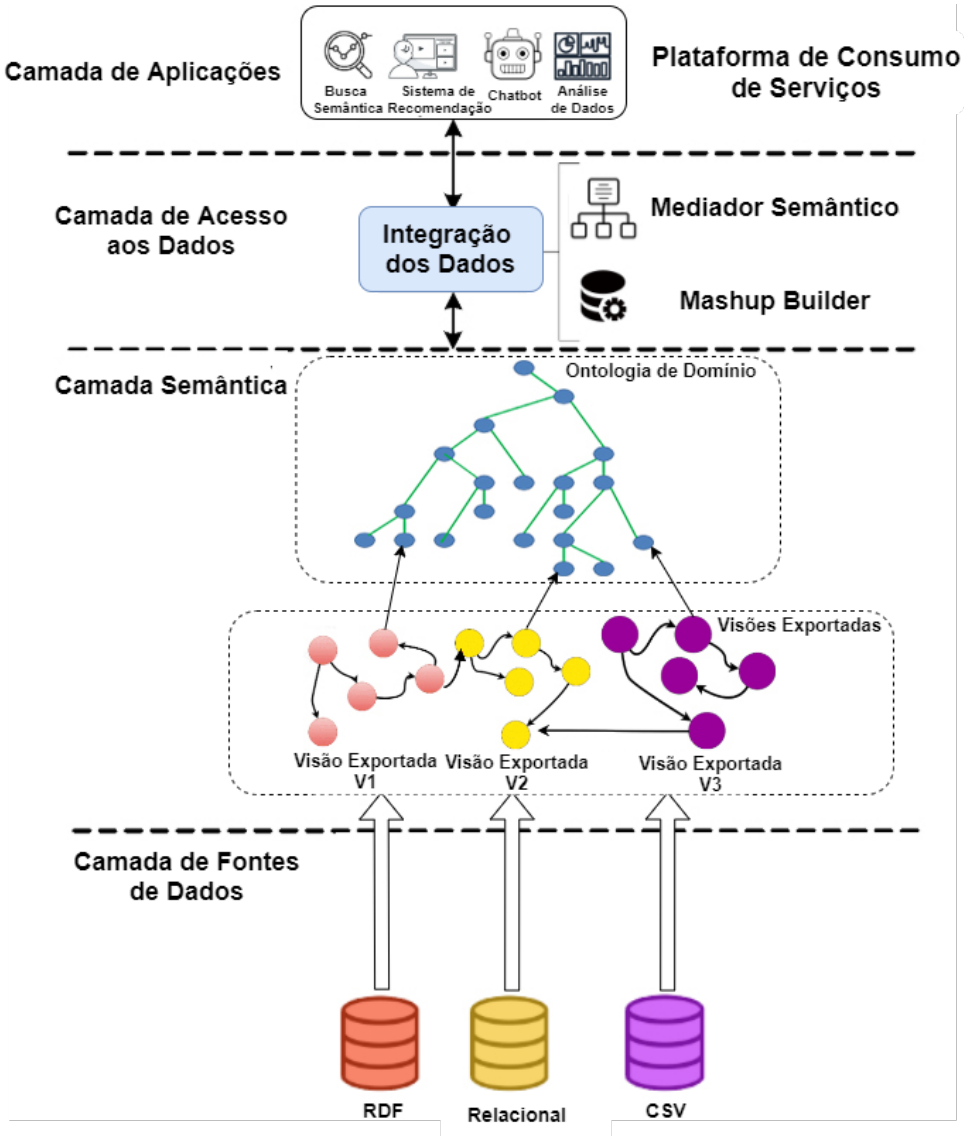
No enfoque virtual, os dados da visão exportada são mantidos nas fontes de dados e o mapeamento é usado para acessar os dados (sob demanda). Nesse caso, o acesso aos dados é realizado através de um Mediador Semântico (MS), que é responsável por traduzir as consultas sob um*wrapper* utilizando a especificação das visões exportadas do EKG, em uma consulta sobre o *schema* das fontes de dados. A vantagem do enfoque virtual é a garantia de que os dados disponibilizados estejam sempre atualizados com relação às fontes originais. Porém, tem a desvantagem do baixo desempenho ao realizar consultas envolvendo várias fontes de dados.

4.1 Visão Arquitetural de SISIFO

Nesta seção é apresentada uma arquitetura para a implantação de plataformas de *Enterprise Knowledge Graphs* com base em SISIFO. Esta arquitetura permite que múltiplas fontes de dados heterogêneas sejam acessadas de uma maneira semanticamente integrada. O uso desta arquitetura permite que aplicações consumam uma visão semântica presente em um EKG, representada por uma ontologia de domínio que integra todas as fontes subjacentes da organização.

A arquitetura apresentada na Figura 4 é dividida em quatro camadas, sendo estas: Camada de Fontes de Dados; Camada de Publicação de Dados; Camada de Acesso aos Dados; e, por fim, a Camada de Aplicações. A seguir estas camadas são apresentadas em mais detalhes.

Figura 4 – Visão Arquitetural de SISIFO.



Fonte: Próprio autor.

4.1.1 Camada de Fontes de Dados

Nesta camada estão as diversas fontes de dados presentes na empresa. Estas fontes podem ser de diferentes formatos, *e.g.*, bancos de dados relacionais, planinhas CSV, *triple stores* RDF, documentos JSON, etc.

Estas fontes podem conter informações complementares sobre objetos em comum, onde a recuperação de todas as informações disponíveis para um mesmo objeto possa possibilitar a implantação de aplicações e estudos mais sofisticados e impactantes. Deste modo, surge a necessidade do acesso integrado a estas fontes.

No entanto, o acesso integrado a essas fontes é um desafio, pois cada uma possui um mecanismo de acesso diferente, o que necessita o uso de diferentes técnicas. Além disso, cada fonte pode ser estruturada seguindo um vocabulário diferente, o que dificulta a compreensão e a relação entre os dados de fontes diferentes.

4.1.2 Camada Semântica

Esta camada torna as fontes de dados subjacentes transparentes, onde nela os dados são publicados através de uma única visão semântica do grafo RDF. A Camada Semântica fornece todos os metadados e informações de especificação e a publicação em si de um EKG. Este EKG expõe todos os dados seguindo um único vocabulário comum, definido pela ontologia de domínio. Onde o acesso aos dados é realizado através de um único método de acesso (consultas SPARQL).

Além disso, no EKG, as diferentes representações de um mesmo objeto do mundo real através das diferentes fontes são identificadas e conectadas por visões de *links owl:sameAs*, o que permite a ligação destas. Mais detalhes sobre a camada semântica e o processo de construção de um EKG são apresentados nas Seções **5.1** e **5.2**.

4.1.3 Camada de Acesso aos Dados

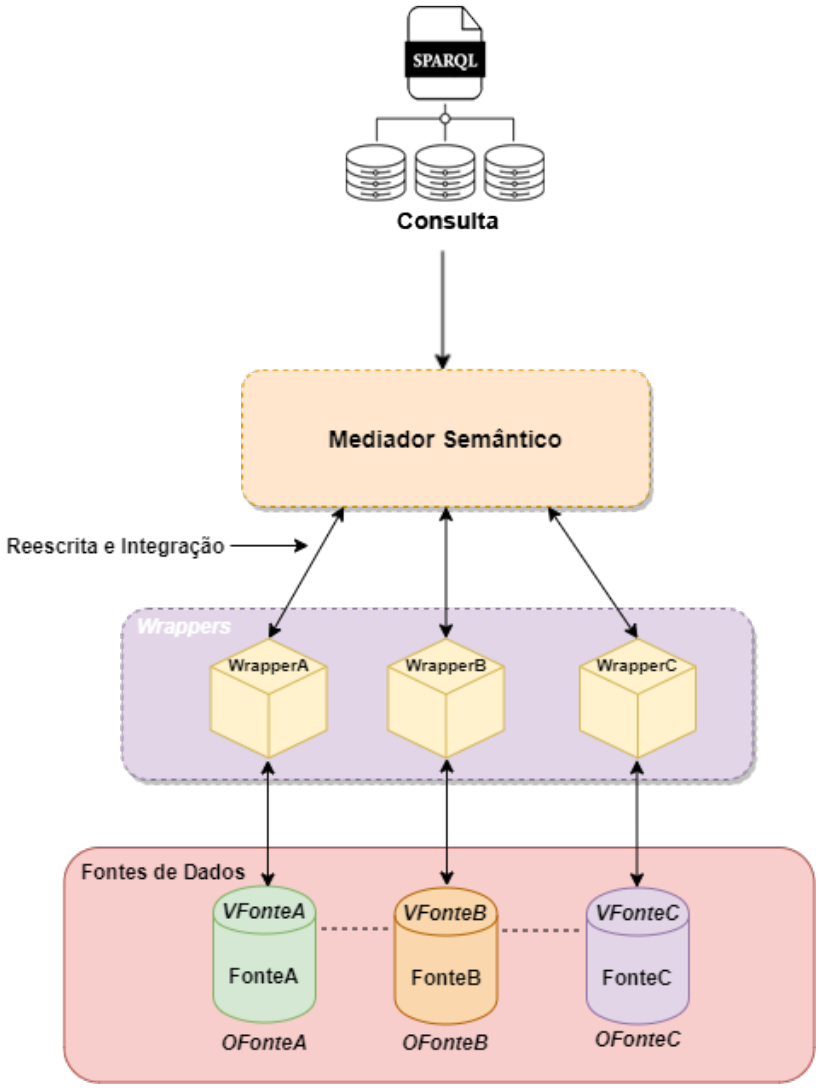
Esta camada permite o acesso ao EKG, onde este pode ser realizado através de dois componentes: o Mediador Semântico MS (*Semantic Mediator*) e o Construtor de *Mashups (Mashup Builder)*.

4.1.3.1 Mediador Semântico

Este componente permite a realização de consultas SPARQL diretamente no EKG. Estas consultas são realizadas de modo federado, com processo seguido sendo representado na Figura 5, onde o mediador trata de:

1. Receber uma consulta SPARQL Q sobre a ontologia de domínio do EKG;
2. Decompor Q, reescrevendo-a como consultas $Q'_1, \ldots, Q'_n$ sobre as visões exportadas publicadas de cada fonte através de *wrappers*;
3. Realizar cada consulta Q'_i, onde $1 \leq i \leq n$, em sua respectiva visão exportada publicada;
4. Agregar os resultadas obtidos pelas consultas $Q'_1, \ldots, Q'_n$ por meio de *links owl:sameAs*;
5. Por fim, retornando uma resposta semanticamente integrada para a consulta Q.

Figura 5 – Processo seguido pelo Mediador Semântico.

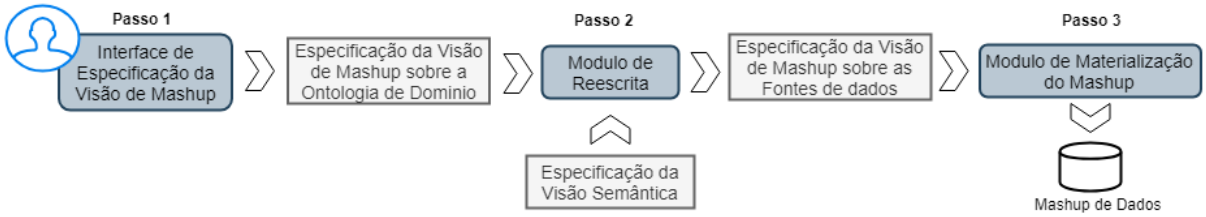


Fonte: Próprio autor.

4.1.3.2 Construtor de Mashups

Mashup Builder é uma ferramenta que permite a construção automática de *mashup* de dados especializado usando a integração semântica à priori das fontes de dados. Como mostrado na Figura 6, o *Mashup Builder* recebe como entrada a especificação da visão de aplicação sobre a ontologia de domínio, e baseado na integração semântica, realiza a materialização do *mashup* especializado requerido pelo usuário.

Figura 6 – Construção de *Mashup* de Dados Especializado.



Fonte: **(CRUZ *et al.*, 2019)**.

O processo de construção do *Mashup* de dados especializado é dado da seguinte forma:

1. Primeiro, o usuário deve especificar a necessidade de informação da aplicação através de uma consulta (facetada) definida sobre a ontologia de domínio;
2. Em seguida, baseado na consulta definida no passo 1 e no resultado da integração semântica, a ferramenta gera automaticamente a especificação da visão do *mashup* sobre as fontes de dados;
3. Por fim, a ferramenta realiza a materialização da Visão do *mashup* em 3 passos:
   - Primeiro, as visões exportadas são materializadas utilizando uma ferramenta como o *Ontop* ou *D2RQ* através mapeamentos *RML*, *R2RML* ou em outra linguagem que mapeiam os dados das fontes de dados em instâncias das visões exportadas por estas fontes;
   - Em seguida, os *links owl:sameAs* entre instâncias que representam um mesmo objeto, seguindo as regras de *linkage* definidas nas especificações de *linkset*;
   - Por último, é realizada a fusão dos objetos - relacionados através de um link *owl:sameAS* - em uma única representação. O processo de fusão é realizado pela ferramenta SIEVE.

4.1.4 Camada de Aplicações

Nesta camada, encontram-se aplicações construídas que consomem os dados semanticamente integrados fornecidos pela camada de acesso. Estas aplicações podem ser implementadas por terceiros que não tenham conhecimento sobre as fontes de dados subjacentes ou que estiveram envolvidas no processo de construção do EKG. Isto ocorre porque todo o processo de construção do EKG é formalmente especificado e publicado como uma ontologia de ***EKG Ontology***, apresentada na Seção **5.3**. Dessa forma, qualquer um pode reusar a integração semântica feita *à priori*, sem que haja a necessidade do auxílio da equipe de construção do EKG.

Esta característica resulta em uma plataforma *Self-Service*, onde o usuário tem acesso a todos os conhecimentos e recursos necessários para a produção de aplicações de maneira independente. Isto diminui o esforço necessário para a criação de novas e ricas aplicações, o que pode gerar avanços disruptivos no fluxo de trabalho e nas atividades executadas na empresa.

5 SISIFO: ABORDAGEM SEMÂNTICA PARA CONSTRUÇÃO DE EKGS

Este capítulo apresenta toda a Abordagem Semântica proposta nesta dissertação para construção de EKGs. Na Seção **5.1**, é apresentada a Camada Semântica para Construção de EKGs proposta neste trabalho; Na Seção **5.2**, a Abordagem Semântica SISIFO é apresentada detalhando todas as suas etapas para construção de um EKG. Na Seção **5.3** expõe-se a Ontologia para representação de EKGs para suporte ao processo de construção do EKG.

5.1 Camada Semântica para Construção de EKGs

Uma das principais dificuldades no gerenciamento de dados corporativos é fornecer acesso integrado às formas complexas de dados armazenados em vários tipos diferentes de bancos e fontes de dados legado (CALVANESE *et al.*, 2017).

Nessa conjuntura, grafos semânticos podem ser utilizados para representar dados armazenados nativamente em outras estruturas e conectar todos os metadados e informações relevantes. Portanto, todo EKG deve possuir um grafo semântico associado. Este trabalho adota a utilização de um grafo semântico representado em uma Camada Semântica para auxiliar a abordagem para construção de EKGs descrita maiores detalhes na subseção 5.2.

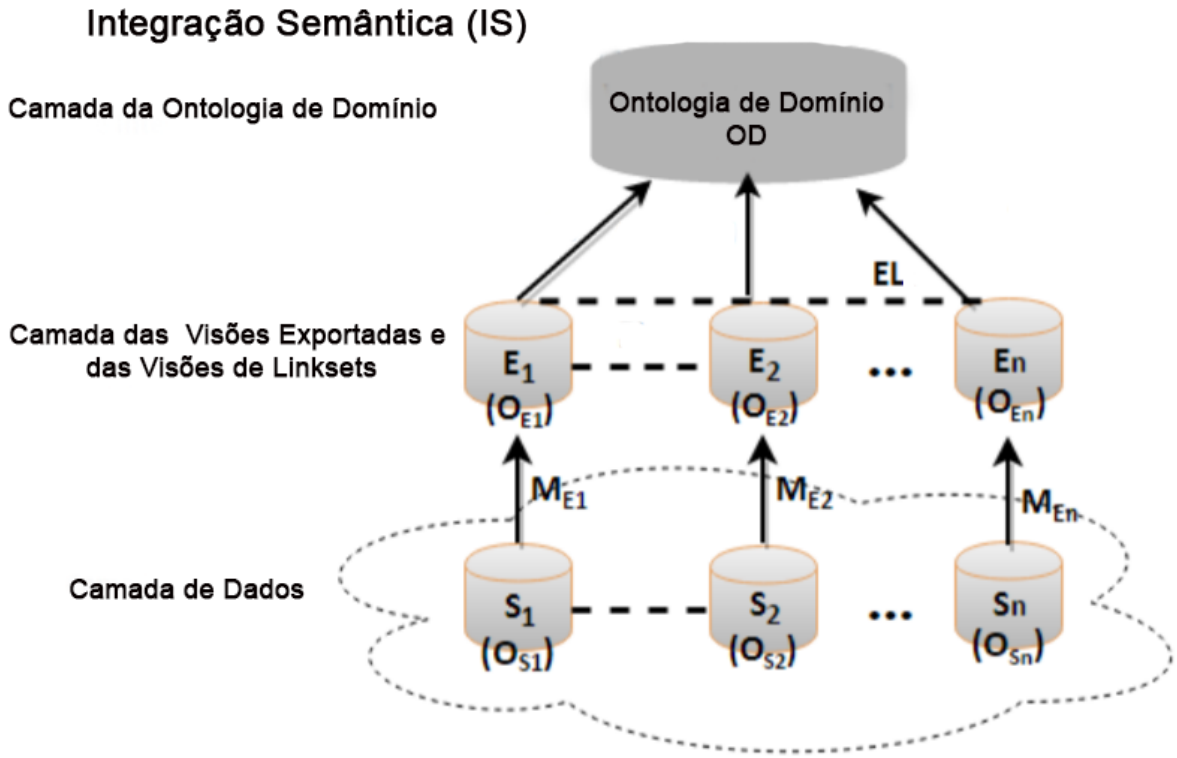
A Camada Semântica provê uma base organizada e modularizada de todo o conhecimento gerado na IS, concentrando metadados e especificações semânticas das visões exportadas a partir das fontes de dados, permitindo um acesso integrado aos dados, bem como o reúso e verificação posterior por parte de usuários. Os principais benefícios da camada semântica, são:

- Fornecer um vocabulário Comum para a Integração e acesso aos dados;
- Fontes de Dados Heterogêneas podem ser integrada mais facilmente;
- Remoção de Barreiras Sintáticas;
- Prover um *framework* robusto para reconciliação de discrepâncias semânticas
- Suporta diferentes visões dos dados (Interoperabilidade semântica);
- Representação de conhecimento Dedutivo e Abdutivo ;
- Oferece flexibilidade para adicionar novas fontes de dados e descoberta de conhecimento;
- Suporta a construção incremental da Semântica do EKG;
- Dados podem ser reusados para vários propósitos e aplicações;
- Facilita a construção sob demanda de *data mashups* através de grafos especializados para análise (visões de dados “population-centric”);

– Suporta o desenvolvimento de aplicações Inteligentes baseadas nas ontologias (Consulta Semântica, Consultas Facetadas, Busca Semântica, Exploração do Grafo, Mineração de Dados Semântica, *Question Answering* e outras aplicações).

Ainda, a Camada Semântica é responsável por determinar a estrutura de representação do conhecimento, e esclarecer o significado semântico em alto nível para os dados de um EKG. A Figura 7 apresenta um *framework* da Integração Semântica adaptado de Vidal *et al.* (2015) utilizado pela Camada Semântica do EKG.

Figura 7 – Framework de Integração Semântica.



Fonte: Adaptado de Vidal *et al.* (2015)

A Camada Semântica do EKG é então definida formalmente como uma tripla $\lambda = (O_D, E, EL)$.

– O_D representa a ontologia de domínio. O_D é responsável por estabelecer um vocabulário comum a ser compartilhado a fim de descrever os dados integrados advindos das fontes de dados;
– E é Uma Visão Exportada definida sobre uma fonte de dados S, utilizando uma ontologia exportada O_E e mapeamentos M_E, formando uma Especificação da Visão Exportada E_S descrita na Subseção 5.3.2.1;
– E_L é uma visão de *linkset*, as quais especificam como identificar objetos em diferentes

fontes de dados que representam o mesmo objeto no mundo real. Cada E_L possui uma especificações da visão de *linkets* E_L_S descrita na Subseção 5.3.6.

Cada Visão Exportada E possui uma Especificação de Visão Exportada E_S contendo uma fonte de dados, ontologia exportada e mapeamentos. Uma E_S é uma tripla $\{S, O_E, M_E\}$, onde:

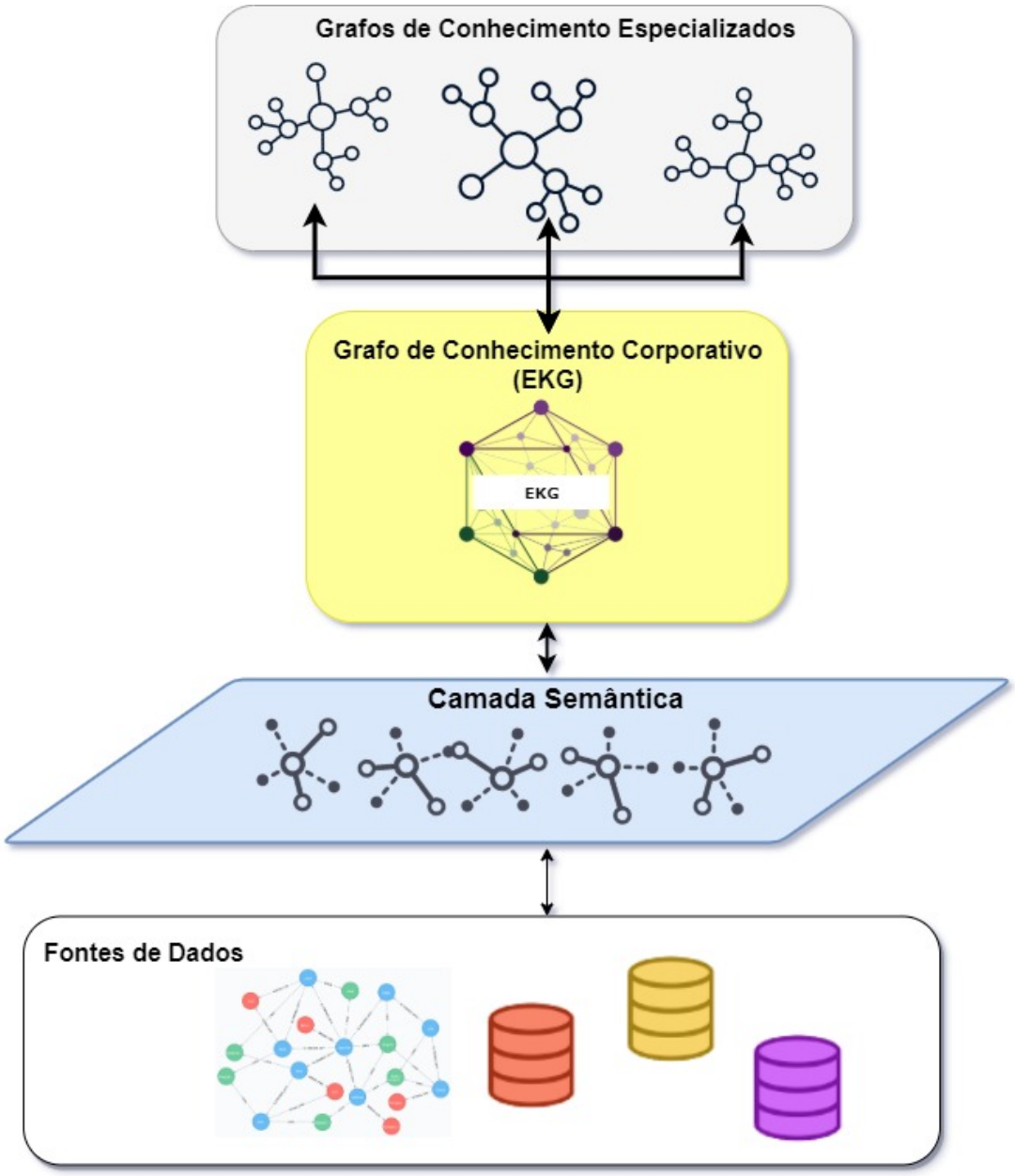
- S**:** é uma fonte de dados;
- O_E**:** é a ontologia exportada, sendo um recorte de O_D, o qual contém os termos de O_D que descrevem o *schema* de uma fonte S;
- M_E**:** é um conjunto de mapeamentos que relacionam termos do vocabulário de O_E com termos do *schema* de uma fonte S.

A compreensão da Especificação da Visão de *Linksets* do EKG E_L_S descrita neste trabalho é vista através de uma adaptação de Vidal *et al.* (2016) como uma quíntupla = $\{E_L\_S = P, OL, F, G, \mu\}$, onde:

- P**:** é um *link type*, este conceito é um link *owl:sameAs*;
- O_E**:** é o vocabulário de correspondência de E_L. Uma O_E está contida em uma E_S de uma E, outros detalhes dessa relação são apresentados na Subseção 5.3.2.1;
- F e G são definições de visões exportadas ***ekgo:ExportedView***. Então, O_E é a ontologia comum para as visões exportadas F e G;
- μ é uma relação $2n$, uma (*Linkage Rule*) de E_L_S é definida por uma tripla = M_C, M_P, L_F, onde:
  - M_C**:** é a *Match Class*, classe utilizada para estabelecer a relação de link *owl:sameAs* entre duas instâncias;
  - M_P**:** é a *Match Property*, propriedade de uma O_D utilizada para calcular uma relação de link entre duas instâncias;
  - L_F**:** é uma *Link Function*, função de link, podendo ser uma Função de Agregação (*Aggregation Function*), Comparação (*Comparator Function*) ou Transformação (*Transformation Function*);

O processo de construção da Camada Semântica é baseado no *pay-as-you-go* (PINKEL *et al.*, 2013) através de SISIFO. Nessa abordagem a camada semântica é construída de forma incremental, e a medida que mais esforço é investido na integração semântica de novas fontes de dados, a camada semântica poderá atender as necessidades de um maior número de aplicações. A Figura 8 expõe a relação entre o EKG e sua Camada Semântica.

Figura 8 – Camada Semântica.



Fonte: Próprio autor.

5.1.1 Acesso aos Dados através da Camada Semântica

A Camada Semântica provê um único ponto de acesso aos dados, e permite que consultas sejam formuladas em termos da ontologia de domínio, de forma que o usuário não precisa entender das fontes de dados, nem das relações entre elas. Existem duas abordagens para acesso aos dados através da Camada Semântica:

Abordagem Materializada: Na abordagem Materializada, ou abordagem ETL (Extrair, Transformar, Carregar), os dados relevantes são extraídos das fontes de dados originais,

transformados em representação RDF de acordo os mapeamentos para a ontologia de domínio, e armazenados em um banco de dados em grafo. Nesse enfoque, as consultas sobre a visão são processadas diretamente sobre a base de dados RDF (visão RDF materializada). Alguns exemplos de *framework* para a materialização da visão RDF são o LDIF (SCHULTZ *et al.*, 2012), ODCleanStore (KNAP *et al.*, 2012) e o Karma (GUPTA *et al.*, 2012).

Abordagem Virtual: Nessa abordagem a visão semântica é uma visão virtual, e consultas definidas sobre a visão semântica devem ser reescrita em consultas sobre as fontes de dados originais. O uso da abordagem virtual garante que os dados disponibilizados estejam sempre atualizados com relação às fontes originais. Porém, a desvantagem é a do desempenho de consultas sobre múltiplas fontes de dados (CALVANESE *et al.*, 2017). Alguns exemplos de ferramentas para o acesso virtual de visões RDF são o Ontop (CALVANESE *et al.*, 2017) e MASTRO (CALVANESE *et al.*, 2011).

5.2 Visão Geral da Abordagem Semântica SISIFO

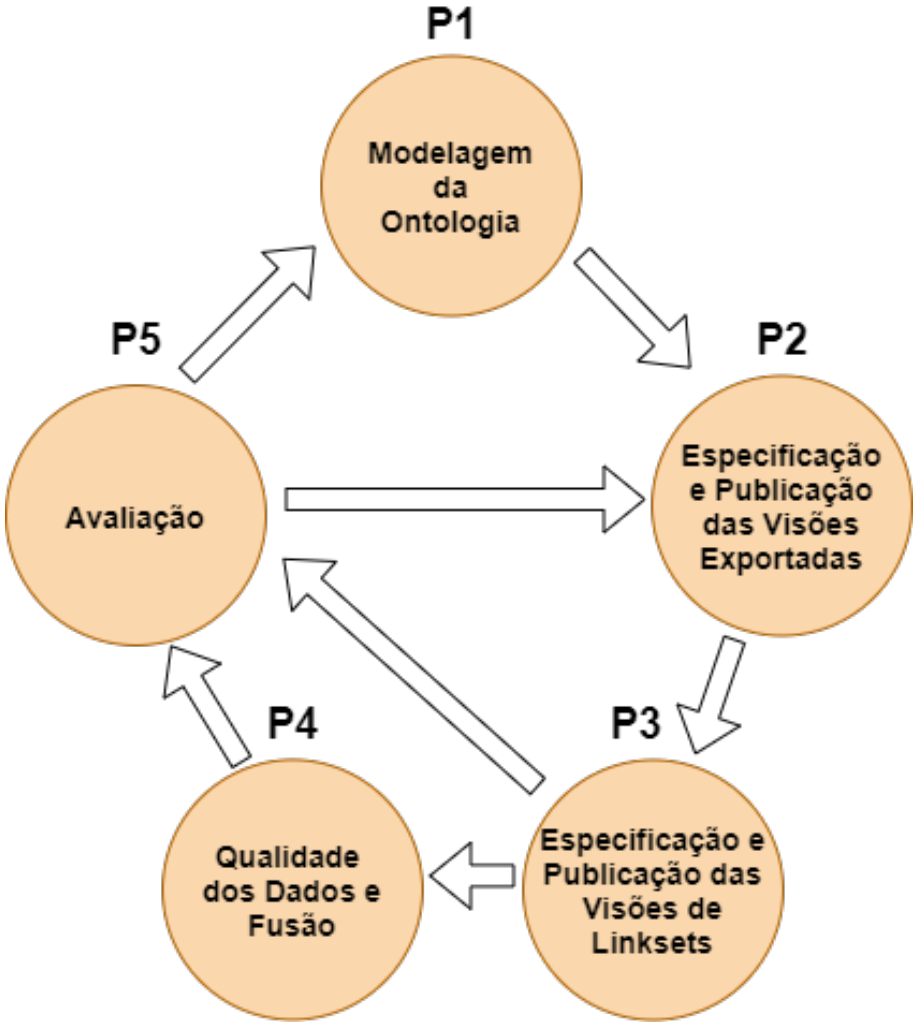
A Abordagem Semântica SISIFO consiste em apresentar um processo para construção de EKGs utilizando-se de sua Camada Semântica para gerir os artefatos envolvidos no processo de IS .

Com base no *pay-as-you-go* SISIFO propõe que EKG seja construído de forma incremental modularizada, isto é, após a construção de um EKG novas inclusões podem ser feitas sob demanda (fontes de dados, *links* e mapeamentos, por exemplo) de modo a não necessitar que seja necessário fazer todo o processo desde o início, evitando assim um esforço e demanda de tempo. Logo, a medida que mais esforço for investido na integração semântica, o EKG poderá atender a essas necessidades através de sua Camada Semântica. Ainda, sua modularização se dá pela organização do EKG em metadados (Camada Semântica) e dados (Grafo de Conhecimento Povoado).

SISIFO também utiliza de uma ontologia (descrita na Seção 5.3) para auxiliar a construção da camada semântica de EKGs, provendo uma conceituação ontológica e utilizando-se de sua capacidade semântica para descrever e representar os componentes de um EKG, (i.e ontologia de domínio, especificação da visão exportada e de *linksets*). Os passos propostos por SISIFO são apresentados conforme Figura 9, sendo descritos logo abaixo.

1. **Modelagem da Ontologia de Domínio:** Etapa contendo o processo de construção da ontologia de domínio, com base no conceitos levantados na etapa de Aquisição do Conhe-

Figura 9 – Passos da Abordagem Semântica SISIFO.



cimento;

2. **Especificação das Visões Exportadas:** Etapa de especificação com base nas fontes de dados, ontologias exportadas e mapeamentos, posteriormente ocorrendo a **Publicação das Visões Exportadas** em um *Endpoint* no caso de visão materializada ou mediante uso de um *Mediador Semântico* tratando-se de visões virtuais;
3. **Especificação das Visões de *Linksets*:** as quais serão usadas para gerar *links* entre instâncias em diferentes fontes de dados e por conseguinte a **Publicação das Visões de *Linksets*** em um *TripleStore* acessível em um *endpoint* ou através de um *Mediador Semântico*;
4. **Qualidade dos Dados e Fusão:** Etapa de identificação e definição de métricas de Qualidade e estratégia de fusão dos dados;
5. **Avaliação:** Etapa de avaliação do EKG com base em tarefas e qualidade do EKG;

5.2.1 Modelagem da Ontologia de Domínio

O primeiro passo consiste em modelar a ontologia de domínio O_D. Os passos para modelagem são basicamente orientados através da representação dos conceitos identificados através de questões de competência (REN *et al.*, 2014) e/ou na reutilização de schemas, vocabulários e outras fontes de dados.

Contudo, o processo de modelagem de uma ontologia é tido como um processo que demanda tempo e esforço, visto a necessidade de compreensão dos conceitos e dados e posterior construção manual, especialmente no que tange à domínios corporativos com grandes bancos de dados (KHARLAMOV *et al.*, 2015).

Entretanto, utilizando o enfoque do *pay-as-you-go* de SISIFO, a ontologia pode ser construída pelo usuário de forma incremental, isto é, partes da ontologia podem ser modeladas, e posteriormente novos pedaços podem ser acoplados as partes já existentes, isso simplifica o processo de construção da ontologia além de viabilizar o reúso. Nessa abordagem, uma ontologia de domínio O_D pode ser construída, e por conseguinte ampliada por outras ontologias O_Dn.

O usuário, como alternativa, pode extrair o conhecimento a partir de fontes de dados de forma automática através da técnica de *bootstrapping* (JIMÉNEZ-RUIZ *et al.*, 2015) e posteriormente utilizar essas ontologias exportadas para construção da ontologia de domínio utilizando algum software como Protégé / WebProtégé (TUDORACHE *et al.*, 2008). Para auxiliar nesse processo, ferramentas podem extrair uma ontologia juntamente com os mapeamentos dos *schemas* a partir de fontes de dados. Na literatura, pode-se encontrar uma ampla variedade de ferramentas como: Ontop (CALVANESE *et al.*, 2017), RDB2OWL (ČERĀNS; BŪMANS, 2011), MIRROR (MEDEIROS *et al.*, 2015) e RDOTE (VAVLIAKIS *et al.*, 2013).

SISIFO propõe dois cenários *(i)* geração automática com base em *schemas* ou outra fonte de dados estruturada e *(ii)* construção manual com base no reúso de recursos não ontológicos como dicionários de dados, tesauros e outros.

1. Uso do cenário de *bootstrapping*, onde havendo fontes de dados *S* relacionais pode-se adotar o uso do RDB2OWL ou *Ontop* para geração da *OD*. SPARQL2XQuery (BIKAKIS *et al.*, 2015) também pode ser utilizado através do módulo XSOWL[1] para lidar com XML. Já para conversão apartir de CSV, CSV2RDF (HAIDER; HOSSAIN, 2018) pode ser utilizado. Ainda, o OpenRefine pode ser utilizado para lidar com planilhas Excel / ODF (HAM, 2013);

[1] https://github.com/istavrak/XS2OWL

2. Alternativamente, não ocorrendo a presença de fontes de dados e *schemas*, a construção da Ontologia pode ser apoiada pela Engenharia de Ontologias (EO), visto que fornece um conjunto de orientações e diretrizes para modelar e construir ontologias de modo formal. Nesse sentido, a EO pode ser direcionada através de metodologias colaborativas e não colaborativas. Em SISIFO são recomendadas as metodologias colaborativas NeOn (SUÁREZ-FIGUEROA *et al.*, 2012) e GOSPL (DEBRUYNE *et al.*, 2013). Já como exemplo de metodologias não-colaborativas são sugeridas as metodologias: METHONTOLOGY (FERNÁNDEZ-LÓPEZ *et al.*, 1997), *On-To-Knowledge Methodology* (SURE *et al.*, 2004) e FMCLGO (GIUNCHIGLIA *et al.*, 2012).

Para a modelagem manual de ontologias, são sugeridas ao usuário as recomendações propostas por Gruber (2003):

- Reusar dados de fontes existentes;
- Gerenciar a integridade dos dados através de um conjunto de restrições;
- Permitir ao usuário a capacidade de realização de consultas sofisticada;
- Representar os dados do negócio advindos de fontes heterogêneas;
- Possibilitar a federação de consultas sob um vocabulário único de representação dos objetos em um negócio e suas definições;

Assim, o processo de criação de ontologias é visto como uma etapa chave para provimento semântico do conhecimento ao EKG. Devendo uma ontologia ser consistente, no que tange a pelo menos três níveis:

1. **Consistência Sintática:** Compreendida com ênfase na representação correta da síntaxe utilizada na linguagem da ontologia;
2. **Consistência Lógica:** A ontologia não pode conter informações contraditórias. Por exemplo, o indivíduo não pode ao mesmo tempo ser uma Pessoa Física e Jurídica, tal situação geraria uma inconformidade lógica semântica, já que ambos os conceitos devem ser disjuntos entre si. Raciocinadores como HermiT (SHEARER *et al.*, 2008) e Pellet (SIRIN *et al.*, 2007) podem ser utilizados para identificar a inconsistência lógica e ferramentas como Protégé / WebProtégé podem ser utilizadas para correção dessas inconsistências;
3. **Consistência de Contexto:** Uma ontologia consistente não infere necessariamente que ela representa com precisão o mundo real. Dado por exemplo uma "Pizza de Carne"e uma "Pizza Vegetariana", a ontologia é logicamente consistente, mesmo que tenhamos definido uma pizza "Vegetariana". No entanto, isso é um erro. Para descobrir esse tipo de problema,

a ontologia precisa ser avaliada por especialistas no domínio.

Sem o seguimento de boas práticas, as ontologias podem se tornar insustentáveis e inutilizáveis. SISIFO sugere outras *guidelines* como:

- **Evite ambiguidade**, para isso, recomenda-se investir os esforços iniciais nas idéias subjacentes dos conceitos, evitando os termos. Estabeleça cada idéia de relação envolvendo os conceitos com qualquer termo conveniente por meio de rótulos que sejam de preferência, escolha as idéias importantes e represente-as com os termos apropriados;
- **Escolha partir do *mid* "meio"** ao invés de escolher a abordagem *top-down* iniciando pelos conceitos mais genéricos ou *bottom-up* iniciando pelos conceitos mais específicos. Por exemplo, utilizar a abordagem *bottom-up* pode resultar em um nível muito alto de detalhamento, o que por conseguinte *i)* pode aumentar o esforço geral para construção da ontologia; *ii)* tornar a identificação de relação entre aspectos em comum dos conceitos; *iii)* tende a aumentar o risco de inconsistências e posteriormente *iv)* leva ao re-trabalho, gerando também um maior custo de tempo.
- **Extensibilidade e Reúso:** Uma ontologia deve ser projetada de forma a maximizar o reúso e a extensão. Isso pode ser alcançado obtendo o equilíbrio entre ser específico o suficiente para executar as tarefas necessárias, mas não tão específico que será de pouca utilidade para outros contextos e aplicações. Durante a modelagem, símbolos e termos técnicos ou muito específicos devem ser evitados, como por exemplo os que são feitos por conveniência de notação ou implementação. Também recomenda-se atenção em relação de termos parecidos que significam aproximadamente a mesma coisa; em vez disso, identifique o termo-chave subjacente e reutilize-o para outros termos. Isso gera parcimônia, o que facilita a reutilização;

Posteriormente, seguindo às orientações apresentadas, o usuário responsável pela construção do EKG pode selecionar uma ferramenta de apoio à modelagem. Para essa tarefa, são recomendadas as ferramentas: Gra.fo[2], OnToology (ALOBAID *et al.*, 2015) e Protégé / WebProtégé. Através de uma das ferramentas, o usuário pode implementar a ontologia de domínio para representação dos conceitos importantes no qual devem constituir a *knowledge base* do EKG.

A etapa de modelagem também pode ocorrer posterior à etapa de avaliação e qualidade, quando havendo alguma inconformidade ou incompletude na representação dos conceitos

[2] gra.fo

relativos ao domínio. Por sua vez, havendo necessidade de reúso, através do *Matching* de Ontologias, são recomendadas algumas abordagens e ferramentas como RiMOM (LI *et al.*, 2008), CODI (HUBER *et al.*, 2011), ServOMap (BA; DIALLO, 2012), MapSSS (CHEATHAM, 2011), LogMap (JIMÉNEZ-RUIZ; GRAU, 2011), YAM++ (NGO; BELLAHSENE, 2012). Por sua vez, o *Matching de Schema* / Alinhamento de Ontologias pode ser feito através das ferramentas AMC (PEUKERT *et al.*, 2011), KSMS (ANAM *et al.*, 2016), e AgreementMaker (CRUZ *et al.*, 2009).

5.2.2 Especificação e Publicação das Visões Exportadas

Para especificação das visões exportadas, faz-se necessário que o Engenheiro de Conhecimento selecione as fontes de dados *S* e sua ontologia exportada *OE* juntamente com os mapeamentos *ME* relativos a elas.

Logo, dado uma fonte de dados *S* são computados um conjunto de mapeamentos *ME* relacionados a *S OE* como um sub-vocabulário de uma ontologia de domínio *OD*.

SISIFO, sugere no caso de *S* relacional, os mapeamentos *ME* podem ser expressos em R2RML(DAS, 2011), linguagem recomendada pela W3C[3] que permite mapear bancos de dados relacionais ou consultas SQL em um vocabulário ontológico destino *OE*.

Comumente o modelo relacional é utilizado na maioria dos casos nos mais variados domínios, entretanto, tratando-se de fontes não relacionais, SISIFO sugere os mapeamentos *ME* sejam representados em RML para lidar com *S* nos formatos CSV, JSON e XML. RML estende a estrutura sintática de R2RML, generalizando seu funcionamento relacional para outros tipos de dados. Por exemplo, uma Tabela (*rr:Logical Table*) passa a ser uma (*rml:Logical Source*), permitindo que RML lide com com diversas fontes ao invés de especificar uma tabela de uma fonte de dados relacional para geração da ontologia exportada *OE* (DIMOU *et al.*, 2014), SPARQL-Generate (LEFRANÇOIS *et al.*, 2017).

Para lidar com outros formatos, contextos, e necessidades além do modelo de dados relacional, outras linguagens de mapeamento baseadas em R2RML vem sendo propostas, tais como: D2RML [JSON, REST, SPARQL e XML] (CHORTARAS; STAMOU, 2018), xR2RML [XML, CSV, JSON] (MICHEL *et al.*, 2017), KR2RML (SLEPICKA *et al.*, 2015) [JSON - dados aninhados] além de outras linguagens declarativas.

Contudo, a definição dos mapeamentos e suas regras de forma manual torna-se muitas vezes uma tarefa inviável, tanto em razão do tempo e esforço, como na demanda de se

[3] https://www.w3.org/

possuir um bom conhecimento acerca das linguagens de mapeamento como R2RML e RML, por exemplo.

Nesse sentido, são recomendadas algumas técnicas e abordagens que propõem a geração automática de mapeamentos através padrões utilizados em bancos de dados relacionais, sugerindo-se o uso das ferramentas MIRROR (MEDEIROS *et al.*, 2015) e BootOX (JIMÉNEZ-RUIZ *et al.*, 2015).

Ainda, os mapeamentos podem ser feitos de forma semi-automática, nesse caso, são sugeridos mapeamentos com base no *schema* da fonte dos dados e o usuário guia o processo supervisionando de modo a buscar corrigir eventuais inconsistências geradas. Para dar suporte a criação de mapeamentos supervisionados, são recomendadas as ferramentas Karma (GUPTA *et al.*, 2012), RMLeditor (HEYVAERT *et al.*, 2016), SQUaRE (BĄK *et al.*, 2017), Map-On(SICILIA *et al.*, 2017) e Juma (JUNIOR *et al.*, 2017).

Após a Especificação das Visões Exportadas de um EKG apresentada na Seção 5.2.2, a Publicação das Visões Exportadas pode ser feita pelo Engenheiro do Conhecimento de modo a povoar o EKG com dados advindos das fontes de dados.

Nesta etapa, a publicação das visões exportadas pode ocorrer de forma virtual ou materializada, cabendo ao Engenheiro do Conhecimento a definição de cada visão exportada como virtual ou materializadas. Mami *et al.* (2011) sugere heurísticas para considerar aspectos na escolha, tais como Frequência de Atualização; Custo de Manutenção da Visão; Tamanho da Visão (Leitura); Espaço/Custo de Armazenamento e *workload* (Frequência de Consulta), por exemplo.

Em SISIFO a publicação das visões exportadas bem como das visões de *linksets* utilizadas na construção do EKG apresentada neste trabalho baseia-se no Ontology Based Data Access (OBDA) (CALVANESE *et al.*, 2011), permitindo assim a publicação utilizando os enfoques: materializado e virtual. No enfoque materializado, uma especificação exportada do EKG contendo as fontes de dados, ontologia exportada e os mapeamentos é utilizada para gerar novos fatos, se a visão for materializada a especificação da visão exportada torna-se um RDF publicado em um *triplestore* como GraphDB[4] ou Virtuoso[5].

Já no virtual, as consultas *SPARQL* são traduzidas em consultas sobre a fonte de dados original utilizando a especificação da visão exportada para processamento sob as fontes

[4] http://graphdb.ontotext.com/
[5] http://vos.openlinksw.com/

de dados. Como sugestão de ferramenta, neste passo podem ser utilizados o Ontop[6] como um mediador semântico juntamente como Teiid[7] para federação sob múltiplas fontes de dados relacionais. Ainda, a implementação pode ocorrer de forma híbrida, isto é, o EKG é construído através da publicação das visões virtuais e materializadas de forma conjunta *on-the-fly* através do mediador semântico durante a consulta.

5.2.3 Qualidade dos Dados e Fusão

A avaliação de qualidade dos dados e Fusão é uma etapa opcional de SISIFO, vista hipótese de não materialização da fontes de dados.

A qualidade das visões exportadas e links de são importantes na determinação da qualidade do tripla, levando em conta também a igualdade e a semelhança dos valores conflitantes.

O EKG pode ser avaliado utilizando métricas direcionadas a avaliação dos dados, assim podem ser utilizadas Metodologias como a proposta de Avaliação de Qualidade de Dados apresentada por (PIPINO *et al.*, 2002) permitem identificar as importantes dimensões da qualidade e seus requisitos sob várias perspectivas.

Essas perspectivas possibilitam uma avaliação semiautomática com base em restrições de integridade de dados. através de, por exemplo, uma avaliação guiada pelo usuário (ZAVERI *et al.*, 2013) ou guiada por testes (KONTOKOSTAS *et al.*, 2014) e uma avaliação manual baseada na experiência humana [avaliação guiada por *crowdsourcing*] (ACOSTA *et al.*, 2013).

Para auxiliar nesse passo, algumas ferramentas podem ser úteis, como SWIQA (Estrutura de Avaliação da Qualidade da Informação na Web Semântica) (FÜRBER; HEPP, 2011) define regras de qualidade de dados e escores de qualidade para identificar dados errôneos. Sieve (MENDES *et al.*, 2012) define uma estrutura para expressar de maneira visível os métodos de avaliação da qualidade e a fusão de *Linked Data*. Sieve utiliza metadados(indicadores) de qualidade sobre *named graphs* e funções de pontuação que definem a avaliação do indicador de qualidade com base em sua dimensão para avaliar a qualidade dos dados conforme definido pelo usuário. Sieve usa o resultado da qualidade no processo de fusão de dados, conforme definido pelos usuários.

[6] https://ontop-vkg.org/
[7] https://github.com/ontop/ontop/wiki/teiid

Validata (HANSEN *et al.*, 2015) é uma ferramenta on-line para testar a conformidade dos dados RDF em relação aos esquemas atuais escritos na linguagem ShEx48 (Shape Expressions). Uma outra ferramenta é o *Luzzu*, um *framework* para avaliação da qualidade de *Linked Data*. Luzzu é composto por quatro módulos principais: i) uma interface extensível para definição novas métricas de qualidade, ii) um *back-end* interoperável, orientado pela ontologia para representar metadados e problemas; iii) um processador escalável para lidar com *endpoints* SPARQL, e iv) um algoritmo para classificação customizável considerando pesos por parte do usuário (DEBATTISTA *et al.*, 2016).

Já na fusão, o objetivo consiste em fundir múltiplas representações da mesma entidade do mundo real em uma única representação. As regras de fusão definem como resolver o problema dos conflitos que podem ocorrer em objetos de fusão. Logo, resolver a inconsistência dos dados melhora a qualidade do EKG.

Em Sieve, a especificação dos requisitos de qualidade, com a ajuda do usuário, tende a ajudar na escolha de qual função usar para resolver um determinado tipo de conflito, avaliando-se a qualidade das triplas geradas.

O ODCleanStore (KNAP *et al.*, 2012), contém metadados de qualidade também podem ser usados no processo de fusão de dados. No entanto, o *ODClean* avalia a qualidade dos triplos gerados no processo de fusão de dados com base na qualidade da fonte de dados e a igualdade e semelhança dos valores conflitantes.

No passo de avaliação da qualidade e fusão SISIFO recomenda o uso do *Sieve* ou *ODClean* para orientação dos passos proposto por cada metodologia / *framework*, além de fornecerem ferramentas estáveis, intuitivas e com funcionalidades que cobrem as atividades envolvidas nesses passos.

5.2.4 Avaliação

A avaliação do EKG proposta por SISIFO é feita mediante sua capacidade de executar tarefas através questões em formato de consultas SPARQL relativas ao domínio de aplicação e medir a qualidade do EKG considerando um conjunto de critérios de qualidade através de uma *checklist*.

A avaliação baseada em tarefas (PORZEL; MALAKA, 2004) demonstra a capacidade de extrair conhecimento a partir da ontologia de domínio (O_D) definida no EKG. Nessa avaliação, será medida a capacidade do EKG realizar tarefas com base em consultas SPARQL.

A definição das tarefas realizadas por consultas SPARQL sobre o domínio baseiam-se na utilização de Questões de Competência (Questões de Competência (QCs)) como forma de representar requisitos funcionais nos quais devem ser contemplados pela ontologia de domínio do EKG.

De acordo com Uschold e Gruninger (1996) em sua definição, QCs são compreendidas como expressões em linguagem natural no formato de pergunta no qual a O_D do EKG deve ser capaz de responder.

Também é proposta uma avaliação de qualidade do EKG com base em uma checklist proposta da adaptação de um conjunto de regras e *guidelines* de qualidade para KGs apresentadas em Färber *et al.* (2018). A avaliação através da *checklist* organiza-se em: critérios *(C)* avaliados por requisitos *(R)* agrupados por dimensões *D*, sendo:

$$R \in C \in D$$

Na composição da *checklist* foram exclusas Dimensões consideradas não reproduzíveis ou não adequadas ao contexto de EKGs, definindo-se 8 (oito) delas para utilização na avaliação:

1. *Precisão (D1)*;
2. *Confiabilidade (D2)*;
3. *Consistência (D3)*;
4. *Compreensão (D4)*;
5. *Interoperabilidade (D5)*;
6. *Acessibilidade (D6)*;
7. *Licença (D7)*;
8. *Interlinks (D8)*;

Cada Dimensão é composta por um conjunto $C_1 ... C_n$ de critérios de qualidade de dados, é formalizado e expresso em termos de função com a faixa de valor de [0, 1]. Chamamos essa função de métrica de qualidade dos dados (FÄRBER *et al.*, 2018).

- **C1:** Validade sintática de documentos RDF;
- **C2:** Validade sintática de literais;
- **C3:** Confiabilidade no nível de *schema* do EKG;
- **C4:** Confiabilidade no nível de declaração / triplas;
- **C5:** Verificação de restrições de esquema durante a inserção de novas instruções;
- **C6:** Consistência das declarações relativas a restrições de classe;

- **C7:** Consistência das declarações relativas a restrições de relação;
- **C8:** Descrição dos recursos;
- **C9:** *Labels* em múltiplos idiomas;
- **C10:** Serialização RDF compreensível;
- **C11:** URIs autodescritivas;
- **C12:** Uso de vocabulários externos;
- **C13:** Provisionamento de vários formatos de serialização;
- **C14:** Interoperabilidade em vocabulário proprietário;
- **C15:** Evite nós em branco e reificação RDF;
- **C16:** Disponibilidade do EKG;
- **C17:** Provisionamento de um endpoint SPARQL público;
- **C18:** Provisionamento de exportações RDF;
- **C19:** Vinculando sites HTML a serializações RDF;
- **C20:** Provisionamento de metadados do EKG;
- **C21:** Fornecimento de informações de licença legíveis por máquina;
- **C22:** Interligando via *owl:sameAs*;
- **C23:** Validade de URIs externas;

Como forma de explicitar e simplificar as necessidades cobertas por cada critério, foram elaborados requisitos de qualidade $R_1 ... R_n$ na forma de um recorte de definição mais claro dos critérios de qualidade no qual o EKG deve estar apto a satisfazer , sendo:

- **R1:** Nenhum erro de sintaxe encontrado;
- **R2:** Gerenciamento, inserção (construção de mapeamentos, regras de linkage ou fusão especificadas) são feitas manualmente por um usuário engenheiro do conhecimento, especialista de domínio ou ambos a partir das fontes de dados;
- **R3:** Provê alguma proveniência aos dados em nível de recurso ou schema;
- **R4:** Utiliza alguma restrição de checagem durante a inserção de novas instâncias;
- **R5:** Os dados de instância são consistentes com as restrições de classe (por exemplo, owl:disjointWith) especificadas em nível de schema ou com restrições de relações;
- **R6:** Os dados são claramente compreensíveis (fornecem valores para a propriedade rdfs:label, rdfs:comment, dc:description);
- **R7:** É fornecido um idioma secundário para as instâncias através de rdfs:label;
- **R8:** Os dados são fornecidos em um formato mais compreensível a humanos como Turtle,

N3 ou NT;

- **R9:** As URIs são auto-descritivas i.e memoráveis e explícitas;
- **R10:** São utilizados vocabulários externos;
- **R11:** São fornecidos múltiplos formatos de representação / serialização;
- **R12:** Existem links em nível de schema para vocabulários externos;
- **R13:** Não há ocorrências de blank nodes e RDF reification;
- **R14:** É fornecido um acesso ao Endpoint do EKG de forma acessível e disponível;
- **R15:** Existe uma página para acesso do EKG em formatos distintos de serialização;
- **R16:** São fornecidos metadados do EKG;
- **R17:** É fornecida uma informação do tipo de licença aplicada sob os dados;
- **R18:** São estabelecidos links *owl:sameAs* entre instâncias;
- **R19:** São realizados links externos a nível de *schema* e/ou com documentos na web;

A Tabela 2 expõe a relação de critérios C avaliados por requisitos R agrupados por dimensões D, sendo: $R \in C \in D$.

Tabela 2 – *Checklist* de Avaliação do EKG.

	C1,C2	**C3**	**C4**	**C5**	**C6,C7**	**C8**	**C9**	**C10**	**C11**	**C12**	**C13**	**C14**	**C15**	**C16,C17,C18**	**C19**	**C20**	**C21**	**C22**	**C23**
D1	**R1**																		
D2		**R2**	**R3**																
D3				**R4**	**R5**														
D4						**R6**	**R7**	**R8**	**R9**										
D5										**R10**	**R11**	**R12**	**R13**						
D6														**R14**	**R15**	**R16**			
D7																	**R17**		
D8																		**R18**	**R19**

Fonte: Adaptado de Zaveri *et al.* (2016)

Para o cálculo da qualidade do EKG, é feito um somatório de todos os valores de R = [0,1], multiplicados pelo peso w de cada requisito (conforme Tabela 3), posterior é feita uma classificação com base no valor do índice da qualidade $i(k)$ final do EKG.

Tabela 3 – Distribuição de pesos para cada requisito de qualidade do \EKG

	R1	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**	**R10**	**R11**	**R12**	**R13**	**R14**	**R15**	**R16**	**R17**	**R18**	**R19**
Peso (w)	2	1	2	1	1	1	0.25	0.25	1	1	1	1	0.25	1	0.50	0.5	0.25	3	1

Figura 10 – Fórmula de cálculo da qualidade i(k) do EKG.

$$i(k) = \begin{cases} \geq \mathbf{15} & \text{Qualidade Desejável} \\ \geq \mathbf{12} \;\&\&\; < \mathbf{15} & \text{Qualidade Intermediária} \\ < \mathbf{12} & \text{Qualidade Indesejável} \end{cases}$$

5.3 Ontologia para Representação de EKGs (EKGO)

Para representar o EKG, e apoiar os passos propostos pela abordagem semântica de construção deste trabalho, foi construída uma representação ontológica definida como EKGO. Na construção da EKGO, foi adotada a utilização da *Unified Foundational Ontology (UFO)* (GUIZZARDI, 2005), uma ontologia de fundamentação que provê um sistema de categorias e relações baseada em um conjunto de teorias advindas da Filosofia, Lógica, Linguagem e da Psicologia Cognitiva e formalmente caracterizada por meio de axiomas lógicos, possibilitando assim, dotar modelos com a capacidade de gerar inferências.

UFO serve como base para a OntoUML devido à sua expressabilidade. Ao utilizar da OntoUML objetiva-se em usufruir do método da engenharia de linguagem baseado em ontologia, proposto por (GUIZZARDI, 2005) e posteriormente refinado por Guizzardi *et al.* (2013), sendo uma linguagem de modelagem cujo metamodelo é baseado em propriedades ontológicas e possuindo um grande número de classes, relações estereotipadas e axiomas. Com base nisso, apoia-se a hipótese de que a OntoUML seja uma linguagem adequada para a construção de uma ontologia para representação de EKGs, já que é baseada em padrões, possibilita proveniência da verbalização da ontologia, suporta a representação de restrições formais específicas ao domínio, possui implementação ontológica, além de possibilitar a detecção e retificação de anti-padrões.

A EKGO representa conceitos que compõem e fazem parte de SISIFO apresentada neste trabalho. A EKGO foi modelada propondo-se a ser compreensível, extensível, especialmente em razão de ser uma proposta inicial para apoiar semanticamente os passos da construção de EKGs.

A captura dos conceitos presentes na EKGO foi realizada através de um processo de aquisição do conhecimento, utilizado com fontes sólidas de conhecimento como: Staab e Studer (2010), Qi *et al.* (2013), Bonatti *et al.* (2019) e Villazón-Terrazas e Hidalgo-Delgado (2019)

À vista disso, o processo de formalização da ontologia ocorreu de forma iterativa, com ênfase em representar e refinar diferentes aspectos em cada iteração de forma interativa, para que os Especialistas em Ontologia pudessem discutir a conceitualização dos conceitos / componentes / artefatos que formam um EKG modelados em OntoUML.

Os requisitos funcionais da EKGO foram definidos com base em Questões de Competência (GRÜNINGER; FOX, 1994). De acordo com Suárez-Figueroa (2010), a adoção de Questões de Competência (QCs) é uma boa técnica para identificar requisitos que devem ser cobertos por uma ontologia. As QCs foram levantadas através de perguntas de consulta criadas

por especialistas em ontologia. Algumas questões definidas são apresentadas na Tabela 4.

Tabela 4 – Tabela com as Questões de Competência da EKGO.

Identificador	Questão de Competência
(QC1)	Quais são os conceitos exportados por cada fonte de dados?
(QC2)	Quais visão exportadas são relacionadas por cada linkset?
(QC3)	Qual Fonte de dados apresenta maior completude (quantidade de registros / tuplas) no EKG?
(QC4)	Qual a proveniência de cada Visão Exportada do EKG?
(QC5)	Quais as Visões Virtuais Exportadas do EKG e os respectivos Wrappers que as acessam?
(QC6)	Quais Fontes de Dados e Mapeamentos formam cada Visão Exportada do EKG?
(QC7)	Quais conceitos de cada Ontologia da Visão Exportada estão representados na Ontologia de Domínio?
(QC8)	Quais das Visões Exportadas são advindas de fontes relacionais?
(QC9)	Quais são os Triplestores de cada Visão Exportada Materializada e seus respectivos endereços de endpoints?
(QC10)	Quais Visões Publicadas do EKG são virtuais e quais são materializadas?

A EKGO foi desenvolvida tendo a compreensão e a usabilidade em mente. Por esse motivo, foi aplicado um *schema* consistente para nomes de propriedades, usando "has"seguido do nome da classe. Nessa ontologia os conceitos relativos a um EKG bem como os artefatos que o compõem são apresentados seguindo uma coerência semântica de relação entre os conceitos. As principais Classes e Propriedades da EKGO são: ***ekgo:EKG, ekgo:ExportedView, ekgo:ExportedViewSpecification, ekgo:LinksetView, ekgo:LinksetViewSpecification, ekgo:-Mappings, ekgo:Mappings, ekgo:hasExportedView, ekgo:hasExportedViewSpecification, ekgo:hasLinksetView e ekgo:hasDomainOntology***.

A versão operacional da EKGO pode ser utilizada como uma ontologia de referência para EKGs nos mais variados domínios, tal aspecto é fundamentado através do fato de que a EKGO baseia-se em uma abordagem formal para construção de EKGs por meio de UFO, fornecendo ainda a possibilidade de reúso e expansão. Para expressar metadados sobre a qualidade de uma visualização semântica, foram utilizados termos do *Data Quality Vocabulary dqv:* (DQV)[8] um vocabulário recomendado pela W3C. Os detalhes para computar a qualidade das fontes de dados, visões exportadas e visões de *linksets* utilizando o DQV podem ser encontrados no trabalho de (ARRUDA *et al.*, 2020).

Como forma de representar a proveniência, foram reutilizados conceitos do vocabulário da *PROV-O prov:*[9] uma ontologia de representação dos atributos de proveniência gerados em diferentes sistemas e sobre diferentes domínios e contextos em conjunto com propriedades do PAV[10]. A EKGO é composta por conceituações explicitadas no modelo por meio de estereótipos, conforme disposto na Tabela 5.

[8] https://www.w3.org/TR/vocab-dqv/
[9] https://www.w3.org/TR/prov-o/
[10] https://pav-ontology.github.io/pav/

Tabela 5 – Tabela de definição dos conceitos e estereótipos da EKGO

Estereótipo	Definição	Conceito
«Category»	Um conceito cujas instâncias compartilham propriedades comuns, mas obedecem a diferentes princípios de identidade	ekgo:ExportedViewPublication, ekgo:LinksetViewPublication e outros.
«Kind»	Um conceito funcional formado por partes distintas, com o princípio de identidade do proprietário que é mantido tanto por suas instâncias quanto em todos os mundos possíveis	ekgo:EKG, ekgo:ExportedView, ekgo:LinksetView, ekgo:Domain Ontology e outros.
«Subkind»	Um conceito onde suas instâncias compartilham o mesmo princípio de identidade de um tipo	rdbso:RelationalDatabase, ldp:RDFSource, ldp:NonRDFSource, ekgo:SemanticMediator e outros.
«Collective»	Um conceito formado por partes iguais, com o princípio de identidade próprios, mantido por suas instâncias e em todos os mundos possíveis	ekgo:ExportedViewSpecification, ekgo:LinksetViewSpecification, ekgo:LinkageRule, ekgo:Mappings e drm:Asset.
«memberOf»	É uma relação de parcialidade entre um *«Kind»* ou um *«Collective»* (como parte) e um *«Collective»* (como um todo)	ekgo:hasSpecification, ekgo:hasDomainOntology, ekgo:hasExportedView, ekgo:hasLinksetView e outros

Neste trabalho, o EKG é representado na EKGO como uma classe ***ekgo:EKG*** composto por visões exportadas, visões de *linksets* e respectivas especificações de visões, os principais componentes da Camada Semântica do EKG são apresentados ontologicamente a seguir.

5.3.1 Ontologia de Domínio

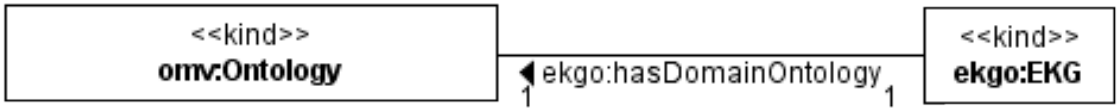
Na EKGO, a ontologia de domínio *OD* é representada como uma classe ***omv:Ontology***. ***omv:Ontology*** é reutilizada do vocabulário *Ontology Metadata Vocabulary omv:*[11] sendo relacionada com ***ekgo:EKG*** através da *Object Property* ***ekgo:hasDomainOntology***. A Figura 11 apresenta uma visão da *OD* representada na EKGO juntamente com sua relação com ***ekgo:EKG***.

5.3.2 Visão Exportada

Uma Visão Exportada E, utilizando a representação semântica da EKGO é definida como:

[11] http://mayor2.dia.fi.upm.es/oeg-upm/index.php/en/downloads/75-omv/index.html

Figura 11 – Representação da Ontologia de Domínio na EKGO.



- ***Visão Exportada (E):*** é uma classe ***ekgo:ExportedView*** que se relaciona com o ***ekgo:EKG*** por meio da *Object Property* ***ekgo:hasExportedView***;
- Uma ***ekgo:ExportedView*** possui uma Especificação de Visão Exportada ES ***ekgo:Exported ViewSpecification*** por meio da *object property* ***ekgo:hasExportedView***;.
- Uma ***ekgo:ExportedView*** pode (<*opcional*>) conter metadados de qualidade representado respectivamente através da relação via *Object Property* ***ekgo:hasQualityMetaData*** com uma ***ekgo:QualityMetaData***;

5.3.2.1 Especificação da Visão Exportada

Uma Especificação de Visão Exportada é uma tripla = S, O_E, M_E, ontologicamente definida como S = ***drm:DataAsset***, O_E = ***omv:Ontology***, e M_E = ***ekgo:Mappings***, formada através da *Object Property* ***«memberOf»*** pela relação das propriedades ***ekgo:hasDataSource, ekgo:hasExportedOntology, ekgo:hasMappings***. A Figura 12 expõe uma visão dos conceitos presentes na especificação da visão exportada.

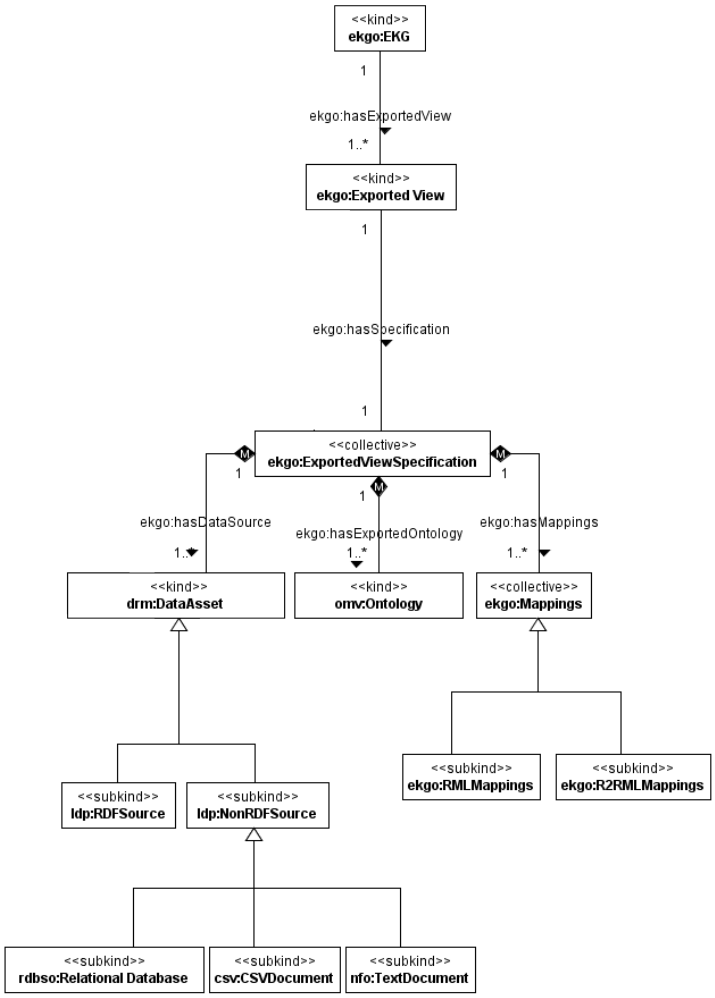
- **Especificação de Visão Exportada (**E_S**):** é representada na EKGO através de uma ***ekgo:Ex***
  portedViewSpecification;
- **Fonte de Dados (**S**:** concerne a um ***«Kind» drm:DataAsset*** na EKGO, sendo um conceito reutilizado do vocabulário *Data Reference Model drm:*[12]. Por sua vez, um ***«Kind> drm:DataAsset*** possui as sub-classes ***ldp:RDFSource*** para representação de fontes de dados RDF e ***ldp:RDFNonSource*** para representação de fontes não-rdf, ambas providas pelo vocabulário *Linked Data Platform ldp:* (LDP) [13]. ***ldp:RDFNonSource*** possui como tipos especializados as subclasses ***rdbso:RelationalDatabase*** para representar fontes de dados relacionais através do vocabulário *Relational Database System Ontology rdbso:* (AGUIAR *et al.*, 2018), ***csv:CSVDocument*** como representação de arquivos no formato *CSV* utilizando como base o vocabulário *CSV Vocabulary csv:*[14] e ***nfo:TextDocument***

[12] https://lov.linkeddata.es/dataset/lov/vocabs/drm
[13] https://www.w3.org/TR/ldp/
[14] https://www.ntnu.no/ub/data/csv-content/

para representar arquivos de texto em formato simples por meio do vocabulário *Nepomuk File Ontology (NFO) nfo:*[15]. Cada ***drm:DataAsset*** possui atributos de proveniência como de onde foi gerada ***prov:wasGeneratedBy*** ou de onde foi derivada ***prov:wasDerivedBy***, além de informações acerca de data de ocorrência.

- **Ontologia Exportada (O_E):** é representada pela classe ***omv:Ontology***;
- **Mapeamentos (M_E):** são representados como uma ***«Collective» ekgo:Mappings*** tendo as subclasses ***ekgo:RMLMappings*** para referenciar mapeamentos genéricos em RML e ***ekgo:R2RMLMappings*** para os mapeamentos de relacional para RDF.

Figura 12 – Representação da Especificação das Visões Exportadas na EKGO.



[15] https://developer.gnome.org/ontology/stable/nfo-TextDocument.html

5.3.3 Publicação da Visão Exportada

Uma publicação de visão exportada refere-se ontologicamente a uma ***«Category» ekgo:ViewPublication***, um conceito genérico ao tipo específico ***«Category» ekgo:ExportedView Publication*** na EKGO.

Uma ***ekgo:ExportedViewPublication*** é acessada diretamente por uma visão exportada ***ekgo:ExportedViewPublication*** através da propriedade ***ekgo:hasPublication***, assim uma ***ekgo:ExportedViewPublication*** obtém a especificação da visão exportada ***ekgo:ExportedView Specification*** apartir da navegação, refletindo a definição de que: <*uma visão exportada possui uma especificação da visão exportada que é posteriormente publicada*>.

5.3.3.1 Publicação da Visão Materializada Exportada

Dada uma especificação das visões exportadas *ES* a materialização é realizada com base através da conversão dos dados de origem no vocabulário da visão exportada E, conforme especificado pelas regras de mapeamento M_E. Por conseguinte a publicação é feita em uma fonte de dados RDF triplificada *Triplestore* contendo um *endpoint*.

No EKGO a publicação da visão materializada exportada é representada através da classe ***ekgo:MaterializedPublication*** que é subtipo de ***ekgo:ExportedViewPublication***, logo tendo relação com uma ***ekgo:ExportedView*** E através da *Object Property* ***ekgo:hasPublication*** que por conseguinte possui uma especificação ***ekgo:ExportedViewSpecification***. Dessa forma, uma ***ekgo:MaterializedPublication*** pode utilizar da especificação para transformação dos dados advindos de E utilizando as regras de mapeamento M_E ***ekgo:Mappings***. A qualidade de cada publicação de visão materializada exportada é adquirida com base na visão exportada e E ***ekgo:ExportedView*** que por sua vez é computada com base na qualidade dos metadados das fontes de dados S ***drm:DataAsset*** e das regras de mapeamento M_E ***ekgo:Mappings***.

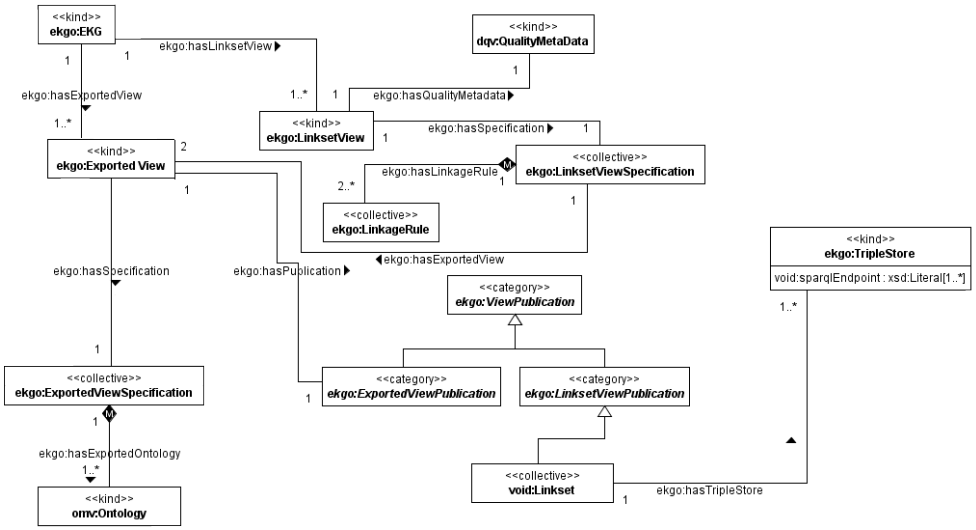
A qualidade dos metadados é representada ontologicamente como uma classe ***dqv:QualityMetaData,*** provida através do vocabulário de representação de conceitos de qualidade, métricas, dimensões e categorias *Data Quality Vocabulary :dqv* [16].

Ao final uma publicação de visão materializada exportada ***ekgo:MaterializedPublication*** é publicada em um *Triplestore ekgo:TripleStore* acessível mediante um endereço de *End Point* representado na *EKGO:TripleStore* como uma *Datatype Property*

[16] https://www.w3.org/TR/vocab-dqv/

textbf*void:sparqlEndpoint*. A Figura 13 apresenta os conceitos relacionados à publicação das visões materializadas exportadas na EKGO.

Figura 13 – Representação da Publicação de Visões Materializadas Exportadas na EKGO.



5.3.4 Publicação da Visão Virtual Exportada

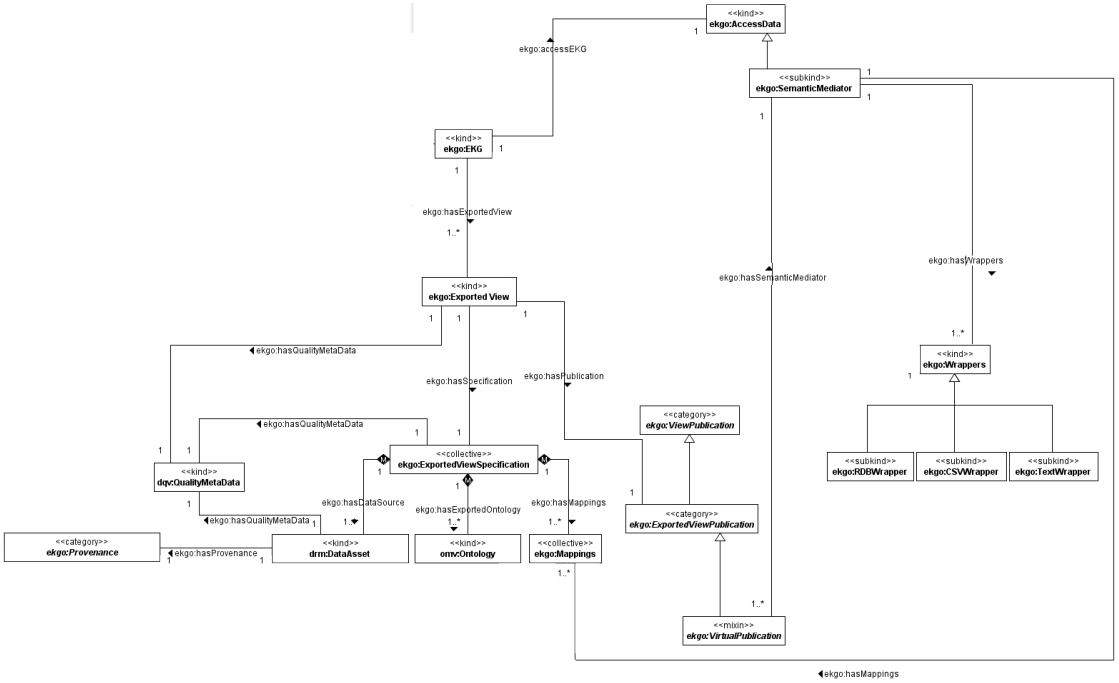
O processo da publicação de visões virtuais do EKG é guiado através da geração de *Virtual Knowledge Graph (VKG)s* (XIAO *et al.*, 2019) conforme abordagem OBDA. No OBDA, através dos mapeamentos uma dada consulta SPARQL é processada através de um *Wrapper* sobre os conceitos e propriedades de uma ontologia exportada, provendo o resultado da consulta em um VKG. Intuitivamente, essa consulta, quando executada sobre as fontes, geraria os dados para atribuição de valor a um conceito / propriedade da visão exportada no qual está associado, obtendo assim um visão virtual exportada.

Logo, um VKG construído com base na abordagem semântica desse estudo, é consistido na publicação de visões virtuais exportadas representada na EKGO através da classe ***ekgo:VirtualPublication***, sendo um conceito *«Mixin»*, dado o fato de uma visão virtual ser algo representativamente abstrato e não materializado.

Uma ***ekgo:VirtualPublication*** possui relação com *Wrapper* representado como classe uma ***ekgo:SemanticMediator*** através da *Object Property* ***ekgo:hasSemanticMediator***. ***ekgo:SemanticMediator*** é responsável por obter uma consulta e redirecionar para o *Wrapper* em respectivo formato da fonte de dados da publicação da visão virtual exportada, sendo: ***ekgo:RDBWrapper*** responsável por lidar com fontes no formato relacional, ***ekgo:CSVWrapper***

para o tratamento de fontes em formato CSV e ***ekgo:TextWrapper*** voltado à arquivos em formato de texto. Um *Semantic Mediator* possui mapeamentos utilizados para processamento das consultas sob os *Wrappers*, sendo expressos na EKGO pela *Object Property* ***ekgo:hasMappings***. Na Figura 14 são mostrados os conceitos do processo de publicação da visão virtual exportada descritos anteriormente na EKGO.

Figura 14 – Representação da Publicação de Visões Virtuais Exportadas na EKGO.



5.3.5 Visão de Linkset

E_L é uma Visão de *Linkset* para especificar um conjunto de regras de ligação entre duas entidades de fontes de dados distintas, sendo representadas no EKGO da seguinte forma:

- **Visão de Linkset (*E_L*):** é representada como uma ***ekgo:LinksetView***;
- Uma ***ekgo:LinksetView*** tem uma especificação *E_LS*;
- ***ekgo:LinksetView*** pode ter sua qualidade *(<opcional>)* mensurada através da relação expressa pela *Object Property* ***ekgo:hasQualityMetaData***;

5.3.6 Especificação da Visão de Linksets

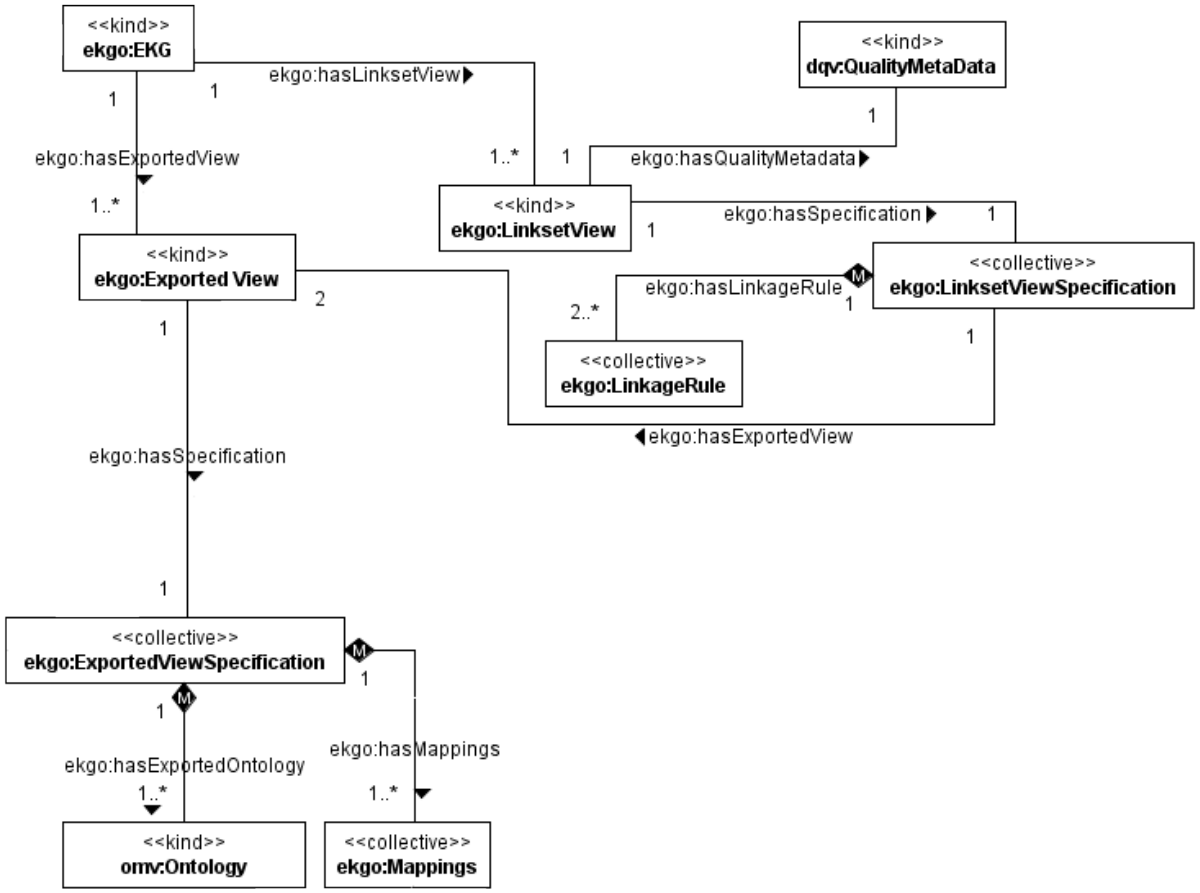
Uma Especificação da Visão de *Linksets* E_LS baseia-se em representar formalmente as definições de linksets e suas regras, descrita da seguinte forma em EKGO:

- **Especificação da Visão de Linksets (*E_LS*):** é uma ***«Collective» ekgo:LinksetViewSpecifi***

cation composta por recursos de uma ***ekgo:LinkageRule*** relacionando-se através da *Object Property* ***ekgo:hasLinkageRule***, conforme EKGO apresentada na Figura 15;

- **Tipo de Link (P):** é uma *object property owl:sameAs*;
- **Ontologia Exportada (O_E):** é representada através de uma classe ***omv:Ontology*** contida em uma ***ekgo:ExportedViewSpecification*** a partir uma instância de ***ekgo:ExportedView***;
- F e G são instâncias de ***ekgo:ExportedView***.
- μ é uma instância de ***ekgo:LinkageRule***.

Figura 15 – Representação da Especificação das Visões de *Linksets* na EKGO.



5.3.7 Publicação da Visão de Linkset

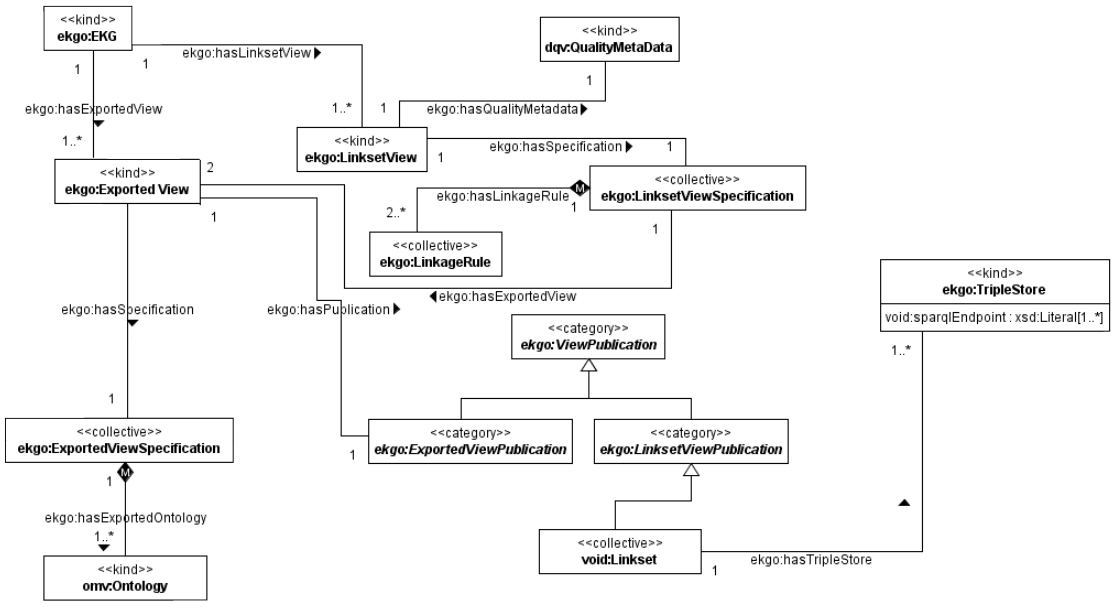
Dada uma E_L do EKG esta etapa publica os links *owl:sameAs* com base na *ELS* exportando-os para a um *Triplestore* ou para um *Mediador Semântico*. Na ontologia EKGO têm-se que a publicação das visões de *linksets* é uma ***ekgo:LinksetViewPublication*** subclasse de ***ekgo:ViewPublication***.

5.3.7.1 Publicação da Visão Materializada de Linksets

Uma publicação de visão de *linkset* materializada é representada ontologicamente como uma ***«Collective» void:Linkset*** um conceito provido pelo *Vocabulary of Interlinked Datasets* (VOiD)[17] para descrever vocabulários e *linksets*. Uma classe ***void:Linkset*** consiste em uma classe para representar *linksets* materializados em RDF. Definida como uma *«Collective»* uma publicação de Links pode possuir 1 ou *n linksets*. ***void:Linkset*** herda de ***ekgo:LinksetViewPublication*** e relaciona-se com ***void:TripleStore*** através da *Object Property* ***ekgo:hasTripleStore***.

A Figura 16 apresenta a publicação das visões de linksets materializadas e suas relações.

Figura 16 – Representação das Visões de *Linksets* Materializados na EKGO.



5.3.7.2 Publicação da Visão Virtual de Linksets

Uma publicação de visão de *linkset* virtual é representada na EKGO como uma ***«Collective» ekgo:LinksetVirtual***. Uma ***ekgo:LinksetVirtual*** estabelece uma relação com um ***ekgo:SemanticMediator*** por meio da *Object Property* ***ekgo:hasSemanticMediator***.

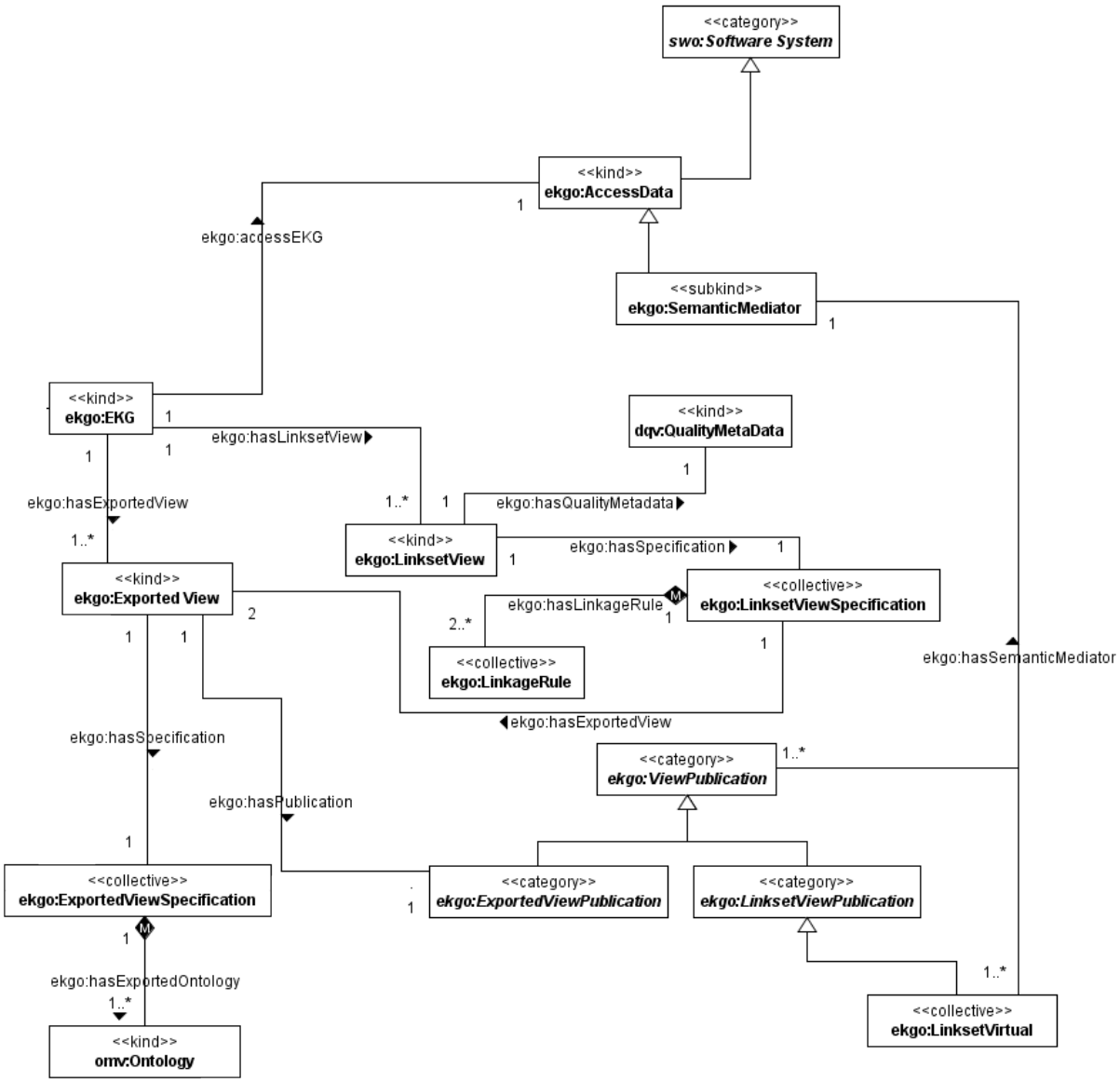
Um ***ekgo:SemanticMediator*** é um *software* para processamento de consultas em fontes heterogêneas descrito pelas classes ***swo:Software System*** e ***ekgo:AccessData***. ***swo:Software System*** é uma conceituação para representação de *softwares*, sendo descrito pela ontologia

[17] https://www.w3.org/TR/void/

apresentada em (AGUIAR *et al.*, 2018). ***ekgo:AccessData*** representa uma classe para acesso ao EKG, mediante uso do Mediador Semântico. Uma instância de ***ekgo:SemanticMediator*** possui uma relação com ***ekgo:Mappings***, que são os mapeamentos que fornecem a definição dos links *owl:sameAs*.

A Figura 17 apresenta um fragmento da EKGO contendo os conceitos relacionados a etapa de publicação das visões de *linksets* virtuais.

Figura 17 – Representação das Visões de *Linksets* Virtuais na EKGO.



6 RESULTADOS

Neste capítulo, é detalhada a avaliação de SISIFO através de um estudo de caso, juntamente com a validação da *EKG Ontology* e seus principais resultados.

6.1 Estudo de Caso - Empresas e Sócios (SEFAZ)

Para validação da Abordagem Semântica proposta em SISIFO, foi utilizado um estudo de caso para o domínio fiscal, especificamente aos dados de empresas e quadros de sócios com ênfase no processo feito por uma Secretaria da Fazenda (SEFAZ).

Coletar e integrar os dados de empresas não é uma tarefa trivial, devido ao tamanho e complexidade crescentes dessas informações. Essas fontes de dados são comumente mantidas por algum órgão público sendo adaptadas ao longo dos anos às necessidades dos aplicativos que às atendem.

Tal característica leva à situação de que as convenções de nomenclatura para elementos de esquema, restrições e a estrutura dos esquemas das fontes de dados são muito complexas e a documentação pode ser limitada ou inexistente, dificultando a extração de dados.

Seguindo práticas comuns das grandes empresas, os auditores da SEFAZ analisam os dados em duas etapas: **(i)** Primeiro acessam e coletam dados relevantes dos bancos de dados disponíveis (Visões de Acesso) e, em seguida, **(ii)** aplicam ferramentas de relatórios analíticos sobre a visão obtida da integração dos dados coletados. A etapa de construção das visões de acesso é geralmente é realizada por meio de uma variedade de interfaces de consulta e ferramentas especializadas de extração e manipulação de dados.

Contudo, flexibilidade das visões de acesso é limitada e, em geral, os usuários podem controlá-las apenas inserindo valores para determinados parâmetros de consulta. Quando as necessidades de informação não podem ser satisfeitas por nenhuma das visões de acesso disponíveis, os auditores, possivelmente com a ajuda da equipe de TI, tentam combinar as respostas obtidas de várias visões de acesso. Em alguns casos, os auditores precisam entrar em contato com a equipe de TI para fornecer uma nova visão de acesso.

As fontes de dados utilizados foram: o Cadastro de Contribuintes do Maranhão e Receita Federal do Brasil (RFB). A Receita Federal do Brasil (RFB) é responsável por cadastrar e gerenciar os dados de empresas e seus sócios regularmente cadastrados. Estes dados são

disponibilizados publicamente na internet[1]. Já o Cadastro de Contribuintes da SEFAZ-MA armazena dados sobre Empresas, Estabelecimentos e seus Endereços, Relações Profissionais e de Representações no estado do Maranhão[2].

Neste estudo de caso, o propósito de construção do EKG foi organizado em duas sub-etapas: *i)* Construir a Camada Semântica utilizada pelo EKG; *ii)* Implantar o EKG com os dados das fontes. SISIFO foi utilizado para organizar essas etapas nos passos apresentados a seguir.

6.1.1 Modelagem da Ontologia de Domínio

Primeiramente, a Ontologia de Domínio foi modelada manualmente através do *Protégé* contendo o vocabulário compartilhado entre as fontes da RFB e do Cadastro da SEFAZ-MA. Para tanto, uma representação dessa ontologia é acessada através do link

Para construção da ontologia de domínio que compõe um EKG, as fontes de dados Cadastro, RFB foram armazenadas no *SGBD PostgreSQL*, visto que na literatura existem um bom conjunto de ferramentas para lidar com *schema* relacional e posterior transformação para RDF.

Com base nisso, a ontologia de domínio foi modelada em formato diagramático através do *Draw.io* seguindo uma representação com base nos conceitos "intermediários - mid", onde, visou-se representar primeiro os conceitos principais / com mais relações, para então criar os conceitos *top* (mais gerais) e *bottom* (mais específicos).

Por conseguinte, a ontologia foi implementada em OWL através do *Protégé*, fornecendo o vocabulário compartilhado entre as fontes da RFB e do Cadastro da SEFAZ-MA. Para tanto, uma representação dessa ontologia é acessada através do link[3]:

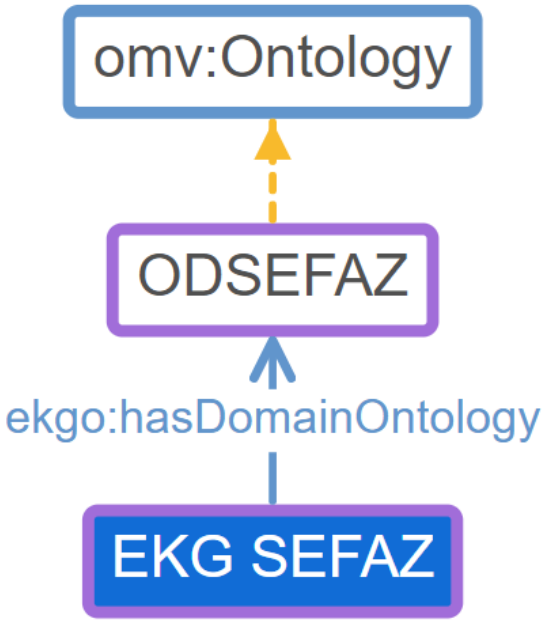
Uma instância de ***ekgo:EKG*** chamada ***EKG SEFAZ*** foi criada, relacionado-se a uma instância da ontologia de domínio instanciada como ***ODSEFAZ*** do tipo ***omv:Ontology*** através da *Object Property* ***ekgo:hasDomainOntology***, conforme apresentado na Figura 18.

[1] http://receita.economia.gov.br/orientacao/tributaria/cadastros/cadastro-nacional-de-pessoas-juridicas-cnpj/dados-publicos-cnpj

[2] https://sistemas1.sefaz.ma.gov.br/portalsefaz/jsp/principal/principal.jsf

[3] tinyurl.com/49gtq9lo

Figura 18 – Instâncias e Classes do passo de Modelagem da Ontologia de Domínio na EKGO.



Nesse passo, SISIFO forneceu uma sequência lógica para construção da ontologia de domínio que será publicada juntamente com as fontes de dados e mapeamentos do EKG, também com base em EKGO, o axioma de inferência na classe EKG:*hasDomainOntology only DomainOntology AND hasDomainOntology exactly 1* - restringe o EKG a ter somente uma relação de *hasDomainOntology* e essa deve ser uma ***DomainOntology*** (Ontologia de Domínio).

6.1.2 Especificação das Visões Exportadas

Neste passo de Especificação das Visões Exportadas, foram criadas duas Visões Exportadas para as fontes de dados do Cadastro e da RFB. A fonte de dados da RFB é publicada trimestralmente no Portal do Ministério da Economia através do Cadastro Nacional de Pessoas Jurídicas[4]. Já a Fonte de Dados do Cadastro de Contribuintes foi obtido através de parceria com a SEFAZ do Maranhão mediante termo de autorização aos dados.

Para cada fonte, foi criada uma Visão Exportada de modo a estabelecer uma representação do Cadastro e RFB junto com suas especificações, tendo sido criadas as instâncias ***VisaoExportadaCadastro*** e ***VisaoExportadaRFB***, da classe *ekgo:ExportedView*.

[4] http://receita.economia.gov.br/orientacao/tributaria/cadastros/cadastro-nacional-de-pessoas-juridicas-cnpj/dados-publicos-cnpj

Após isso, cada Especificação de Visão Exportada foi instanciada como ***EspecificacaoVisaoExportadaCadastro*** e ***EspecificacaoVisaoExportadaRFB*** respectivamente, sendo utilizada a propriedade ***ekgo:hasExportedViewSpecification*** para relacionar as instâncias ***VisaoExportadaCadastro*** e ***VisaoExportadaRFB*** com suas especificações.

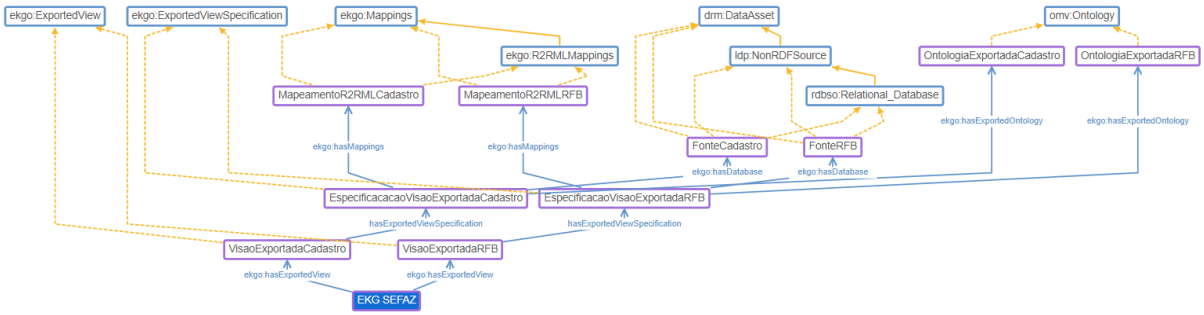
As duas fontes de dados Cadastro e RFB foram representadas como instâncias ***FonteCadastro, FonteRFB*** do tipo ***rdbso:RelationalDatabase*** por meio da propriedade ***rdf:type***. A Ontologia Exportada de cada foi extraída a partir de um recorde da Ontologia de Domínio e o *schema* relacional das fontes de dados com base nos conceitos pertencentes a cada fonte de forma específica. Nesse passo, foram criadas as instâncias ***OntologiaExportadaCadastro e OntologiaExportadaRFB*** da classe ***omv:Ontology***.

Os mapeamentos foram implementados através do *R2RML* e revisados através da ferramenta *Map-On*. Para representar os mapeamentos voltados as fontes relacionais, foram criadas as instâncias ***MapeamentoR2MLCadastro e MapeamentoR2RMLRFB*** da classe ***ekgo:R2RMLMappings***.

Após a definição dos mapeamentos, a especificação de cada visão exportada foi construída, relacionando as especificações através das propriedades ***ekgo:hasDataSource, ekgo:hasExportedOntology, ekgo:hasMappings*** com as instâncias das fontes de dados = ***{FonteCadastro e FonteRFB}***, ontologias exportadas = ***OntologiaExportadaCadastro e OntologiaExportadaRFB*** e mapeamentos = {***MapeamentoR2MLCadastro e MapeamentoR2RMLRFB***}.

Ao final, é estabelecida a relação entre um EKG e uma Visão Exportada através da ***EKG SEFAZ*** utilizando a propriedade ***ekgo:hasExportedView*** tendo com *rdfs:range* as instâncias ***VisaoExportadaCadastro*** e ***VisaoExportadaRFB***. A Figura 19 apresenta uma visão das instâncias envolvidas na especificação das visões exportadas e suas relações com base na EKGO.

Figura 19 – Instâncias e Classes do passo de Especificação das Visões Exportadas na EKGO.



Por utilizar SISIFO, a especificação da visão exportada foi realizada de modo formal,

a fim de guiar o usuário durante cada um dos passos na especificação. Sem essas orientações, dúvidas poderiam ser geradas no que tange ao tipo de mapeamento a se usar para cada fonte, ou como representar uma visão exportada e semântica de cada fonte, por exemplo.

6.1.3 Publicação das Visões Exportadas

Após a realização dos passos de Especificação das Visões Exportadas, é realizada a publicação das visão das fontes *Cadastro e RFB*. Em razão da constante frequência de atualização da fonte de Cadastro, optou-se pelo uso do enfoque virtual para Publicação das Visões Exportadas de modo a visar a recência dos dados. Por outro lado, a Fonte da Receita é publicada trimestralmente e não demandaria desse requisito, contudo, seu volume é enorme, o que torna impraticável seu armazenamento e manutenção em um ambiente não tão robusto como a máquina onde foi executada esse estudo de caso.

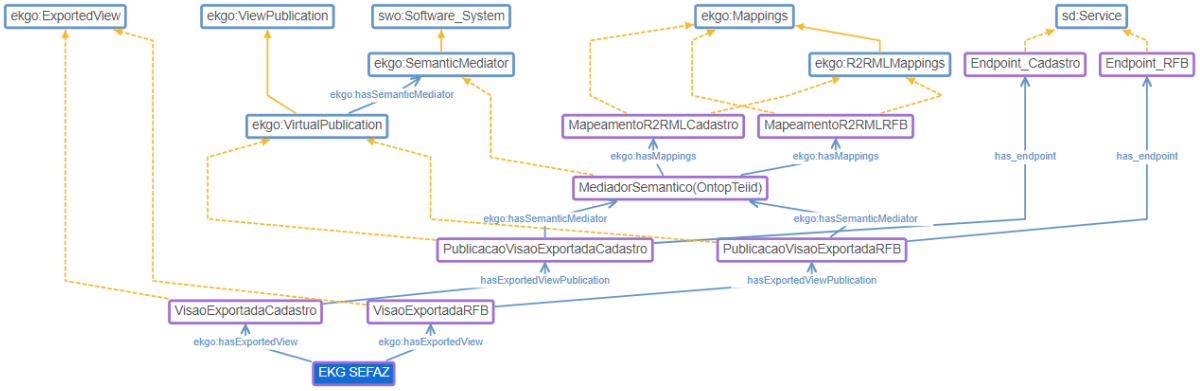
Tendo uma publicação no enfoque virtual, esse estudo de caso não realizou o passo de Qualidade e Fusão, tendo em vista que os dados não são materializados, não tornando viável a adoção de métricas de qualidade e fusão em tempo de consulta.

Nesse passo, a publicação das visões exportadas virtuais foi realizada utilizando o *software* Ontop juntamente com o Teiid, uma proposta de Mediador Semântico para realização de consultas federadas sob fontes relacionais. Com base na ontologia EKGO criou-se então uma instância ***PublicacaoVisaoExportadaCadastro*** e outra ***PublicacaoVisaoExportadaRFB*** do tipo ***ekgo:VirtualPublication*** relacionada com a instância ***MediadorSemantico(OntopTeiid)*** do tipo ***ekgo:SemanticMediator*** através da propriedade ***ekgo:hasSemanticMediator***. A instância ***EKG SEFAZ*** relaciona-se com uma publicação de visão exportada utilizando a propriedade ***ekgo:hasExportedView*** tendo o *range* as instâncias ***PublicacaoVisaoExportadaCadastro*** e ***PublicacaoVisaoExportadaRFB***.

Na Figura 20 são apresentados os detalhes da relação das instâncias durante a publicação das visões virtuais exportadas.

Ao realizar uma publicação seguindo SISIFO, um usuário construtor do EKG já obtém uma estrutura semanticamente organizada para realizar a publicação a partir da especificação das visões exportadas de maneira correta, permitindo a utilização das ferramentas do *Ontop* e *Teiid* com base nos passos propostos e o reúso.

Figura 20 – Instâncias e Classes do passo de Publicação das Visões Exportadas na EKGO.



6.1.4 Especificação das Visões de Linksets

Na Visão de Linkset desenvolvida neste estudo de caso, o intuito foi identificar se duas empresas representam um mesmo indivíduo no mundo real. Foi definida uma Visão de *Linkset* usando instâncias relativas ao processo de especificação foram criadas como ***VisaoLinksetCadastroRFB*** do tipo ***ekgo:LinksetView*** e sua especificação ***ekgo:EspecificacaoVisaoLinksetsCadastroRFB***.

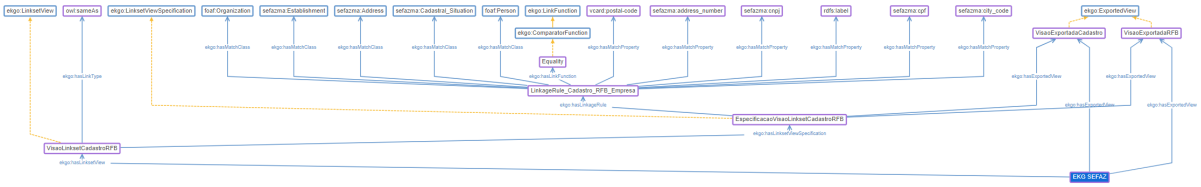
ekgo:EspecificacaoVisaoLinksetsCadastroRFB possui uma *Linkage Rule* composta pelas seguintes *Match Class's*: ***foaf:Organization***, ***foaf:Person***, ***sefazma:Establishment***, ***sefazma:Cadastral_Situation***, ***sefazma:Address***, ***sefazma:City***, ***sefazma:Federative_Unit*** e as seguintes *Match Properties*: ***vcad:cnpj***, ***sefazma:cpf*** , ***sefazma:address_number***, ***sefazma:city_code***, ***vcard:postal-code***, ***rdfs:label*** um *Link Type* ***Equality***.

Após, uma Visão de *Linkset* foi instanciada como ***VisaoLinksetCadastroRFB*** do tipo ***ekgo:LinksetView*** tendo uma ***EspecificacaoVisaoLinksetsCadastroRFB*** através da propriedade ***ekgo:hasLinksetViewSpecification***.

As visões exportadas *source* e *target* que fazem parte da especificação são acessadas através da visões exportadas ***VisaoExportadaCadastro*** e ***VisaoExportadaRFB*** mediante propriedade ***ekgo:hasExportedView***. Por conseguinte, uma instância ***EKG SEFAZ*** adquire acesso ao conteúdo das visões exportadas por meio da propriedade ***ekgo:hasLinksetView*** possuindo como *range* uma instância ***VisaoLinksetCadastroRFB***. A Figura 21 apresenta na EKGO as instâncias e classes da camada semântica povoada no EKG descritas nesta fase.

Tendo as orientações, o usuário pode especificar as regras de *linkage* identificando as propriedades adequadas e definindo os *links* via *owl:sameAs*, evitando problemas na resolução de identidade dos conceitos.

Figura 21 – Instâncias e Classes do passo de Especificação das Visões de Linksets na EKGO.



6.1.5 Publicação das Visões de Linksets

A publicação das visões de *linksets* ocorreu em um MS, que tratou de obter a especificação das visões de links e realizar o processo de identificar as relações do tipo $< S, owl : sameAsO >$ associando as instâncias em comum através dos mapeamentos R2RML de cada visão exportada.

Logo, com base na EKGO, primeiro é criada uma instância ***VisaoLinksetCadastroRFB*** do tipo ***ekgo:LinksetView*** associada com uma instância ***EspecificacaoVisaoLinksetsCadastroRFB*** através da propriedade ***ekgo:hasLinkSetViewSpecification***. Após estabelecer a relação entre uma Visão de *Linkset* e uma Especificação ocorre a realização da publicação. Uma publicação de visão de *linkset* virtual é representada por meio de uma instância do tipo ***ekgo:VirtualPublication*** chamada ***PublicacaoLinksetsCadastroRFB***. Uma ***PublicacaoLinksetsCadastroRFB*** é publicada em uma instância de um ***MS*** do tipo ***ekgo:SemanticMediator***.

O MS realiza a virtualização dos links de forma *on-the-fly* através dos mapeamentos R2RML, onde ocorre um mapeamento de um *subject* com um *object* via propriedade *owl:sameAs* com base em cláusulas de $j_o in$ ou $where x = y$.

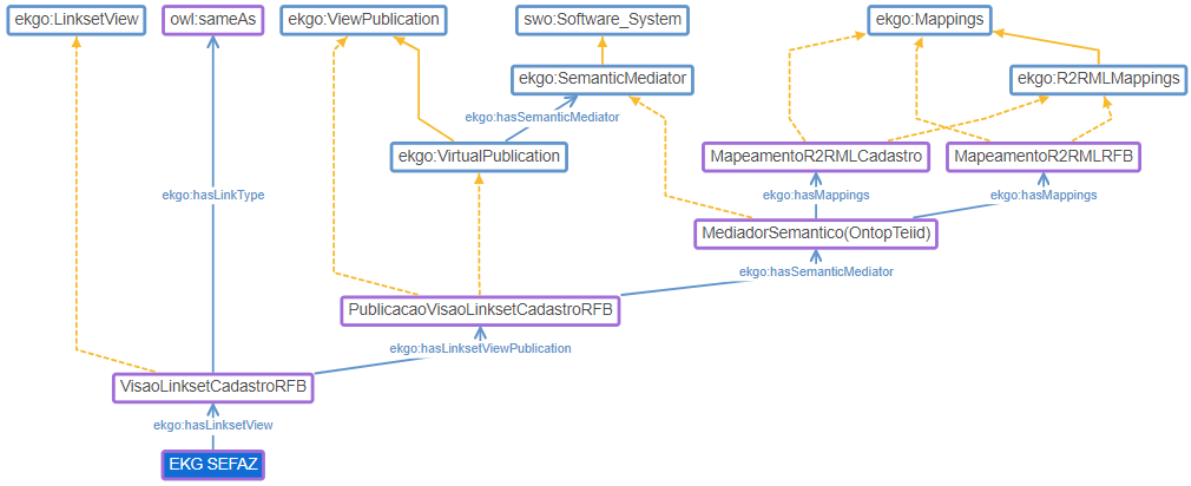
Para publicação dessa visão de *linksets* no EKG, a instância ***EKG SEFAZ*** adquire acesso a uma visão de *linkset* ***VisaoLinksetCadastroRFB*** contendo uma especificação das visões de *linksets* mediante uso da propriedade ***ekgo:hasLinksetView*** possuindo como *rdfs:range* uma instância ***VisaoLinksetCadastroRFB***. Na Figura 22 são apresentadas as instâncias e classes descritas no passo de publicação da Visão com base na EKGO.

6.1.6 Avaliação

Neste passo, o objetivo consistiu em avaliar EKG em relação a sua capacidade de cobertura do domínio através de tarefas (PORZEL; MALAKA, 2004), bem como identificar se as relações entre as instâncias das fontes foram feitas de forma adequada.

O EKG SEFAZ, inicialmente foi feita a carga do EKG em um repositório no *TripleS-*

Figura 22 – Instâncias e Classes da etapa de Publicação das Visões de Linksets na EKGO.



tore GraphDB[5]. Posteriormente foram formuladas algumas questões de competência (conforme abaixo) em linguagem natural, sendo:

- Quais as informações integradas de uma Empresa entre as fontes?;
- Quais Estabelecimentos de uma Empresa que apresentam portes distintos nas fontes da RFB e de Cadastro?;
- Quais Estabelecimentos estão com situações cadastrais diferentes com relação às fontes da RFB e de Cadastro da SEFAZ?;
- Quais Estabelecimentos de uma Empresa que não estão situados no Maranhão?;
- Quais dados de endereços de um mesmo estabelecimento nas bases da RFB e Cadastro são divergentes?;
- Quais Sócios de uma Empresa existem na RFB que não estão presentes no Cadastro da SEFAZ e vice-versa?;
- Com base nos dados integrados, existem diferentes códigos de município para uma mesma cidade?;

Para realização das consultas em SPARQL sob o EKG foram feitos os seguintes passos sequencialmente descritos a seguir:

1. O usuário realiza sua consulta com base no *schema* da ontologia utilizando um Mediador Semântico;
2. É construída uma consulta SPARQL Q que representa a intenção do usuário ;
3. O Mediador Semântico traduz a consulta SPARQL Q para uma consulta SQL Q' sobre um *schema* global que representa todas as fontes de dados como uma única fonte;

[5] https://www.ontotext.com/products/graphdb/

4. O MS identifica as fontes não triplificadas NS (Cadastro e RFB) Q' e para cada fonte NS_j , onde $1 \leq j \leq n$, produz uma sub-consulta Q'_j sobre as visões exportadas do EKG SEFAZ advindas das fontes NS_j através do *Ontop* juntamente com o Teiid que traduz a subconsulta em Q'_j uma consulta sobre o *schema* das fontes;
5. O MS executa cada uma das consultas $Q'_1, Q'_2, ..., Q'_n$ em suas respectivas fontes de dados e recupera as tuplas de resposta com base nas visões exportadas do Cadastro e RFB;
6. O MS realiza o processo de linkagem com base nas visões dos *linksets owl:sameAs* a partir dos mapeamentos presentes nas especificações de cada visão exportada;
7. O resultado é retornado ao usuário / aplicação;

As questões de competência foram realizadas com base em uma Organização: Supermercado Mateus. Abaixo são apresentadas as consultas SPARQL equivalentes a cada questão de competência de tarefa e seus respectivos resultados:

Consulta 1 – Quais as informações integradas de uma Empresa entre as fontes?

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX sefazma: <http://www.sefaz.ma.gov.br/ontology/>

SELECT DISTINCT ?organizacao ?propriedade ?objeto {
   ?organizacao a foaf:Organization;
      owl:sameAs ?organizacao2;
      ?propriedade ?objeto.
}
```

Figura 23 – Resultados da Consulta para a Consulta 1.

	organizacao	propriedade	objeto
1	foaf:Organization/039955150/	rdf:type	foaf:Organization
2	foaf:Organization/039955150/	rdf:type	foaf:Agent
3	foaf:Organization/039955150/	owl:sameAs	foaf:Organization/039955150/
4	foaf:Organization/039955150/	owl:sameAs	foaf:Organization/39955150-MATEUS%20SUPERMERCADOS%20S.A
5	foaf:Organization/039955150/	owl:sameAs	http://www.sefaz.ma.gov.br/RFB/resource/Organization/0399 MATEUS_SUPERMERCADOS_SA
6	foaf:Organization/039955150/	rdfs:label	"MATEUS SUPERMERCADOS S.A."
7	foaf:Organization/039955150/	rdfs:label	"MATEUS SUPERMERCADOS S A "
8	foaf:Organization/039955150/	rdfs:label	"39955150-MATEUS SUPERMERCADOS S.A"
9	foaf:Organization/039955150/	foaf:homepage	"VAZIO"
10	foaf:Organization/039955150/	:company_name	"MATEUS SUPERMERCADOS S.A."
11	foaf:Organization/039955150/	:company_name	"MATEUS SUPERMERCADOS S A "
12	foaf:Organization/039955150/	:company_name	"39955150-MATEUS SUPERMERCADOS S.A"
13	foaf:Organization/039955150/	:person_identifier	"03995515"
14	foaf:Organization/039955150/	:cnpj	"03995515"
15	foaf:Organization/039955150/	:cnpj	"039955150"

Consulta 2 – Quais Estabelecimentos que apresentam portes distintos nas fontes da RFB e de Cadastro?

```
1  PREFIX foaf: <http://xmlns.com/foaf/0.1/>
2  PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
3  PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
4  PREFIX owl: <http://www.w3.org/2002/07/owl#>
5  PREFIX sefazma: <http://www.sefaz.ma.gov.br/ontology/>
6  PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
7
8  SELECT DISTINCT ?estabelecimento ?estabelecimentoSameAs ?
     cnpj ?label_porte_cadastro ?label_porteSameAs_RFB WHERE
     {
9      ?organizacao a foaf:Organization;
10                sefazma:has_establishment ?estabelecimento.
11     ?estabelecimento sefazma:cnpj ?cnpj;
12              sefazma:has_size ?porte.
13     ?porte rdfs:label ?label_porte_cadastro.
14
15     ?organizacao owl:sameAs ?organizacaoSameAs.
```

```
    ?estabelecimento owl:sameAs ?estabelecimentoSameAs.
    ?estabelecimentoSameAs sefazma:cnpj ?cnpj2;
                sefazma:has_size ?porteSameAs.
    ?porteSameAs rdfs:label ?label_porteSameAs_RFB.
}
```

Figura 24 – Resultados da Consulta para a Consulta 2.

	estabelecimento	estabelecimentoSameAs	cnpj	label_porte_cadastro	label_porteSameAs_RFB
1	http://www.sefaz.ma.gov.br/Cadastr 12444861/	http://www.sefaz.ma.gov.br/Cadastr 12444861/	"03995515006521"	"EPP"	"EPP"
2	http://www.sefaz.ma.gov.br/RFB/res ELETRO_MATEUS	http://www.sefaz.ma.gov.br/Cadastr 12444861/	"03995515006521"	"EPP"	"EPP"
3	http://www.sefaz.ma.gov.br/Cadastr 12444861/	http://www.sefaz.ma.gov.br/RFB/res ELETRO_MATEUS	"03995515006521"	"EPP"	"EPP"
4	http://www.sefaz.ma.gov.br/RFB/res ELETRO_MATEUS	http://www.sefaz.ma.gov.br/RFB/res ELETRO_MATEUS	"03995515006521"	"EPP"	"EPP"
5	http://www.sefaz.ma.gov.br/Cadastr 12627415/	http://www.sefaz.ma.gov.br/Cadastr 12627415/	"03995515015199"	"EPP"	"EPP"
6	http://www.sefaz.ma.gov.br/RFB/res ELETRO_MATEUS	http://www.sefaz.ma.gov.br/Cadastr 12627415/	"03995515015199"	"EPP"	"EPP"
7	http://www.sefaz.ma.gov.br/Cadastr 12627415/	http://www.sefaz.ma.gov.br/RFB/res ELETRO_MATEUS	"03995515015199"	"EPP"	"EPP"
8	http://www.sefaz.ma.gov.br/RFB/res ELETRO_MATEUS	http://www.sefaz.ma.gov.br/RFB/res ELETRO_MATEUS	"03995515015199"	"EPP"	"EPP"
9	http://www.sefaz.ma.gov.br/Cadastr 12376404/	http://www.sefaz.ma.gov.br/Cadastr 12376404/	"03995515004405"	"EPP"	"EPP"
10	http://www.sefaz.ma.gov.br/RFB/res MATEUS_SUPERMERCADOS	http://www.sefaz.ma.gov.br/Cadastr 12376404/	"03995515004405"	"EPP"	"EPP"
11	http://www.sefaz.ma.gov.br/Cadastr 12376404/	http://www.sefaz.ma.gov.br/RFB/res MATEUS_SUPERMERCADOS	"03995515004405"	"EPP"	"EPP"

Consulta 3 – Quais Estabelecimentos estão com situações cadastrais diferentes com relação às fontes da RFB e de Cadastro da SEFAZ?

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX sefazma: <http://www.sefaz.ma.gov.br/ontology/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT DISTINCT ?cnpj_estabelecimento ?situacao_cadastral ?
   situacao_cadastralSameAs WHERE {

  ?estabelecimento a sefazma:Establishment.

```

```
  ?estabelecimento owl:sameAs ?estabelecimentoSameAs;
     sefazma:cnpj ?cnpj_estabelecimento;
     sefazma:has_cadastral_situation ?situacao_cadastral.

  ?estabelecimentoSameAs sefazma:cnpj ?
     cnpj_estabelecimento2;
      sefazma:has_cadastral_situation ?
         situacao_cadastralSameAs.

  FILTER ( (!CONTAINS(STR(?situacao_cadastralSameAs),
     SUBSTR(STR(?situacao_cadastral), 77, 4))))
}
```

Figura 25 – Resultados da Consulta para a Consulta 3.

	cnpj_estabelecimento	situacao_cadastral	situacao_cadastralSameAs
1	"03995515015601"	http://www.sefaz.ma.gov.br/Cadastral_Situation/resource/Ca 3995515015601-21%2F01%2F20/	http://www.sefaz.ma.gov.br/RFB/resource/Cadastral_Situatio 03995515015601-2020_01_21

Consulta 4 – Quais Estabelecimentos de uma Empresa que não estão situados no Maranhão?

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX sefazma: <http://www.sefaz.ma.gov.br/ontology/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT DISTINCT ?cnpj ?codigo_estado {
        ?organizacao a foaf:Organization;
            sefazma:has_establishment ?estabelecimento.
  ?estabelecimento sefazma:cnpj ?cnpj;
            sefazma:has_address ?endereco.
    ?endereco sefazma:has_federative_unit ?estado.
    ?estado sefazma:abbreviation_UF ?codigo_estado.
}
```

Figura 26 – Resultados da Consulta para a Consulta 4.

	cnpj	codigo_estado
1	"03995515004073"	"TO"
2	"03995515004820"	"TO"
3	"03995515012092"	"PA"
4	"03995515005479"	"PA"
5	"03995515005398"	"PA"
6	"03995515008656"	"PA"
7	"03995515015512"	"PA"
8	"03995515005207"	"PA"
9	"03995515012173"	"PA"
10	"03995515008737"	"PA"
11	"03995515012840"	"PA"
12	"03995515008907"	"PA"
13	"03995515009113"	"PA"
14	"03995515014893"	"PA"
15	"03995515013307"	"PA"
16	"03995515009466"	"PA"
17	"03995515014540"	"PA"

Consulta 5 – Quais dados de endereços de um mesmo estabelecimento nas bases da RFB e Cadastro são divergentes?

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX sefazma: <http://www.sefaz.ma.gov.br/ontology/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>

PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT DISTINCT ?endereco1 ?descricao {
   ?organizacao a foaf:Organization;
            sefazma:has_establishment ?estabelecimento.
   ?estabelecimento owl:sameAs ?estabelecimento2;
             sefazma:has_address ?endereco1.
    ?endereco1 rdfs:label ?descricao.
  OPTIONAL { ?endereco2 owl:sameAs ?endereco1
  FILTER (?endereco1 != ?endereco2)
  }
  FILTER(!BOUND(?endereco2))
}
```

Figura 27 – Resultados da Consulta para a Consulta 5.

	endereco1	descricao
1	http://www.sefaz.ma.gov.br/Cadastral_Situation/resource/Address/AVE-NEWTON%20BELO-541-VAZIO-EM%20FRENTE%20A%20CONSTRUEL%20RIBEIRO-CENTRO-SANTA%20LUZIA/	"AVE-NEWTON BELO-541--EM FRENTE A CONSTRUEL RIBEIRO-CENTRO"
2	http://www.sefaz.ma.gov.br/Cadastral_Situation/resource/Address/AVE-ALEXANDRE%20COSTA-1-LETRA%20A%20%20%20%20%20%20%20%20%20%20%20%20%20%20%20%20%20%20%20%20(SHOPPING%20ACAILANDIA-RESIDENCIAL%20TROPICAL%20PRANCHA%2002-COCALZINHO%20DE%20GOIAS/	"AVE-ALEXANDRE COSTA-1-LETRA A QUADRA56 LOTE 1 A-SHOPPING A CAILANDIA-RESIDENCIAL TROPICAL PRANCHA 02"
3	http://www.sefaz.ma.gov.br/Cadastral_Situation/resource/Address/AVE-DORGIVAL%20PINHEIRO%20DE%20SOUSA%201278-1400-%3A%20SHOPPING%20IMPERATRIZ%3B-SHOPPING%20IMPERATRIZ-CENTRO-IMPERATRIZ/	"AVE-DORGIVAL PINHEIRO DE SOUSA 1278-1400-: SHOPPING IMPERATRIZ;-SHOPPING IMPE RATRIZ-CENTRO"
4	http://www.sefaz.ma.gov.br/Cadastral_Situation/resource/Address/RUA-PROFESSOR%20ANTONIO%20OLIVIO%20RODRIGUES-1-VAZIO-PROXIMO%20A%20RODOVIARIA-PICARRA-UBERABA/	"RUA-PROFESSOR ANTONIO OLIVIO RODRIGUES-1--PROXIMO A RODOVIARIA-PICARRA"
5	http://www.sefaz.ma.gov.br/Cadastral_Situation/resource/Address/AVE-CASTELO%20BRANCO-2790-VAZIO-VAZIO-LARANJEIRAS-SANTA%20INES/	"AVE-CASTELO BRANCO-2790---LARANJEIRAS"
6	http://www.sefaz.ma.gov.br/Cadastral_Situation/resource/Address/RUA-LEONCIO%20PIRES%20DOURADO%201084-1765-%3A%20A%3B-PROXIMO%20A%20ANTIGA%20COOPERLEITE-BACURI-IMPERATRIZ/	"RUA-LEONCIO PIRES DOURADO 1084-1765-: A;-PROXIMO A ANTIGA COOPERLEITE-BACUR I"
7	http://www.sefaz.ma.gov.br/Cadastral_Situation/resource/Address/AVE-JOAO%20DO%20VALE-1-VAZIO-NA%20ENTRADA%20DA%20CIDADE-SAO%20FRANCISCO-MASSARANDUBA/	"AVE-JOAO DO VALE-1--NA ENTRADA DA CIDADE-SAO FRANCISCO"

Consulta 6 – Quais Sócios de uma Empresa existem na RFB que não estão presentes no Cadastro da SEFAZ e vice-versa?

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX sefazma: <http://www.sefaz.ma.gov.br/ontology/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT DISTINCT ?nome_socio WHERE{
    ?o1 sefazma:has_partnership ?p1.
    ?p1 sefazma:has_partner ?par1.
    ?par1 foaf:name ?nome_socio.

    OPTIONAL {
      ?o1 owl:sameAs ?o2.
      ?p2 sefazma:has_partner ?par2.
    ?par1 owl:sameAs ?par2.
    }
  FILTER(!BOUND(?par2))
}
```

Figura 28 – Resultados da Consulta para a Consulta 6.

	nome_socio
1	"MARIA BARROS PINHEIRO"
2	"PAULO ERLANDIO GERALDO RODRIGUES"
3	"GRUPO MATEUS S A"
4	"TOCANTINS PARTICIPACOES E EMPREENDIMENTOS LTDA"

Consulta 7 – Existem diferentes códigos de município para uma mesma cidade?

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX sefazma: <http://www.sefaz.ma.gov.br/ontology/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT (COUNT(?codigo) AS ?quantidade_codigos) ?cidade ?
   nome_cidade {
  ?cidade a sefazma:City;
          sefazma:city_code ?codigo;
       rdfs:label ?nome_cidade.
  OPTIONAL {?cidade2 owl:sameAs ?cidade.}

} GROUP BY ?cidade ?nome_cidade
HAVING(?quantidade_codigos >1)
```

Figura 29 – Resultados da Consulta para a Consulta 7.

	quantidade_codigos	cidade	nome_cidade
1	"2"^^xsd:integer	http://www.sefaz.ma.gov.br/RFB/resource/City/NOVA_IPIXUN PARA	"NOVA IPIXUNA"
2	"2"^^xsd:integer	http://www.sefaz.ma.gov.br/IBGE/Location/resource/City/NO PARA	"NOVA IPIXUNA"
3	"2"^^xsd:integer	http://www.sefaz.ma.gov.br/RFB/resource/City/ESTREITO- MARANHAO	"ESTREITO"
4	"2"^^xsd:integer	http://www.sefaz.ma.gov.br/IBGE/Location/resource/City/ES MARANHAO	"ESTREITO"
5	"2"^^xsd:integer	http://www.sefaz.ma.gov.br/RFB/resource/City/ACAILANDIA- MARANHAO	"ACAILANDIA"
6	"2"^^xsd:integer	http://www.sefaz.ma.gov.br/IBGE/Location/resource/City/AC MARANHAO	"ACAILANDIA"
7	"2"^^xsd:integer	http://www.sefaz.ma.gov.br/RFB/resource/City/BARRA_DO_ MARANHAO	"BARRA DO CORDA"

Na figura 30 é apresentada uma visão geral de todo o EKG construído com base na

Tabela 6 – Avaliação do EKG-SEFAZMA com base na checklist.

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16	R17	R18	R19
Sat.	S	S	S	S	S	S	S	S	S	S	S	N	S	S	N	S	S	S	N
Vt	2	1	2	1	1	1	0.25	0.25	1	1	1	0	0.25	1	0	0.5	0.25	3	0

Sat. = Satisfaz requisito?;
S = Sim;
N = Não;
V_t = Valor Total.

utilização da EKGO sob o estudo de caso.

Figura 30 – Visão Geral da Camada Semântica do EKG construído com base na EKGO.

Considerando as consultas realizadas, o EKG retornou resultados para todas as consultas, demonstrando cobertura do domínio na representação dos tipos, propriedades e relações, expressando sua capacidade de realização de tarefas.

Posteriormente é feita a avaliação do EKG através de uma *checklist* de qualidade adaptada de (FÄRBER *et al.*, 2018). Mais detalhes da avaliação são descritos na Subseção 5.2.4.

Sumarização dos Resultados:

– **Requisito (R1)**: Para avaliar a sintaxe do EKG foram utilizadas as ferramentas *Turtle Validator*[6] e *RDF-Validator*[7], não havendo quaisquer erro de sintaxe quanto ao arquivo do

[6] http://ttl.summerofcode.be/
[7] https://www.w3.org/RDF/Validator/

EKG-SEFAZMA e valores de literais ;

- **Requisito (R2):** EKG-SEFAZMA foi construí com base em SISIFO de forma manual com apoio de ferramentas nas etapas de modelagem da ontologia de domínio (Protégé), escrita de mapeamentos utilizando o R2RML, especificação dos links através do SILK e realização de consultas através do Ontop e Teiid;
- **Requisito (R3):** Como forma de viabilizar a proveniência, EKG-SEFAZMA possui metadados de proveniência em nível de recurso para cada Fonte de Dados através da relação *Object Property prov:wasDerivedFrom* com uma instância de *ekgo:ProvenienciaRFB* e *ekgo:ProvenienciaSEFAZ*;
- **Requisito (R4):** Para satisfazer esse requisito são consideradas restrições da EKGO, e.g (hasDataSource only ;
- **Requisito (R5):** A ontologia utilizada para descrever EKG-SEFAZMA (EKGO) fornece restrições de disjunção entre tipos e.g *ekgo:Exported_View owl:disjointWith ekgo:Links, ekgo:LinksetView , ekgo:Mappings , ekgo:Materialized , ekgo:Rules, ekgo:Virtual* e outros. Para relações (propriedades) um exemplo de restrição que cobre este requisito é expresso por: *ekgo:hasExportedView owl:propertyDisjointWith ekgo:hasDomainOntology, ekgo:hasLinksetView e outros.*
- **Requisito (R6):** Todas as instâncias recebem um valor para rdfs:label com base nos mapeamentos R2RML a partir das fontes, e.g

```
http://www.sefaz.ma.gov.br/Cadastral_Situation/resource/
Address/AVE-GETULIO\%20VARGAS-1205-VAZIO-ANTIGA%20ROMA%20
DECORE-CENTRO-IMPERATRIZ/\> rdfs:label \"RUA AFONSO PENA,
nº 537, CENTRO CODO - MA 65400000")};
```

- **Requisito (R8):** O EKG-SEFAZMA é disponibilizado em Formato Turtle através do Link https://tinyurl.com/24e5busq;
- **Requisito (R9):** O padrão de URI adotado foi: "http://www.sefazma.gov.br/resource/RFB/" para a fonte de dados da RFB, e "http://www.sefazma.gov.br/resource/SEFAZMA-Cadastral/" para a fonte de Cadastro;

- **Requisito (R10):** Dentre os vocabulários utilizados, estão o **foaf**[8], **skos**[9], **prov**[10] e o **dc**[11];
- **Requisito (R11):** O EKG foi disponibilizado em múltiplas representações como em Turtle[12] e Nquad[13];
- **Requisito (R13):** EKG-SEFAZMA não possui nenhuma ocorrência de *Blank nodes* e *reification* não estão presentes;
- **Requisito (R14):** Para acesso disponível, um *Endpoint* público foi disponibilizado no seguinte link: https://tinyurl.com/yb9lcho6;
- **Requisito (R16):** Através da representação de EKGO são fornecidos diversos metadados ao EKG, e.g *(rdfs:label, dc:description, dc:issued, dc:modified, dc:format e outros)*;
- **Requisito (R17):** Os dados são publicados com base na licença *CC BY-NC-ND 3.0 PT* através de relações do tipo: *(:X dc:license "http://rdflicense.appspot.com/rdflicense/cc-by-nc-nd3.0pt")* ;
- **Requisito (R18):** O EKG-SEFAZMA possui links através do *owl:sameAs*, ocorrendo: $:X\ owl:sameAs:Y$, expressando que no mínimo 1 caso de *sameAs* foi identificado, satisfazendo o requisito proposto por (ZAVERI *et al.*, 2016);

Considerando a fórmula para definição do índice de qualidade do EKG, $i(k)$ apresentou valor super

Ao final do estudo de caso, os resultados sugerem quem SISIFO demonstrou conseguir contemplar as fases do processo de construção de um EKG com base em uma especificação da EKGO. Outros achados denotam que a experiência realizada evitou a construção de um EKG sem qualidade, argumento justificável pelo conjunto de passos sequenciais e bem definidos propostos por SISIFO juntamente com seu suporte semântico através do conjunto de especificações da Camada Semântica.

6.2 Avaliação da EKGO

Como forma de validar a EKGO utilizando questões já definidas, as questões utilizadas para validar a RDBSO (*Relational Database Ontology*) proposta em Aguiar *et al.* (2018) para bancos de dados relacionais foram adaptadas três questões de competência para EKGs. Na

[8] http://xmlns.com/foaf/spec/
[9] https://www.w3.org/2009/08/skos-reference/skos.html
[10] https://www.w3.org/TR/prov-o/
[11] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[12] https://tinyurl.com/24e5busq
[13]

Tabela 7 – Questões de Competência adaptadas para a EKGO

Identificador	Questão de Competência	Resultado

(QC1)	SELECT DISTINCT ?instance WHERE {
	?instance a drm:DataAsset.
	}
	ORDER BY ?instance
		FonteCadastro, FonteRFB

(QC2)	SELECT DISTINCT ?instance WHERE {
	
	?instance a ekgo:ExportedViewPublication;
	ekgo:hasPublicationDate ?date.
	FILTER (xsd:dateTime(?date) = xsd:dateTime("2020-01-01"))
	}
	ORDER BY ?instance
		VisaoExportadaCadastro e VisaoExportadaRFB

(QC3)	SELECT DISTINCT ?instance WHERE {
	?instance a ekgo:hasLinksetView;
	ekgo:hasExportedView ?ev.
	?ev ekgo:hasDataSource ?source.
	FILTER regex(?source, "RFB")
	}
	ORDER BY ?source
		VisaoLinksetCadastroRFB

Tabela 7 são dispostas essas questões e os respectivos resultados da EKGO.

Com base na avaliação baseada nas questões de competência adaptadas, a EKGO demonstrou ser capaz de responder questões já formuladas anteriormente em outros estudos adaptando-se para o contexto de EKGs.

Para complementar a avaliação, após realizar buscas em repositórios de *datasets* e ontologias, foi encontrado um vocabulário para descrever grafos, o RDFP *The RDF Presentation ontology*[14]. Com base nisso buscou-se identificar a capacidade de satisfatibilidade do vocabulário EKGO frente ao RDFP baseando-se em uma avaliação *gold standard* (PORZEL; MALAKA, 2004). Para isso foram utilizados aspectos de cobertura para analisar a capacidade de respostas de cada uma das ontologias, conforme apresentado na Tabela 8.

Considerando os critérios na Tabela 8 EKGO apresentou uma maior cobertura no que tange aos principais elementos presentes em um EKG, dessa forma EKGO demonstra ser válida para especificar EKGs e considerando-se também como uma proposta original.

6.2.1 Avaliação Sintática

A consistência e qualidade de uma ontologia podem ser afetadas pelos obstáculos apresentados durante o processo de modelagem; portanto, é essencial evitar falhas indesejáveis (POVEDA-VILLALÓN *et al.*, 2012). No que diz respeito à Avaliação Sintática da ontologia implementada, a ferramenta *OntOlogy Pitfall Scanner!* (*OOPS!*) (POVEDA-VILLALÓN *et al.*, 2014) foi adotada para identificar possíveis problemas de sintaxe. Essa técnica permite

[14] https://lov.linkeddata.es/dataset/lov/vocabs/rdfp

Tabela 8 – Tabela de Análise da Cobertura dos Conceitos nas ontologias.

Identificador	Conceitos Representados	RDFP	EKGO
(RC1)	Fontes de Dados	-	X
(RC2)	Ontologia de Domínio	-	X
(RC3)	Mapeamentos	X	X
(RC4)	Linksets	X	X
(RC5)	Proveniência	-	X
(RC6)	Regras de Validação	X	-
(RC7)	Atributos de Qualidade	-	X
(RC8)	Visões Virtuais	-	X
(RC9)	Visões Materializadas	X	X
(RC10)	Apresentação	X	X
(RC11)	Elementos da Etapa de Publicação KG (i.e TripleStore, Mediador Semântico)	-	X
Média:		45.45	91.66

analisar uma ontologia e verificar sua conformidade com os padrões de modelagem recomendados. OOPS! somente encontrou duas inconsistências na EKG Ontology, sendo ***P21****: Using a miscellaneous class* e ***P08****: Missing Annotations*. Os erros encontrados foram resolvidos através do Protégé Web.

Além disso, também foram utilizadas as ferramentas *OWL Validator* [15] para verificar erros na sintaxe e *OLED Validator* [16] para validar as relações baseadas em UFO como *part-whole*, de indentidade, ciclos hierárquicos e outros. Ambas as ferramentas não relataram nenhum alerta ou erro de sintaxe. Por fim, foi utilizada a ferramenta *online* Vapour (BERRUETA *et al.*, 2008), demonstrando que a ontologia está sintaticamente consistente.

[15] http://visualdataweb.de/validator/
[16] https://github.com/nemo-ufes/ontouml-lightweight-editor

7 CONCLUSÕES E TRABALHOS FUTUROS

A construção de EKGs não é uma tarefa trivial, para isso são necessários conhecimentos acerca do processo de integração semântica bem como das tecnologias. Para isso, foi proposto nesta dissertação a utilização de SISIFO uma abordagem semântica guiada por uma representação formal e ontológica de especificação de um EKG.

Para tanto, foram apresentados sequencialmente os passos propostos de SISIFO, bem como uma validação com base em um estudo de caso no domínio fiscal juntamente com uma avaliação da EKGO a fim de identificar sua capacidade funcional sob o uso de questões de competência.

Os resultados sugerem que SISIFO conseguiu ser utilizada de modo satisfatório para construção de um EKG, fornecendo uma semântica adequada aos dados bem como a EKGO foi capaz de ser utilizada para especificação e formalização demonstrando viabilidade para uso conjunto de SISIFO.

As limitações da abordagem podem ser apoiadas através de um ambiente contendo interface e elementos gráficos para viabilizar um melhor gerenciamento do EKG, além de necessitar de outras validações voltadas a outros cenários e contextos . No estudo, houve uma limitação quanto ao tipo de *link* utilizado, sendo somente o *owl:sameAs*, Como trabalhos futuros pretende-se expandir a abordagem e implementar um Ambiente Semântico para suportar todos os passos de SISIFO. Ainda, almeja-se que esse sistema alinhe-se como um sistema de recomendação para auxiliar o usuário duante a construção do EKG.

REFERÊNCIAS

ACOSTA, Maribel et al. Crowdsourcing linked data quality assessment. In: **Anais da International semantic web conference**. Springer, Berlin, Heidelberg, 2013. p. 260-276.

DE AGUIAR, Camila Zacché; DE ALMEIDA FALBO, Ricardo; SOUZA, Vítor E. Silva. Ontological Representation of Relational Databases. In: **Anais do ONTOBRAS**. São Paulo, Brasil. 2018. p. 140-151.

ALOBAID, Ahmad et al. OnToology, a tool for collaborative development of ontologies. In: **Anais do ICBO**. [*S. l.*: *s. n.*], 2015.

ANAM, Sarawat et al. Adapting a knowledge-based schema matching system for ontology mapping. In: **Anais da Australasian Computer Science Week Multiconference**. 2016. p. 1-10.

ARRUDA, Narciso et al. Publishing and consuming semantic views for construction of knowledge graphs. In: **Anais da 22nd International Conference on Enterprise Information Systems (ICEIS)**. SCITEPRESS Digital Library, 2020. p. 197-204.

BA, Mouhamadou; DIALLO, Gayo. ServOMap and ServOMap: lt results for OAEI 2012. In: **Anais da 7th International Conference on Ontology Matching-Volume 946**. Boston, EUA. 2012. p. 197-204.

BĄK, Jarosław; BLINKIEWICZ, Michał; ŁAWRYNOWICZ, Agnieszka. User-friendly visual creation of R2RML mappings in SQuaRE. In: **Anais da Third International Workshop on Visualization and Interaction for Ontologies and Linked Data co-located with the 16th International Semantic Web Conference ISWC 2017.** Vienna, Austria. 2017. p. 139-150.

BERNERS-LEE, Tim. Linked data-design issues. Disponível em: **http://www. w3. org/DesignIssues/LinkedData.html**. Acesso em: 10 dez. 2019.

BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora. The semantic web. **Scientific american**, v. 284, n. 5, p. 34-43, 2001.

BERRUETA, Diego; FERNÁNDEZ, Sergio; FRADE, Iván. Cooking HTTP content negotiation with Vapour. In: **Anais do 4th Workshop on Scripting for the Semantic Web (SFSW2008)**. Tenerife, Espanha: Springer, 2008. v. 72, p. 1–6.

BIKAKIS, Nikos et al. The SPARQL2XQuery interoperability framework. **World Wide Web**, v. 18, n. 2, p. 403-490, 2015.

BIZER, Christian; HEATH, Tom; BERNERS-LEE, Tim. Linked data: The story so far. In: **Anais do emantic services, interoperability and web applications: emerging concepts**. IGI global, 2011. p. 205-227.

BONATTI, Piero Andrea et al. V. Knowledge graphs: new directions for knowledge representation on the semantic web (dagstuhl seminar 18371). In: **Anais do SCHLOSS DAGSTUHL-LEIBNIZ-ZENTRUM FUER INFORMATIK**. [*S. l.*], 2019.

CALVANESE, Diego et al. Ontop: Answering SPARQL queries over relational databases. **Semantic Web**, v. 8, n. 3, p. 471-487, 2017.

CALVANESE, Diego et al. The MASTRO system for ontology-based data access. **Semantic Web**, v. 2, n. 1, p. 43-53, 2011.

CASANOVA, Marco A. et al. On materialized sameAs linksets. In: **International Conference on Database and Expert Systems Applications**. Springer, Cham, 2014. p. 377-384.

ČERĀNS, Kārlis; BŪMANS, Guntars. Rdb2owl: a rdb-to-rdf/owl mapping specification language. In: **Databases and Information Systems VI**. IOS Press, 2011. p. 139-152.

CHEATHAM, Michelle. MapSSS results for OAEI 2011. In: **Anais do ISWC 2011 workshop on ontology matching**. 2011. p. 184-190.

CHORTARAS, Alexandros; STAMOU, Giorgos. D2RML: Integrating Heterogeneous Data and Web Services into Custom RDF Graphs. In: **LDOW@ WWW**. 2018.

CRUZ, Isabel F.; ANTONELLI, Flavio Palandri; STROE, Cosmin. AgreementMaker: efficient matching for large real-world schemas and ontologies. In: **Anais do of the VLDB Endowment**, v. 2, n. 2, p. 1586-1589, 2009.

DA CRUZ, Matheus Mayron Lima et al. 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**. [*S. l.*], 2019. p. 13–18.

DAS, S. **R2rml:** Rdb to rdf mapping language. Disponível em: http://www. w3. org/TR/r2rml/. Acesso em: 16 dez. 2019.

DEBATTISTA, Jeremy; AUER, Sören; LANGE, Christoph. Luzzu—a methodology and framework for linked data quality assessment. **Journal of Data and Information Quality (JDIQ)**, v. 8, n. 1, p. 1-32, 2016.

DEBRUYNE, Christophe; TRAN, Trung-Kien; MEERSMAN, Robert. Grounding ontologies with social processes and natural language. **Journal on Data Semantics**, v. 2, n. 2, p. 89-118, 2013.

DIMOU, Anastasia et al. RML: a generic language for integrated RDF mappings of heterogeneous data. In: **Anais do Ldow**. 2014.

DUAN, Rong; XIAO, Yanghua. Enterprise Knowledge Graph From Specific Business Task to Enterprise Knowledge Management. In: **Anais do 28th ACM International Conference on Information and Knowledge Management**. 2019. p. 2965-2966.

FÄRBER, Michael et al. Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago. **Semantic Web**, v. 9, n. 1, p. 77-129, 2018.

FENSEL, Dieter et al. **Knowledge Graphs: Methodology, Tools andSelected Use Cases**. [*S. l.*]: Springer Nature, 2020.

FERNÁNDEZ-LÓPEZ, Mariano; GÓMEZ-PÉREZ, Asunción; JURISTO, Natalia. Methontology: from ontological art towards ontological engineering. 1997.

FORBES. **Is The Enterprise Knowledge Graph Finally Going To Make All Data Usable?** 2018. Disponível em: https://www.forbes.com/sites/danwoods/2018/09/19/is-the-enterprise-knowledge-graph-going-to-finally-make-all-data-usable/#2d7048db7d39. Acesso em: 12 jan. 2020.

FRAUNHOFER. **Enterprise Knowledge Graphs**. 2020. Disponível em: https://www.iais.fraunhofer.de/en/ business-areas/enterprise-information-integration/enterprise-knowledge-graphs.html. Acesso em: 13 jan. 2020.

FÜRBER, C.; HEPP, M. Swiqa–a semantic web information quality assessment framework. 2011.

GALKIN, M.; AUER, S.; SCERRI, S. Enterprise knowledge graphs: A survey. In: **Anais da 37th International Conference on Information Systems**. [*S. l.*: *s. n.*], 2016.

GALKIN, Mikhail et al. Enterprise knowledge graphs: A Semantic Approach for Knowledge Management in the Next Generation of Enterprise Information Systems. In: **Anais do ICEIS (2)**. [*S. l.*: *s. n.*], 2017. p. 88–98.

GIUNCHIGLIA, Fausto et al. A facet-based methodology for the construction of a large-scale geospatial ontology. **Journal on data semantics**, v. 1, n. 1, p. 57-73, 2012.

GOMEZ-PEREZ, Jose Manuel et al. Enterprise knowledge graph: An introduction. In: **Anais Exploiting linked data and knowledge graphs in large organisations**. Springer, Cham, 2017. p. 1-14.

GRUBER, Tom. It is what it does: The pragmatics of ontology for knowledge sharing. In: **Anais do International CIDOC CRM Symposium, Available online at.** [*S. l.*: *s. n.*], 2003.

GRÜNINGER, Michael; FOX, Mark S. The role of competency questions in enterprise engineering. In: **Benchmarking—Theory and practice**. Springer, Boston, MA, 1995. p. 22-31.

GUIZZARDI, G. Ontological foundations for structural conceptual models. Tese (Doutorado) – University of Twente, Enschede, 2005.

GUIZZARDI, Giancarlo et al. Towards ontological foundations for the conceptual modeling of events. In: **International Conference on Conceptual Modeling**. Springer, Berlin, Heidelberg, 2013. p. 327-341.

GUPTA, Shubham et al. Karma: A system for mapping structured sources into the Semantic Web. In: **Anais da Extended Semantic Web Conference**. Springer, Berlin, Heidelberg, 2012. p. 430-434, 2012. p. 430–434.

HAIDER, Noori et al. CSV2RDF: Generating RDF data from CSV file using semantic web technologies. **Journal of Theoretical and Applied Information Technology**, v. 96, n. 20, p. 6889-6902, 2018.

HAM, K. et al. Free open-source tool for cleaning and transforming data. **Journal of the Medical Library Association: JMLA**, v. 101, p. 233-234, 2013.

BAUNGARD HANSEN, J. et al. Validata: an online tool for testing RDF data conformance. In: **Anais da 8th semantic web applications and tools for life sciences international conference. Cambridge UK**. 2015. p. 157-166.

HEYVAERT, Pieter et al. RMLEditor: a graph-based mapping editor for linked data mappings. In: **Anais da European Semantic Web Conference**. Springer, Cham, 2016. p. 709-723.

HUBER, Jakob et al. CODI: Combinatorial optimization for data integration–results for OAEI 2011. **Anais do Ontology Matching**, v. 134, 2011.

ICSC. **ICSC 2014 16/6/2014 Enterprise “Knowledge Graphs“ when "Web of Data"technologies make a lot of sense in business scenarios. Dr. Giovanni Tummarello.** 2014. Disponível em: https://slideplayer.com/slide/3727468/. Acesso em: 01 abr. 2020.

JETSCHNI, Jonas; MEISTER, Vera G. Schema engineering for enterprise knowledge graphs: A reflecting survey and case study. In: **2017 eighth international conference on intelligent computing and information systems (icicis)**. [*S. l.*], IEEE, 2017. p. 271-277.

JIA, Junzhi From data to knowledge: the relationships between vocabularies, linked data and knowledge graphs. **Journal of Documentation**, Emerald Publishing Limited, 2020.

JIMÉNEZ-RUIZ, Ernesto; CUENCA GRAU, Bernardo. Logmap: Logic-based and scalable ontology matching. In: **International Semantic Web Conference**. Springer, Berlin, Heidelberg, 2011. p. 273-288.

JIMÉNEZ-RUIZ, Ernesto et al. BootOX: Practical mapping of RDBs to OWL 2. In: **International Semantic Web Conference**. Springer, Cham, 2015. p. 113-132.

CROTTI JUNIOR, Ademar; DEBRUYNE, Christophe; O’SULLIVAN, Declan. Juma: An editor that uses a block metaphor to facilitate the creation and editing of R2RML mappings. In: **European Semantic Web Conference**. Springer, Cham, 2017. p. 87-92.

KHARLAMOV, Evgeny et al. Optique: Ontology-based data access platform. Citeseer 2015.

KITCHENHAM, Barbara; CHARTERS, Stuart. Guidelines for performing systematic literature reviews in software engineering. 2007.

KITCHENHAM, Barbara et al. Systematic literature reviews in software engineering–a systematic literature review. **Information and software technology**, v. 51, n. 1, p. 7-15, 2009.

KNAP, Tomáš et al. ODCleanStore: a framework for managing and providing integrated linked data on the web. In: **Anais da International Conference on Web Information Systems Engineering**. Springer, Berlin, Heidelberg, 2012. p. 815-816.

KONTOKOSTAS, Dimitris et al. Test-driven evaluation of linked data quality. In: **Anais da 23rd international conference on World Wide Web**. 2014. p. 747-758.

KROETSCH, M.; WEIKUM, Gerhard. Special issue on knowledge graphs. **Journal of Web Semantics**, v. 37, n. 38, p. 53-54, 2016.

LEFRANÇOIS, Maxime; ZIMMERMANN, Antoine; BAKERALLY, Noorani. A SPARQL extension for generating RDF from heterogeneous formats. In: **Anais da European Semantic Web Conference**. Springer, Cham, 2017. p. 35-50.

LI, Juanzi et al. Rimom: A dynamic multistrategy ontology alignment framework. **IEEE Transactions on Knowledge and data Engineering**, v. 21, n. 8, p. 1218-1232, 2008.

MAMI, Imene; COLETTA, Remi; BELLAHSENE, Zohra. Modeling view selection as a constraint satisfaction problem. In: **Anais da International Conference on Database and Expert Systems Applications**. Springer, Berlin, Heidelberg, 2011. p. 396-410.

MEDEIROS, Luciano Frontino de; PRIYATNA, Freddy; CORCHO, Oscar. MIRROR: Automatic R2RML mapping generation from relational databases. In: **Anais da International Conference on Web Engineering**. Springer, Cham, 2015. p. 326-343.

MENDES, Pablo N.; MÜHLEISEN, Hannes; BIZER, Christian. Sieve: linked data quality assessment and fusion. In: **Anais do EDBT/ICDT workshops 2012**. 2012. p. 116-123.

MICHEL, Franck et al. **xR2RML: Relational and non-relational databases to RDF mapping language**. 2017. Tese de Doutorado. CNRS.
NGO, D.; BELLAHSENE, Z. Yam++: a multi-strategy based approach for ontology matchingtask. In: SPRINGER. **Anais da International Conference on Knowledge Engineering and KnowledgeManagement**. [*S. l.*], 2012. p. 421–425.

PÉREZ, Jorge; ARENAS, Marcelo; GUTIERREZ, Claudio. Semantics and complexity of SPARQL. **ACM Transactions on Database Systems (TODS)**, v. 34, n. 3, p. 1-45, 2009.

PEUKERT, Eric; EBERIUS, Julian; RAHM, Erhard. AMC-A framework for modelling and comparing matching systems as matching processes. In: **Anais da IEEE 27th International Conference on Data Engineering**. IEEE, 2011. p. 1304-1307.

PINKEL, Christoph et al. IncMap: pay as you go matching of relational schemata to OWL ontologies. In: **Anais do OM**. [*S. l.*], 2013. p. 37-48.. 2013. p. 37–48.

PIPINO, Leo L.; LEE, Yang W.; WANG, Richard Y. Data quality assessment. **Anais da Communications of the ACM**, v. 45, n. 4, p. 211-218, 2002.

POGGI, Antonella; RODRIGUEZ, Mariano; RUZZI, Marco. Ontology-based database access with DIG-Mastro and the OBDA Plugin for Protégé. In: **Anais do 4th Int. Workshop on OWL: Experiences and Directions (OWLED 2008 DC)**. [*S. l.*], 2008.

POOLPARTY. **The Enterprise Knowledge Graph - A Definition.** 2019. Disponível em: https://help.poolparty.biz/pp72/white-papers-release-notes/poolparty-technical-white- paper/an-enterprise-knowledge-graph-life-cycle-a-summary/the-enterprise-knowledge-graph- a-definition. Acesso em: 13 jan. 2020.

PORZEL, Robert; MALAKA, Rainer. A task-based approach for ontology evaluation. In: **Anais do ECAI Workshop on Ontology Learning and Population, Valencia, Spain**. Valencia, Spain: Citeseer, 2004. p. 1-6.

POVEDA-VILLALÓN, María; GÓMEZ-PÉREZ, Asunción; SUÁREZ-FIGUEROA, Mari Carmen. Oops!(ontology pitfall scanner!): An on-line tool for ontology evaluation. **International Journal on Semantic Web and Information Systems (IJSWIS)**, v. 10, n. 2, p. 7-34, 2014.

POVEDA-VILLALÓN, María; SUÁREZ-FIGUEROA, Mari Carmen; GÓMEZ-PÉREZ, Asunción. Validating ontologies with oops!. In: **Anais do International conference on knowledge engineering and knowledge management**. Springer, Berlin, Heidelberg, 2012. p. 267-281.

QI, Guilin et al. (Ed.). **Linked Data and Knowledge Graph: Seventh Chinese Semantic Web Symposium and the Second Chinese Web Science Conference, CSWS 2013, Shanghai, China, August 12-16, 2013. Revised Selected Papers**. Springer, 2013.

QIAO, Bo et al. Building thesaurus-based knowledge graph based on schema layer. **Cluster Computing**, v. 20, n. 1, p. 81-91, 2017.

REN, Yuan et al. Towards competency question-driven ontology authoring. In: **European Semantic Web Conference**. Springer, Cham, 2014. p. 752-767.

SCHULTZ, Andreas. et al. Ldif a framework for large-scale linked data integration. In: **Anais do 21st International World Wide WebConference (WWW 2012).** Lyon, France. 2012.

SEQUEDA, Juan F.; MIRANKER, Daniel P. A pay-as-you-go methodology for ontology-based data access. **IEEE Internet Computing**, v. 21, n. 2, p. 92-96, 2017.

SHADBOLT, Nigel; BERNERS-LEE, Tim; HALL, Wendy. The semantic web revisited. **IEEE intelligent systems**, v. 21, n. 3, p. 96-101, 2006.

SHEARER, Robert DC; MOTIK, Boris; HORROCKS, Ian. Hermit: A highly-efficient OWL reasoner. In: **Anais do Owled**. [*S. l.*: *s. n.*], 2008. p. 91.

SICILIA, Álvaro. A. Map-on: A web-based editor for visual ontologymapping. **Semantic Web**, IOS Press, v. 8, n. 6, p. 969–980, 2017.

SINGHAL, A. **Introducing the Knowledge Graph: things, not strings**. 2012. Disponível em: https://blog.google/products/search/introducing-knowledge-graph-things-not/. Acesso em: 04 jan. 2020

SIRIN, Evren et al. Pellet: A practical owl-dl reasoner. **Journal of Web Semantics**, v. 5, n. 2, p. 51-53, 2007.

SLEPICKA, Jason et al. KR2RML: An Alternative Interpretation of R2RML for Heterogenous Sources. In: **Anais do Cold.** [*S. l.*: *s. n.*], 2015.

STAAB, Steffen; STUDER, Rudi (Ed.). **Handbook on ontologies**. [*S. l.*]: Springer Science & Business Media, 2010.

STUDER, Rudi; BENJAMINS, V. Richard; FENSEL, Dieter. Knowledge engineering: principles and methods. **Data & knowledge engineering**, v. 25, n. 1-2, p. 161-197, 1998.

SUÁREZ-FIGUEROA, Mari Carmen. **NeOn Methodology for building ontology networks: specification, scheduling and reuse**. 2010. Tese de Doutorado. Informatica.

SUÁREZ-FIGUEROA, Mari Carmen; GÓMEZ-PÉREZ, Asunción; FERNÁNDEZ-LÓPEZ, Mariano. The NeOn methodology for ontology engineering. In: **Ontology engineering in a networked world**. Springer, Berlin, Heidelberg, 2012. p. 9-34.

SURE, York; STAAB, Steffen; STUDER, Rudi. On-to-knowledge methodology (OTKM). In: Anais do **Handbook on ontologies**. Springer, Berlin, Heidelberg, 2004. p. 117-132.

TUDORACHE, Tania et al. Supporting collaborative ontology development in Protégé. In: **Anais da International Semantic Web Conference**. Springer, Berlin, Heidelberg, 2008. p. 17-32.

USCHOLD, Mike; GRUNINGER, Michael. Ontologies: Principles, methods and applications. **The knowledge engineering review**, Cambridge University Press, v. 11, n. 2, p. 93–136, 1996.

VAVLIAKIS, Konstantinos N.; GROLLIOS, Theofanis K.; MITKAS, Pericles A. RDOTE–publishing relational databases into the semantic web. **Journal of Systems and Software**, v. 86, n. 1, p. 89-99, 2013.

VIDAL, Vânia MP et al. Specification and incremental maintenance of linked data mashup views. In: **International Conference on Advanced Information Systems Engineering**. Springer, Cham, 2015. p. 214-229.

VIDAL, Vânia MP et al. Using changesets for incremental maintenance of linkset views. In: **International Conference on Web Information Systems Engineering**. Springer, Cham, 2016. p. 196-204.

VILLAZÓN-TERRAZAS, Boris; HIDALGO-DELGADO, Yusniel (Ed.). **Knowledge Graphs and Semantic Web: First Iberoamerican Conference, KGSWC 2019, Villa Clara, Cuba, June 23-30, 2019, Proceedings**. Springer, 2019.

W3C. **Resource Description Framework (RDF): Concepts And Abstract Syntax.** 2004. Disponível em: https://www.w3.org/TR/rdf-concepts/. Acesso em: 20 nov. 2019.

W3C. **SPARQL Query Language for RDF - Formal Definitions.** 2006. Disponível em: https://www.w3.org/2001/sw/DataAccess/rq23/sparql-defns.html. Acesso em: 12 mar. 2019.

XIAO, Guohui et al. Virtual knowledge graphs: An overview of systems and use cases. **Data Intelligence**, v. 1, n. 3, p. 201-223, 2019.

YU, L. **A developer's guide to the semantic Web**. [*S. l.*]: Springer Science & Business Media,2011.

ZAVERI, Amrapali et al. User-driven quality evaluation of dbpedia. In: **Anais da 9th International Conference on Semantic Systems**. [*S. l.*: *s. n.*], 2013. p. 97–104.

ZAVERI, Amrapali et al. Quality assessment for linked data: A survey. **Semantic Web**, v. 7, n. 1, p. 63-93, 2016.