

Uso das Tecnologias da Web Semântica na Construção de Grafos de Conhecimento Semântico baseado no Enfoque Híbrido

Tulio Vidal Rolim¹, Caio Viktor S. Avila¹,
Roberval Gomes Mariano², Tainan Calixto², Pedro Ivo²,
José M. M. Filho¹, Ângelo R. Brayner, Vânia Maria Ponte Vidal¹

¹Universidade Federal do Ceará (UFC)
Fortaleza, CE, Brasil.

²Secretaria da Fazenda do Maranhão
São Luís, MA, Brasil

{tulio.xcrtf, arlaass, vaniap.vidal}@gmail.com

{mariano, tainan.calixto}@sefaz.ma.gov.br

{monteiro, brayner}@dc.ufc.br

Abstract. Semantic Knowledge Graph (SKG) is a new paradigm based on semantic web technologies and knowledge graph to integrate heterogeneous data sources. The main objective of a SKG is to provide an ontological and unified view, so that applications can have integrated access to source data through the Semantic View. This article presents an incremental process for building Semantic Knowledge Graphs based on a hybrid approach. In this approach, the RDF views exported by data sources to the semantic layer can be virtual or materialized. The proposed approach makes use of ontologies and the RDF model, in order to guarantee the easy integration of multiple heterogeneous data sources, and to provide a formal semantic representation to allow inference and machine processing. The article describes the construction of a large knowledge graph, GC_SEFAZMA, which semantically integrates internal and external data sources relevant to the Secretariat of Finance of Maranhão.

Resumo. Grafos de Conhecimento Semântico (GCS) é um novo paradigma baseado nas tecnologias da web semântica e grafo de conhecimento para integrar fontes de dados heterogêneas. O principal objetivo de um GCS é fornecer uma visão ontológica e unificada, para que aplicações possam ter acesso integrado aos dados das fontes através da Visão Semântica. Esse artigo apresenta um processo incremental para construção de Grafos de Conhecimento Semânticos baseado em um enfoque híbrido. Nesse enfoque, as visões RDF exportadas pelas fontes de dados para a camada semântica podem ser virtuais ou materializadas. O enfoque proposto faz uso de ontologias e o modelo RDF, com o objetivo de garantir a fácil integração de múltiplas fontes heterogêneas de dados, e prover uma representação semântica formal para permitir inferência e processamento de máquina. O artigo descreve a construção de um grande grafo de conhecimento, GC_SEFAZMA, o qual integra semanticamente as fontes de dados internas e externas relevantes para a Secretaria da Fazenda do Maranhão.



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

1. Introdução

Grafos de Conhecimento Semântico (GCS) é um novo paradigma que está sendo usado para consolidar e integrar semanticamente um grande número de dados advindos de fontes de dados heterogêneas. Nos últimos anos, as tecnologias de grafo de conhecimento estabeleceram uma posição sólida no mundo corporativo, servindo como um elemento central na infraestrutura de gerenciamento de dados organizacionais [Grainger et al. 2016].

Um elemento chave de um GCS é a ontologia de domínio que permite combinar e enriquecer informações armazenadas em fontes de dados em uma visão unificada, que representa não apenas como os dados são organizados, mas também qual é o seu significado pretendido. Além de facilitar a integração de fontes de dados heterogêneas, o uso de ontologias provê uma representação semântica formal, permitindo assim, inferência e processamento de máquina.

Entretanto, a construção de um GCS não é uma tarefa trivial, sendo um processo complexo com muitos desafios, tais como, heterogeneidade de vocabulários e formatos de armazenamento, a descoberta de *links* entre recursos em diferentes fontes de dados, além da resolução de inconsistências e conflitos para melhorar a qualidade dos dados.

Para explorar as capacidades das tecnologias do GCS ao máximo, é necessário um suporte de métodos e tecnologias semânticas para construção, consulta, e uso de aplicações inteligentes. Nessa perspectiva, esse trabalho propõe um enfoque híbrido para construção de GCSs, onde apenas um subconjunto dos dados do GCS é materializado. No enfoque proposto, o processo de construção do GCS é incremental, e tem 5 passos principais: **I**) Aquisição do Conhecimento; **II**) Modelagem da ontologia de domínio ou modelo semântico; **III**) Construção dos grafos de conhecimentos das fontes de dados locais; **IV**) Criação de ligações entre instâncias dos grafos de conhecimento locais; **V**) Inferência e limpeza dos dados. No enfoque híbrido, os grafos de conhecimento locais exportados pelas fontes de dados para a camada semântica podem ser virtuais ou materializados.

Como estudo de caso, discutiremos a construção do grafo de conhecimento corporativo da Secretaria de Fazenda do Maranhão (SEFAZ-MA). As principais contribuições deste artigo são:

- Arquitetura de 4 camadas para GCS baseada no Enfoque Híbrido;
- Processo Incremental para construção da Camada Semântica;
- Ontologia de Domínio para representação e integração de dados de Pessoas Jurídicas;
- Grafo de Conhecimento para integrar fontes de dados internas e externas da SEFAZ-MA.

O restante do artigo está organizado como se segue: A **Seção 2** apresenta a Arquitetura de 4 camadas para construção de GCS. A **Seção 3** discute as camadas do GC_SEFAZMA. Na **Seção 4** são apresentados os trabalhos relacionados. Por fim, na **Seção 5** são apresentadas as conclusões.

2. Arquitetura em 4 Camadas

Nessa seção apresentamos uma Arquitetura em Camadas para a construção de GCS. A Figura 1 apresenta os principais componentes de um GCS agrupados em 4 camadas, descritas a seguir.



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

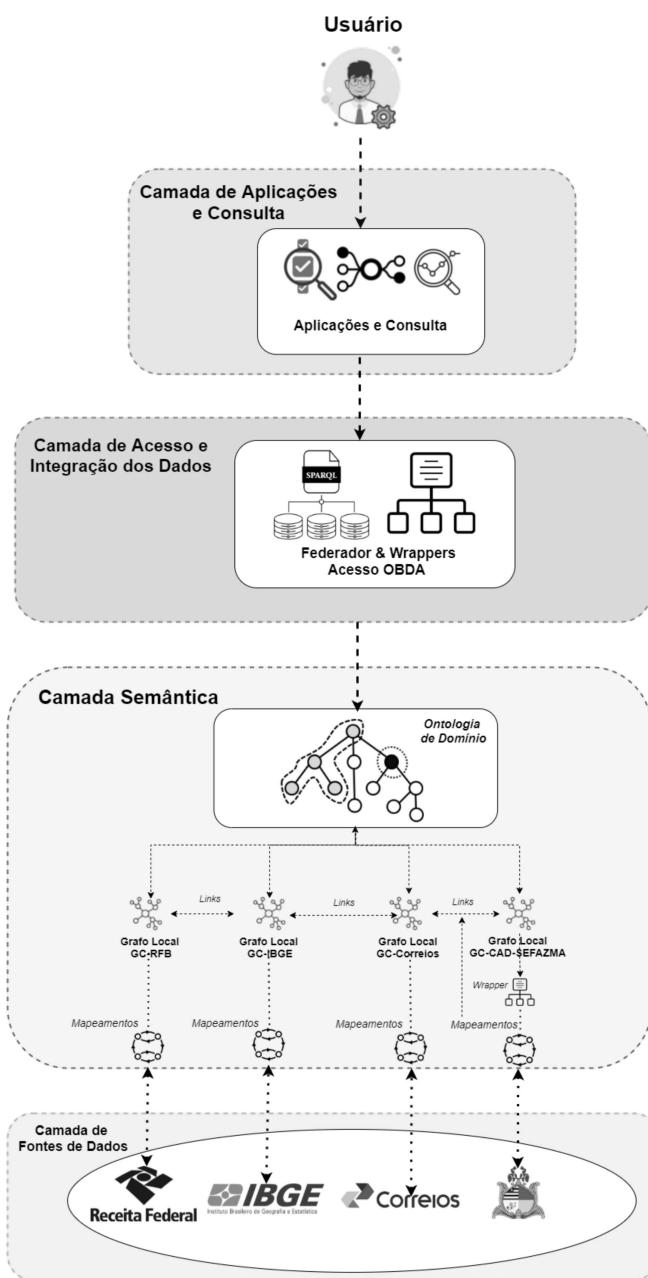


Figura 1. Arquitetura em 4 Camadas.

- **Camada de Fontes de Dados:** Na camada das fontes de dados estão armazenadas as fontes de dados utilizadas na construção do GCS. As fontes de dados na Figura 1, referem-se às fontes de dados integradas no grafo de conhecimento da SEFAZMA que será apresentado na Seção 4.

- **Camada Semântica:** Na camada semântica, é publicado um grafo de conhecimento local para cada fonte de dados usando o mesmo vocabulário da ontologia de domínio. A publicação é realizada através da criação de mapeamentos entre a ontologia de domínio e a fonte dos dados. Nessa camada, também são construídas ligações semânticas entre diferentes grafos locais. Conceitualmente, o GCS é definido a partir da união dos grafos locais em conjunto com as visões de ligações.

- Camada de Acesso e Integração de Dados: Nessa camada o acesso integrado aos dados é feito usando a Ontologia de Domínio. Essa abordagem permite que usuários e aplicações possam realizar consultas de forma transparente, sem ter que entender sobre as fontes de dados e suas características, como estrutura ou formato.

Os principais componentes dessa camada são o Federador (Mediador) e os *Wrappers*. O **Federador** permite a realização de consultas SPARQL sob múltiplos Grafos de conhecimentos locais (Consulta Federada). Enquanto que os **Wrappers** são responsáveis por traduzir consultas para a linguagem nativa das fontes de dados locais (em fontes RDF, o federador é um *endpoint* SPARQL). Para um grafo local materializado, o *Wrapper* é substituído pelo *endpoint* do *triplestore* que hospeda as triplas.

- Camada de Aplicações e Consulta Semântica: Um GC Semântico suporta o desenvolvimento de interfaces e aplicações Inteligentes baseadas na ontologia, tais como: Busca, Consulta, Mineração de dados, Chatbots, Visualizações, etc. A seção 4.4 apresenta as vantagens do uso de ontologias e do GCS em algumas dessas aplicações.

3. Construção da Camada Semântica

A Camada Semântica é obtida da integração semântica das fontes de dados. Chamamos de integração semântica o processo que faz uso de uma representação conceitual dos dados e seus relacionamentos para eliminar possíveis heterogeneidades. A construção da camada semântica é uma tarefa complexa que envolve quatro grandes desafios: (1) seleção das fontes de dados relevantes; (2) extração e tradução de dados de fontes de dados diferentes, possivelmente heterogêneas, para um vocabulário comum; (3) identificação de links entre recursos em diferentes fontes de dado; (4) Limpeza e resolução de inconsistências para melhorar a qualidade dos dados.

Os principais componentes da camada semântica são: A Ontologia de Domínio, um conjunto de grafos de conhecimentos exportados pelas fontes de dados locais; e um Conjunto de Visões de Ligações Semânticas;

A *ontologia de domínio* (O_D) é responsável por estabelecer um vocabulário a ser compartilhado para publicação dos grafos de conhecimento locais. Além de facilitar a integração de múltiplas fontes heterogêneas de dados, a ontologia provê uma representação semântica formal, o que permite inferências e processamento de máquina.

Um *Grafo de Conhecimento Local GCL* é uma visão RDF publicada, na camada semântica, por uma fonte de dados local, usando o vocabulário da ontologia de domínio. Um *GCL* é definido como uma tupla (M_L, O_L), onde:

- M_L especifica um conjunto de mapeamentos que relacionam os termos da fonte de dados S aos termos da ontologia de domínio.
- O_L estabelece a ontologia local que descreve a visão RDF publicada pela fonte de dados local; O vocabulário da ontologia local é um recorte do vocabulário da ontologia de domínio, e contém as classes e propriedades da fonte de dados que têm mapeamento para o vocabulário da O_D .

Uma *Visão de Ligação Semântica* especifica relacionamentos *owl:sameAs* entre instâncias em diferentes *GCLs*. É importante a criação dessas ligações devido ao problema de resolução de identidade, quando instâncias com diferentes *URIs* representam o mesmo objeto no mundo real.



Formalmente, uma visão de ligação semântica é uma n -tupla $L_i = (G_s, G_t, C, T, \mu)$, onde:

- G_s e G_t são Grafos de Conhecimentos (GCs) locais;
- C é uma classe comum no vocabulário de G_s e G_t ;
- T é um conjunto de propriedades comuns no vocabulário de G_s e G_t ;
- μ é uma 2_n -relação chamada de predicado de correspondência de instâncias.

O processo de construção da camada semântica é realizado de forma incremental em 5 passos (Figura 2), descritos a seguir:

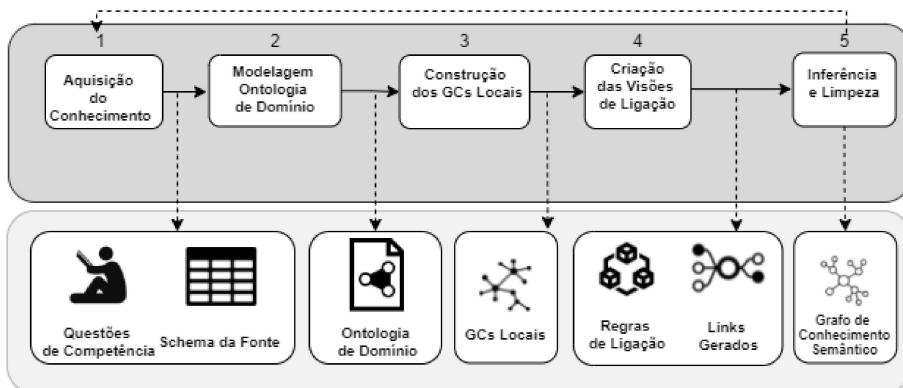


Figura 2. Processo para Construção da Camada Semântica.

Passo 1: Na etapa de Aquisição do conhecimento, em primeiro lugar deve-se estabelecer a meta por trás da coleta e integração dos dados, assim como definir as questões de competência que o usuário deseja que sejam respondidas. Em seguida, procurar descobrir quais as fontes de dados seriam mais úteis para atingir seu objetivo em termos de domínio, escopo, proveniência, manutenção, etc.

Passo 2: O processo de Modelagem da Ontologia de Domínio requer uma análise minuciosa dos diferentes esquemas das fontes de dados, objetivando a criação de um modelo comum para resolver o problema da heterogeneidade das fontes. Nessa atividade, deve-se procurar reusar ontologias já existentes, e formalizar o modelo usando os padrões como *RDF Schema* e *OWL*.

Passo 3: Para cada fonte de dados, é construído um grafo de conhecimento local usando o vocabulário da ontologia de domínio. A construção de um *GCL* é realizada em 3 passos: I) Primeiro, é gerado o mapeamento entre a fonte de dados e a ontologia de domínio; II) Então, baseado nos mapeamentos gerados, é especificada a ontologia do GC local; III) Por último, o *GCL* é implementado usando o enfoque virtual ou materializado de forma independente.

Considerando a **Abordagem Materializada** 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 *triplesstore*. As consultas sobre o grafo são processadas diretamente nos dados (visão RDF materializada). A desvantagem desse enfoque é que o grafo precisa ser atualizado quando atualizações são efetuadas na fonte de dados.

Já na **Abordagem Virtual** o *GCL* é uma visão virtual, onde as consultas realizadas sob o *endpoint* do *GCL* são reescritas em consultas sobre as fontes de dados originais. O uso dessa abordagem garante que os dados disponibilizados estejam sempre atualizados com relação às fontes originais. Porém, a desvantagem é o desempenho de consultas sobre fontes com grande volume de dados [Calvanese et al. 2017].

Passo 4: Nesse passo são criadas visões de ligação relacionando recursos em diferentes *GCLs* que tenham classes em comum. Para construir uma visão de ligação, o usuário terá que criar primeiro a regra de ligação (*linkage*) que especifica as condições para inferência de links “*same-As*”.

As vantagens desses links é permitir o aumento e enriquecimento de informações contidas nos dados através do *linked data*, de modo que novas informações e fatos possam ser descobertas entre fontes de dados distintas. Nessa perspectiva, essas fontes são interligadas seguindo uma ontologia, tendo como uma das principais práticas, a adoção da propriedade *owl:sameAs* [Paris 2018], propriedade utilizada para indicar que dois recursos são os mesmos. Caso o *GCL* seja materializado, sugere-se que os links possam ser descobertos através de alguma ferramenta de apoio tal como o Silk [Volz et al. 2009] e posteriormente gere-se as triplas de relações *owl:sameAs*.

De forma alternativa, os links podem ser estabelecidos em tempo de consulta - *on-the-fly*, onde através de mapeamentos (*e.g.*, R2RML) são definidas regras de equivalência entre duas tabelas/visões em uma visão SQL (*rr:sqlQuery*) ocorrendo um mapeamento de um *subject* com um *object* via predicado *owl:sameAs*. Nesse caso, conforme estado atual das fontes uma consulta consegue obter uma visão atualizada das relações *owl:sameAs*, possibilitando a inclusão ou remoção de novos *links* de forma virtual.

Passo 5: Neste passo, é realizado o processo de limpeza dos dados, aumentando a qualidade do GCS gerado. Esta limpeza busca tratar os problemas de conflito e redundância de dados, visto que os dados podem advir de várias fontes.

A limpeza se dá inicialmente pelo reconhecimento de classes e propriedades que possam conter fatos duplicados ou conflituosos, onde o espaço de busca é limitado para recursos que foram conectados através de links *owl:sameAs*. Definindo-se em seguida regras sobre como estes casos devem ser tratados. Este passo pode ser realizado com o uso de ferramentas tais como [Mendes et al. 2012].

A execução de um algoritmo de *Reasoner* sobre o GCS resultante pode ser realizada opcionalmente, gerando a inferência de novos fatos. Vale ressaltar que a etapa de *reasoning* é executada de diferentes maneiras dependendo do tipo *GCL* sendo tratado. No caso deste sendo um *GCL* materializado, o processo de *reasoning* pode acontecer durante o processo de materialização das triplas ou durante a carga destas no *triplesstore*. No primeiro, ferramentas como o *Ontop* [Calvanese et al. 2017], utilizam a ontologia de domínio para realizar a saturação dos mapeamentos de entrada, gerando um grafo resultante já contendo fatos inferidos seguindo o perfil OWL 2 QL¹. No segundo, *triplesstores* como o GraphDB² possuem motores de *reasoning* que executam este processo durante a carga dos dados, permitindo ao usuário escolher o perfil de inferências a ser realizado. No caso do *GCL* ser virtual, o processo de *reasoning* é executado em tempo de consulta pelo

¹https://www.w3.org/TR/owl2-profiles/#OWL_2_QL_2

²<https://graphdb.ontotext.com/>

mecanismo de OBDA (e.g. Ontop), onde este utiliza a ontologia de domínio para saturar os mapeamentos envolvidos na consulta.

4. Estudo de Caso: GC-SEFAZMA

Esta seção apresenta os resultados obtidos a partir do processo de construção do GC-SEFAZMA. O GC-SEFAZMA integra dados da Receita Federal do Brasil, Instituto Brasileiro de Geografia, Estatística, Correios e Cadastro de Contribuintes da SEFAZ-MA, permitindo um acesso integrado e homogêneo à estas bases.

4.1. Camada das Fontes de Dados

GC-SEFAZMA foi construído utilizando fontes de dados abertos, baseando-se nas premissas do *Linked Open Data*³. Como proposta inicial, GC-SEFAZMA integra 03 fontes de dados públicas e 01 interna a SEFAZ-MA: (1) O Cadastro Nacional de Pessoas Jurídicas da Receita Federal (RFB); (2) O IBGE-CNAE, fornecendo dados do Cadastro Nacional de Atividades Econômicas (CNAE); (3) IBGE-Location, fornecendo dados Localização (País, UF, Meso-região, Micro-região, Cidade, etc.); (4) Cadastro de Contribuintes do Maranhão.

4.2. Camada Semântica

Nesta seção são apresentados os resultados do processo de construção da camada semântica do GC-SEFAZMA, expondo detalhes dos artefatos gerados.

4.2.1. Aquisição do Conhecimento

Nessa etapa utilizou-se a abordagem Criação de Ontologias Orientada a Perguntas (COOP) [Ren et al. 2014] através de Questões de Competência (QCs). Como saída desta etapa foram obtidas questões de competência refinadas fornecidas pelos especialistas de domínio da SEFAZ-MA, conforme demonstrado na Tabela 1.

Questões de Competência (QCs)	
QC1	Quais empresas tem divergências de Sócios na RFB e no Cadastro da SEFAZ?
QC2	Quais empresas na RFB ou SEFAZ não tem sócio pessoa física?
QC3	Quais empresas não estão ativas na RFB, mas estão na SEFAZ?
QC4	Quais empresas são públicas?
QC5	Quais situações cadastrais da RFB são incompatíveis com as existentes na SEFAZ?

Tabela 1. Exemplo de Questões de Competência.

4.2.2. Modelagem e Construção da Ontologia de Domínio

A Modelagem e Construção da Ontologia de Domínio foi orientada através dos conceitos identificados nas QCs e na reutilização dos esquemas das fontes de dados. Foram feitos inicialmente modelos ontológicos em alto nível, que foram então agregados em um esquema global único, por conseguinte, esse modelo ontológico foi implementado em OWL

³<https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>

através da ferramenta Protégé⁴, tendo a definição de regras e axiomas necessários de modo a enriquecer a capacidade de inferência e descoberta sobre os dados. Figura 3 mostra as principais classes e relacionamentos da ontologia de domínio *ONTO_SEFAZMA*. Uma Visão mais detalhada da *ONTO_SEFAZMA* pode ser vista no link⁵.

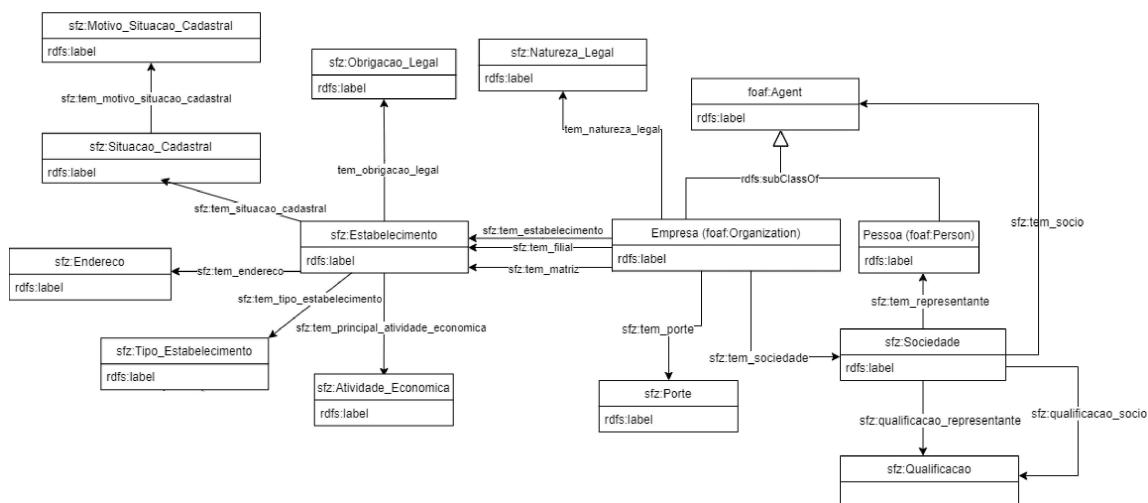


Figura 3. Fragmento da Ontologia de Domínio *ONTO_SEFAZMA*.

4.2.3. Construção dos Grafos de Conhecimento Locais

Nesta seção apresentamos os Grafos de Conhecimento Locais construídos, sendo 3 grafos locais materializados (GC-RFB, GC-IBGE e GC-Correios) e 1 grafo local virtual (GC-CAD_SEFAZMA) descritos a seguir. Outros artefatos dos *GCLs* podem ser vistos no link⁶.

GC-RFB: No tempo da realização dos trabalhos, a Receita liberava os dados de Pessoas Jurídicas publicamente, particionando-os em 20 partes no formato *fixed-width formatted files*, o que dificultava seu consumo. Deste modo, necessitou-se da implementação de um *script* feito em *python* para realizar a decodificação dos arquivos originais, seguindo de sua carga no SGBD PostgreSQL. Neste processo tabelas de apoio as quais os dados referenciavam também foram carregadas para o mesmo banco de dados. Em seguida, o esquema do banco de dados foi mapeado para a ontologia de domínio utilizando a linguagem R2RML. Por fim, os dados foram *triplicados* utilizando o Ontop e carregados no *triplestore*, armazenando também sua informação de proveniência. A Tabela 2 apresenta os conceitos exportados no *GCL* da RFB.

	GC-RFB
Classes	<i>sfz:Empresa, sfz:Estabelecimento, sfz:Sociedade, sfz:Pessoa, sfz:Endereco, sfz:Situacao_Cadastral, sfz:Porte, sfz:Natureza_Legal, sfz:Qualificacao,</i>
Relações	<i>sfz:tem_estabelecimento, sfz:tem_sociedade, sfz:tem_matriz, sfz:tem_endereco, sfz:tem_tipo_estabelecimento, sfz:tem_qualificacao_socio, sfz:tem_situacao_cadastral, sfz:tem_motivo_situacao_cadastral</i>

Tabela 2. Conceitos exportados pelo GC-RFB.

⁴<https://protege.stanford.edu/>

⁵http://tiny.cc/diagrama_od_cnpi

⁶<https://sites.google.com/view/camada-semantic-a-sefazma/in%C3%A7ao>

GC-IBGE: Os dados do IBGE foram utilizados para a construção de dois *GCLs*, IBGE-CNAE e IBGE Location. Os dados originalmente encontravam-se nos formatos *XLS* e *CSV*, respectivamente. Para a *triplificação* dos dois *datasets* foram construídos *scripts python* que liam os arquivos originais, gerando as triplas RDF segundo a ontologia de domínio. Os *GCLs* foram então carregados juntamente com suas informações de proveniência para o *triplestore*. A Tabela 3 apresenta os conceitos exportados no GC-IBGE.

	GC-IBGE
Classes	<i>sfx:Atividade_Economica, sfx:Secao, sfx:Divisao, sfx:Grupo, sfx:Classe, sfx:Subclasse, sfx:País, sfx:Unidade_Federativa, sfx:Cidade, sfx:Bairro, sfx:Logradouro, sfx:Mesoregiao, sfx:Microregiao, sfx:Localizacao, sfx:Localidade</i>
Relações	<i>sfx:tem_atividade_economica, sfx:tem_pais, sfx:tem_unidade_federativa, sfx:tem_cidade, sfx:tem_bairro, sfx:tem_logradouro</i>

Tabela 3. Conceitos exportados pelos GCL's do IBGE.

GC-Correios: Grafo publicado de forma materializada pela fonte de dados dos Correios. Seu processo de *triplificação* foi análogo ao do IBGE. As classes e propriedades do GC-Correios são expostas na Tabela 4.

	GC-Correios
Classes	<i>sfx:País, sfx:Unidade_Federativa, sfx:Cidade, sfx:Bairro, sfx:Logradouro, sfx:Mesoregiao, sfx:Microregiao, sfx:Localizacao, sfx:Localidade</i>
Relações	<i>sfx:tem_pais, sfx:tem_unidade_federativa, sfx:tem_cidade, sfx:tem_bairro, sfx:tem_logradouro</i>

Tabela 4. Conceitos exportados pelo GC-Correios

GC-CAD_SEFAZMA: Grafo publicado de forma virtual pelo Cadastro de Contribuintes do Maranhão. Na Tabela 5 são apresentadas as principais classes e propriedades do vocabulário do GC-CAD_SEFAZMA. Para construção do **GC-CAD_SEFAZMA** foi utilizado o *Ontop* como um *endpoint* virtual.

	GC-CAD_SEFAZMA
Classes	<i>sfx:Empresa, sfx:Estabelecimento, sfx:Sociedade, sfx:Pessoa, sfx:Porte sfx:Situacao_Cadastral, sfx:Natureza_Legal, sfx:Qualificacao, sfx:Endereco, sfx:Obrigacao_Legal, sfx:Contribuinte_Geral</i>
Relações	<i>sfx:tem_estabelecimento, sfx:tem_sociedade, sfx:tem_matriz, sfx:tem_filial, sfx:tem_tipo_estabelecimento, sfx:tem_endereco, sfx:tem_obrigacao_legal, sfx:tem_qualificacao_socio, sfx:tem_situacao_cadastral</i>

Tabela 5. Conceitos exportados pelo GC-CAD_SEFAZMA

4.2.4. Criação das Visões de Ligação

Para a criação das visões de ligação, foram estabelecidas regras através da utilização da ferramenta *Silk* para descoberta dos links entre os 3 grafos de conhecimento locais materializados utilizando o predicado *owl:sameAs*. A escolha de uso *Silk* justifica-se por seu conjunto de filtros e regras, gerenciáveis através de *workflows* automáticos para descoberta e geração de dados ligados [Volz et al. 2009].

No link ⁷ são apresentadas as regras para descoberta de links entre os *GCLs* da RFB e IBGE-CNAE através da propriedade (*sfx: cnae*), comparando-se seus valores através da métrica *equality* (igualdade).

Já para definição dos links envolvendo o **GC-CAD_SEFAZMA**, foram utilizadas regras de junção em mapeamentos para estabelecer relações *owl:sameAs* entre recursos do GC-RFB e do **GC-CAD_SEFAZMA**. Um exemplo pode ser visto no seguinte link⁸.

⁷<https://sites.google.com/view/camada-semantica-sefazma/in%C3%ADcio>

⁸<https://pastebin.com/raw/QxfBSmdv>

4.2.5. Inferências e Limpeza dos dados

Como forma de enriquecimento e aumento da qualidade dos dados de endereço do GC-SEFAZMA, o GC-Correios foi utilizado para validar os endereços declarados, verificando a partir de dois estabelecimentos idênticos relacionados por *owl:sameAs* se estes são válidos com base no CEP. Isto é feito ao verificar se o CEP é válido e se o endereço declarado (logradouro, bairro, município e estado) corresponde à aquele CEP.

A limpeza dos dados, se deu por meio de links com dados de localização, geolocalização e CEPS do GC-Correios de modo a corrigir conflitos, ausências e inconsistências nos dados de endereço, inferindo a informação mais atualizada e correta, com ênfase na melhora da qualidade dos dados contidos no GC-SEFAZMA.

Ao final da construção, GC-SEFAZMA possui um grande conjunto de inferências, onde por exemplo, entidades do tipo “Empresa”, são inferidas como “idênticas” a partir dos links *owl:sameAs*, permitindo a completude e corretude de dados ausentes ou incorretos relacionados a seu quadro de sócios e estabelecimentos, situação cadastral e outros, possibilitando, por exemplo, inferir a descoberta de eventuais contribuintes fantasmas e/ou irregulares.

4.3. Camada de Acesso e Integração dos Dados

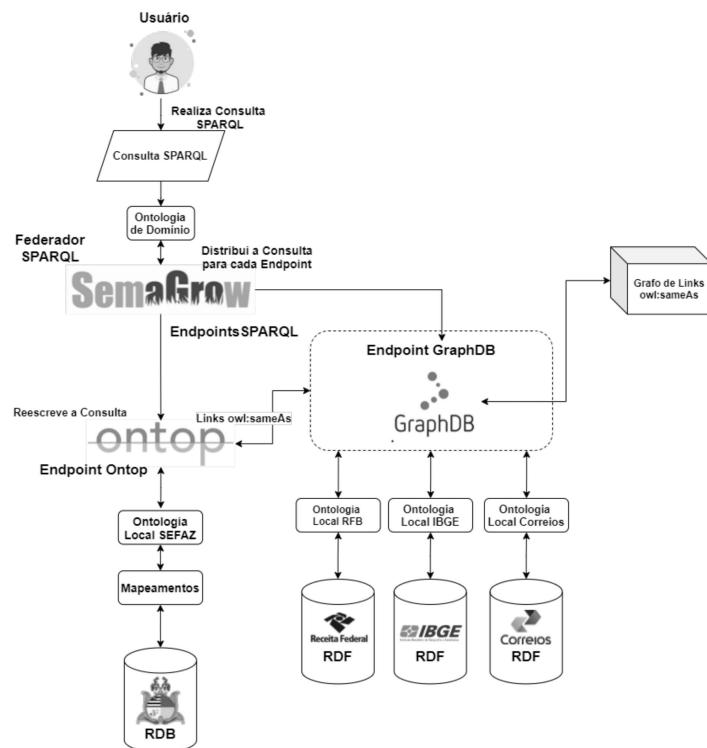


Figura 4. Acesso e Integração dos dados do GCS-SEFAZMA.

A Figura 4 mostra os principais componentes da camada de acesso e Integração dos dados do GC-CAD_SEFAZMA. Para a geração dos *GCLs* materializados, foi utilizado o software *Ontop* na geração dos grafos locais GC-RFB, GC-IBGE e GC-Correios. Nessa etapa o *Ontop* teve como parâmetro de entrada a ontologia e o mapeamento con-

tidos na especificação de cada *GCL*. Foi criado um repositório através do *TripleStore GraphDB* para armazenar os *GCLs*.

Para construção do GC-CAD_SEFAZMA, a abordagem virtual foi realizada através do uso do *Ontop* como um *wrapper*. O ontop segue a abordagem *Ontology-Based Data Access* (OBDA), onde é feito um processamento online de uma dada consulta SPARQL e sua tradução em SQL usando os mapeamentos. Ainda nesse estágio são feitas otimizações na consulta SQL, sendo executada pelo mecanismo de banco de dados e ao final realizando a tradução do resultado para RDF [Calvanese et al. 2017].

O *Semagrow* [Charalambidis et al. 2015] foi utilizado como federador para processamento da consulta sob os múltiplos endpoints disponibilizados, de modo à fornecer a capacidade de realização de consultas SPARQLS sob grafos locais materializados e virtual.

4.4. Camada de Aplicações e Consulta Semântica

Durante nossos experimentos notamos que três principais tipos de aplicações alcançaram um bom nível de aceitação por parte dos usuários, além de poderem ser facilmente aplicados em qualquer cenário, sendo estes:

Navegação do Grafo: Estas ferramentas renderizam de maneira interativa o grafo subjacente de uma maneira gráfica, permitindo que o usuário visualize de maneira concreta as entidades contidas neste grafo e como estas estão relacionadas entre si. Como vantagem deste tipo de ferramenta destaca-se a falta da necessidade de conhecimentos técnicos sobre RDF ou consultas SPARQL, além de não necessitar que o usuário tenha conhecimentos prévios sobre o esquema dos dados sendo navegados. Neste trabalho indicamos o uso da ferramenta baseada em *Web*, *Ontodia*, para a renderização de navegação do GCS [Mouromtsev et al. 2015].

Interface Visual Para a Construção de Consultas Semânticas: Ferramentas de consulta visual podem amenizar a necessidade de conhecimentos técnicos para a construção de consultas, permitindo que a consulta seja definida de maneira visual e interativa com base na ontologia de domínio. Neste trabalho, a ferramenta *Optique VQS* [Soylu et al. 2018] foi selecionada para dar suporte à construção de consultas. A ferramenta permite a construção visual de consultas, atuando de maneira interativa com base na ontologia de domínio.

Visualizações de Dados: Neste trabalho propomos o uso de *dashboards* de visualizações interativas com gráficos interconectados que permitem ao usuário a aplicação de filtros e a visualização de diversas facetas dos dados. Neste estudo de caso foi construído um portal de *dashboards* utilizando uma amostra dos dados gerados na seção 4, com uma demo online disponível no link⁹.

5. Trabalhos Relacionados

Trabalhos como [Collaran et al. 2017], [Bizon et al. 2019] e [Galkin et al. 2016] apresentam ferramentas que abordam pontos específicos do processo de construção de um GCS. Enquanto os dois primeiros tem como foco a descoberta de links e fusão dos dados, o terceiro foca no processo de triplificação automática dos dados.

⁹<https://anonymous.4open.science/w/VizRFB-9E7E/>

Considerando trabalhos no domínio de dados de pessoas jurídicas de fontes públicas, em [Victorino et al. 2018], os autores apresentam um modelo conceitual proposto, junto de sua arquitetura e uma ferramenta demonstrativa para facilitar a busca por dados abertos disponibilizados pelo governo brasileiro. A abordagem faz uso de ontologias para a transformação dos dados abertos para *Linked Open Data*. No entanto, o estudo não aborda aspectos da homogenização e integração semântica das fontes, limitando-se a ligá-las e publicá-las.

Deste modo, podemos observar que até então as pesquisas na área vem focando em pontos específicos dos processos de transformação, integração, publicação ou consumo. Neste contexto, nosso trabalho diferencia-se por tratar todo o processo de construção de grafos de conhecimento semânticos através de um enfoque híbrido, além de fornecer um processo completo para a construção de uma camada semântica integrada.

No contexto das tecnologias da Web Semântica, [Bizer et al. 2012] apresenta o Ldif, um *framework* para facilitar o processo de construção de aplicações de *Linked Data* (LD). O LDIF é capaz juntar dados de múltiplas fontes de dados da LD, gerando um *Linked Data Mashup* (LDM), uma representação única local e limpa dos dados. Um LDM pode ser visto como um GCS materializado específico para uma aplicação.

Em [Calvanese et al. 2017], os autores apresentam a abordagem *Ontology-Based Data Access* (OBDA), onde é criada uma camada semântica virtual sobre um banco de dados relacional. No entanto, esta abordagem não trata aspectos da integração semântica de fontes heterogêneas (e.g. homogenização de esquemas, descoberta de links, etc.). Deste modo, nossa abordagem diferencia-se das demais por tratar-se de um enfoque híbrido, onde o GC resultante é uma integração semântica de GCs locais materializados e virtuais.

6. Conclusões

O artigo apresentou uma arquitetura de quatro camadas para construção de Grafos de conhecimentos Semânticos. Nessa arquitetura, os componentes de um GCS estão agrupados em cinco camadas: Camada das Fontes de Dados, Camada Semântica, Camada de Integração e Acesso e Camada de Aplicações e Consulta Semântica.

Em seguida, o artigo apresentou a um enfoque incremental para construção da Camada Semântica do GCS. No enfoque proposto a camada Semântica é obtida da integração semântica das fontes de dados e envolve 5 passos principais: aquisição de conhecimento, modelagem da ontologia de domínio, construção dos grafos de conhecimento locais, geração das visões de ligações, e limpeza e avaliação.

Por último, o artigo discutiu a construção do grafo de conhecimento corporativo da SEFAZ-MA, usando a arquitetura e processo proposto. Durante a construção do GC-SEFAZMA mostraram-se nítidas as vantagens do uso das tecnologias da web Semântica no processo de construção e manutenção do grafo.

Agradecimentos

Esse trabalho foi financiado pelo BID, Processo: 22831/2021/SEFAZ-MA, referente ao contrato de acordo de parceria: 01/2019 – PROFISCO II/SEFAZ-MA.



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

Referências

- Bizer, C., Becker, C., Mendes, P. N., Isele, R., Matteini, A., and Schultz, A. (2012). Ldif-a framework for large-scale linked data integration.
- Bizon, C., Cox, S., Balhoff, J., Kebede, Y., Wang, P., Morton, K., Fecho, K., and Tropsha, A. (2019). Robokop kg and kgb: integrated knowledge graphs from federated sources. *Journal of chemical information and modeling*, 59(12):4968–4973.
- Calvanese, D., Cogrel, B., Komla-Ebri, S., Kontchakov, R., Lanti, D., Rezk, M., Rodriguez-Muro, M., and Xiao, G. (2017). Ontop: Answering sparql queries over relational databases. *Semantic Web*, 8(3):471–487.
- Charalambidis, A., Troumpoukis, A., and Konstantopoulos, S. (2015). Semagrow: Optimizing federated sparql queries. In *Proceedings of the 11th International Conference on Semantic Systems*, pages 121–128.
- Collarana, D., Galkin, M., Traverso-Ribón, I., Lange, C., Vidal, M.-E., and Auer, S. (2017). Semantic data integration for knowledge graph construction at query time. In *2017 IEEE 11th International Conference on Semantic Computing (ICSC)*, pages 109–116. IEEE.
- Galkin, M., Auer, S., Kim, H., and Scerri, S. (2016). Integration strategies for enterprise knowledge graphs. In *2016 ieee tenth international conference on semantic computing (icsc)*, pages 242–245. IEEE.
- Grainger, T., AlJadda, K., Korayem, M., and Smith, A. (2016). The semantic knowledge graph: A compact, auto-generated model for real-time traversal and ranking of any relationship within a domain. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 420–429. IEEE.
- Mendes, P. N., Mühleisen, H., and Bizer, C. (2012). Sieve: linked data quality assessment and fusion. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, pages 116–123.
- Mouromtsev, D., Pavlov, D. S., Emelyanov, Y., Morozov, A. V., Razdyakonov, D. S., and Galkin, M. (2015). The simple web-based tool for visualization and sharing of semantic data and ontologies. In *International Semantic Web Conference (Posters & Demos)*.
- Paris, P.-H. (2018). Assessing the quality of owl: sameas links. In *European Semantic Web Conference*, pages 304–313. Springer.
- Ren, Y., Parvizi, A., Mellish, C., Pan, J. Z., Van Deemter, K., and Stevens, R. (2014). Towards competency question-driven ontology authoring. In *European Semantic Web Conference*, pages 752–767. Springer.
- Soylu, A., Kharlamov, E., Zheleznyakov, D., Jimenez-Ruiz, E., Giese, M., Skjæveland, M. G., Hovland, D., Schlatte, R., Brandt, S., Lie, H., et al. (2018). Optiquevqs: a visual query system over ontologies for industry. *Semantic Web*, 9(5):627–660.
- Victorino, M., de Holanda, M. T., Ishikawa, E., Oliveira, E. C., and Chhetri, S. (2018). Transforming open data to linked open data using ontologies for information organization in big data environments of the brazilian government: the brazilian database government open linked data–dbgoldbr. *KO KNOWLEDGE ORGANIZATION*, 45(6):443–466.

Volz, J., Bizer, C., Gaedke, M., and Kobilarov, G. (2009). Silk-a link discovery framework for the web of data. In *Ldow*.



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)