Construção do Grafo de Conhecimento Semântico de Dados Abertos de Pessoas Jurídicas

Tulio Vidal Rolim[1,*,†], Caio Viktor S. Avila[1,†], Narciso M. A. Junior[1], Jose J. Dutra[1], Jamires Costa[1], Roberval G. Mariano[2], Angelo R. A. Brayner[1] and Vania M. P. Vidal[1]

[1]*Universidade Federal do Ceará, Ceará - Brasil*

[2]*Secretaria da Fazenda do Maranhão, São Luís, Brasil*

Abstract

This article presents the construction of a semantic knowledge graph of the open data of legal entities (**SKG:CNPJ**). **SKG:CNPJ** is obtained from the semantic integration of four data sources: RFB, IBGE, Correios, TCU and CEIS. The main objective of **SKG:CNPJ** is to provide a semantic layer, so that applications can have integrated access to the data source data through the Semantic Layer. The data and metadata of the **SKG:CNPJ** are made available on a Semantic Portal (**Semantic-CNPJ**) for query and visualization. The article also describes the use of **SKG:CNPJ** for construction of semantic queries.

Keywords

Semantic Knowledge Graph, Open Data, Ontology, Semantic Integration

1. Introdução

Os dados de pessoas jurídicas, são uma importante fonte para questões fiscais, sendo uma fonte importante para diagnosticar eventuais irregularidades, auxiliando na descoberta de ações não saudáveis por parte de empresas no âmbito público. Trabalhos recentes [1], [2], [3] demonstram esforços para se integrar a crescente quantidade de coleções de dados públicos através de uma semântica bem definida na melhoria do processo de descoberta de conhecimento.

Esse trabalho construção de um Grafo de Conhecimento Semântico (**SKG:CNPJ**) que integra 5 fontes de dados abertas públicas: **O Cadastro Nacional de Pessoas Jurídicas (CNPJ) da Receita Federal (RFB), IBGE (IBGE-CNAE** e **IBGE-Localizacao), Correios, Cadastro Nacional de Empresas Inidôneas e Suspensas (CEIS)** e do **Sistema de Inabilitados e Inidôneos (TCU)**. 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. O objetivo principal de uma integração de dados baseadas em GCS é fornecer uma camada de dados unificada, flexível e usável, que é semanticamente

Proceedings of the 15th Seminar on Ontology Research in Brazil (ONTOBRAS) and 6th Doctoral and Masters Consortium on Ontologies (WTDO), November 22-25, 2022

*Corresponding author.

†These authors contributed equally.

tulio.xcrtf@gmail.com (T. V. Rolim); arlaass@gmail.com (C. V. S. Avila); narcisoarruda@gmail.com (N. M. A. Junior); lanodutra@gmail.com (J. J. Dutra); jamirescostaa@gmail.com (J. Costa); mariano@sefaz.ma.gov (R. G. Mariano); brayner@dc.ufc.br (A. R. A. Brayner); vaniap.vidal@gmail.com (V. M. P. Vidal)

conectada à fonte de dados, para que aplicações possam ter acesso integrado aos dados das fontes através de uma Camada Semântica.

Durante a construção do GCS de fontes de fontes externas, foram identificados muitos desafios, tais como, extração dos dados com diferentes formatos de armazenamento, construção de mapeamentos para resolver problema da heterogeneidade de vocabulários, descoberta de *links* entre recursos em diferentes fontes de dados (resolução de entidade), e resolução de inconsistências e conflitos para melhorar a qualidade dos dados. De forma geral, as principais contribuições fornecidas por esse trabalho são:

- Construção de um GCS de Dados Abertos de Pessoas Jurídicas **SKG:CNPJ** a partir da extração de fontes de dados externas e semi-estruturadas;
- Um Portal Semântico de fontes de dados de Pessoas Jurídicas (*Semantic-CNPJ*) para acesso, consulta e visualização dos dados e metadados do **SKG:CNPJ**.

O restante do artigo está organizado da seguinte forma: A Seção 2 apresenta a arquitetura de 5 camadas proposta para representação e organização do **SKG:CNPJ**. A Seção 3 apresenta a Ontologia de Domínio do **SKG:CNPJ**. A Seção 4 descreve os componentes de cada camada do **SKG:CNPJ**. A Seção 5 apresenta os trabalhos relacionados. Por fim, a Seção 6 apresenta as conclusões.

2. Modelo conceitual do SKG:CNPJ

O modelo conceitual é responsável por estabelecer um vocabulário a ser compartilhado para publicação das fontes de dados locais. Além de facilitar a integração de múltiplas fontes heterogêneas de dados, este modelo provê uma representação semântica formal.

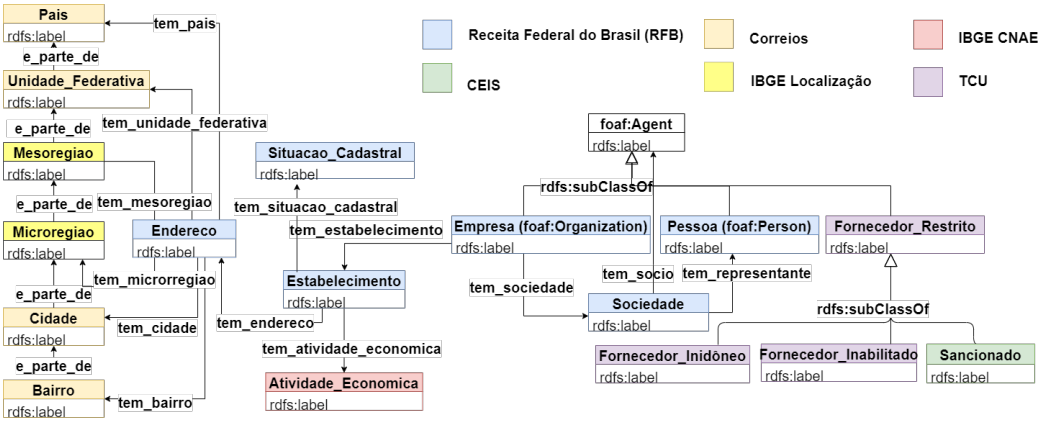

Figure 1: Recorte da Ontologia de Domínio do **SKG:CNPJ**.

A Modelagem e Construção do modelo foi orientada através dos conceitos identificados no domínio de pessoa jurídica e no reuso dos esquemas das fontes de dados, tendo ajuda de especialistas de domínio pertencentes à SEFAZ do Maranhão. Inicialmente, foram criados

modelos ontológicos em alto nível para cada fonte, seguindo os esquemas das fontes originais. Em seguida, este modelos foram refinados com a ajuda dos especialistas, tendo como objetivo a correta representação semântica dos dados de acordo com a comunidade interessada. Posteriormente estes modelos foram unificados em um esquema global único. O modelo resultante foi implementado em OWL através do Protégé[1], tendo as regras e axiomas necessários para permitir inferências sobre os dados. As principais classes e relacionamentos do modelo conceitual final são mostradas na Figura 1, estando disponível em versão completa no Portal Semântico no link[2].

3. Camadas do SKG:CNPJ

Nessa seção detalhamos os componentes de cada uma das camadas do SKG:CNPJ. Esses componentes podem ser acessados no Portal Semântico.

3.1. Camada das Fontes de Dados

Na camada das fontes de dados estão armazenados os dados extraídos das fontes de dados externas. Em muitos casos, os dados de fontes externas semiestruturadas estão armazenados em vários arquivos. Para cada arquivo de uma fonte externa é criada, na camada das fontes de dados, uma tabela de extração com estrutura similar a essa do arquivo.

Web scrapers foram desenvolvidos para detectar se houveram atualizações nos sites das fontes externas. No caso de haver uma atualização dos arquivos de dados das fontes externas, o processo de ingestão desses dados é disparado. Esses arquivos, primeiro são extraídos no seu formato nativo, e, em seguida, são transformados e armazenados nas tabelas de extração.

Nas tabelas de extração, não há restrições de integridade ligadas, tendo em vista que o processo de higienização dos dados será realizado em etapas posteriores. No Portal semântico, na aba Camada das Fontes de Dados, pode-se acessar a descrição, artefatos (Dicionários, Diagramas), origem e proveniência fontes de dados trabalhadas através do link[3].

3.2. Camada Relacional

Na Camada Relacional normalizada, os dados das tabelas de extração devem ser transformados e armazenados em tabelas relacionais normalizadas. O esquema das tabelas normalizadas usam um vocabulário comum, construído a partir do vocabulário do modelo conceitual. O mapeamento das visões relacionais para o modelo conceitual tem a semântica dos mapeamentos diretos [4], o que facilita a criação e manutenção desses mapeamentos. Assim como, a diminuição do "gap" semântico entre o modelo relacional e modelo RDF [5].

A construção do esquema relacional normalizado é realizado ao fazer o *"matching"* do esquema das tabelas de extração para as classes e propriedades do modelo conceitual. Vale salientar que esse processo foi realizado de forma manual, e que o esquema relacional normalizado gerado

[1]https://protege.stanford.edu/
[2]https://semantic-cnpj.github.io/Semantic-CNPJ/ontologia_dominio.html
[3]https://semantic-cnpj.github.io/Semantic-CNPJ/camada_fontes

garante o mapeamento direto para a ontologia de domínio. No Portal Semântico, os diagramas dos esquemas normalizados e scripts de criação podem ser acessados no link[4].

O processo de povoamento das tabelas normalizadas é realizado através de funções que consultam as tabelas de extração, checam a consistência dos registros retornados, e os armazenam nas tabelas normalizadas.

3.3. Camada Semântica

A Camada Semântica[5] 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. Os componentes da Camada Semântica são:

- **Grafos de Conhecimento Locais (GCLs):** Consiste em visões RDF publicadas, na camada semântica, por fonte de dados, mapeamentos e ontologias [6];
- **Visões de Ligação:** Especifica ligações semânticas entre instâncias em diferentes GC_L's das fontes de dados [6];

Na Camada Semântica, um GC_L é gerado para cada fonte de dados usando o mesmo vocabulário do modelo conceitual. Um GC_L é uma visão RDF virtual [7], sendo esta definida por um conjunto de mapeamentos que relacionam os termos do esquema relacional normalizado aos termos do modelo conceitual. Esses mapeamentos são mapeamentos diretos definidos usando o *R2RML* [8]. Adotando os mapeamentos diretos, o esquema e a complexidade das visões *R2RML* contidas nos mapeamentos são refletidas diretamente a partir das tabelas normalizadas.

Nessa camada, também são definidas ligações semânticas entre instâncias em diferentes grafos de conhecimento das fontes de dados. As ligações semânticas são virtuais, definidas a partir de mapeamentos *R2RML* [8].

Conceitualmente, o **SKG:CNPJ** é um grafo virtual definido a partir da união dos grafos das fontes de dados locais juntamente com as ligações semânticas entre suas instâncias. Todos os grafos de conhecimento locais com seus respectivos modelos conceituais locais e mapeamentos podem ser acessados no portal semântico através do link[6]. As visões de ligação podem ser encontradas no link[7].

3.4. Camada de Acesso e Integração dos Dados

O GraphDB foi utilizado como Endpoint e o *Ontop* foi utilizado como *wrapper* acoplado ao GraphDB através de um repositório para construção do **SKG:CNPJ** utilizando a abordagem virtual.

O Ontop adota um *workflow* para construção do grafo local virtual com base em 2 estágios (*offline* e *online*). Onde no estágio *offline*, é realizado o processo de leitura e processamento do modelo conceitual, mapeamentos e verifica as restrições de integridade do banco de dados,

[4]https://semantic-cnpj.github.io/Semantic-CNPJ/camada_relacional.html
[5]https://semantic-cnpj.github.io/Semantic-CNPJ/camada_semantica.html
[6]https://semantic-cnpj.github.io/Semantic-CNPJ/grafos_locais.html
[7]https://semantic-cnpj.github.io/Semantic-CNPJ/visoes_ligacoes_semanticas.html

expandindo os mapeamentos para se adequarem aos axiomas definidos no esquema conceitual. No segundo estágio, o Ontop segue a abordagem *Ontology-Based Data Access (OBDA)*, onde é feito um processamento *online* de uma consulta SPARQL dada como entrada. Em seguida, esta consulta é traduzida para sua equivalente em SQL, utilizando as regras definidas nos mapeamentos. Ainda nesse estágio são feitas otimizações na consulta SQL. Por fim, a consulta SQL é executada pelo mecanismo de consultas do banco de dados relacional, tendo seu resultado retornado para o usuário [7].

3.5. Camada de Aplicações

Nesta seção é apresentada uma aplicação do grafo construído para ilustrar os usos do **SKG:CNPJ** construído. A seção 3.5.1 aborda o cenário da realização de consultas semânticas sobre **SKG:CNPJ**.

3.5.1. Interface de Consulta

Além da opção da escrita direta de uma consulta SPARQL, consultas podem ser construídas com o apoio de um Sistema de Consultas Visuais (*Visual Query System, VQS)*. O uso de um VQS permite que usuários leigos possam consultar os dados de maneira visual e intuitiva. Neste trabalho, a ferramenta *Optique VQS* [9] foi selecionada para dar suporte à construção de consultas. A ferramenta atua de maneira interativa, onde o usuário seleciona uma classe alvo, seguindo da ferramenta apresentado os atributos (podendo estes serem escolhidos como elementos de filtragem ou projeção) e os relacionamentos desta classe (definindo um caminho), guiando o usuário na construção da consulta.

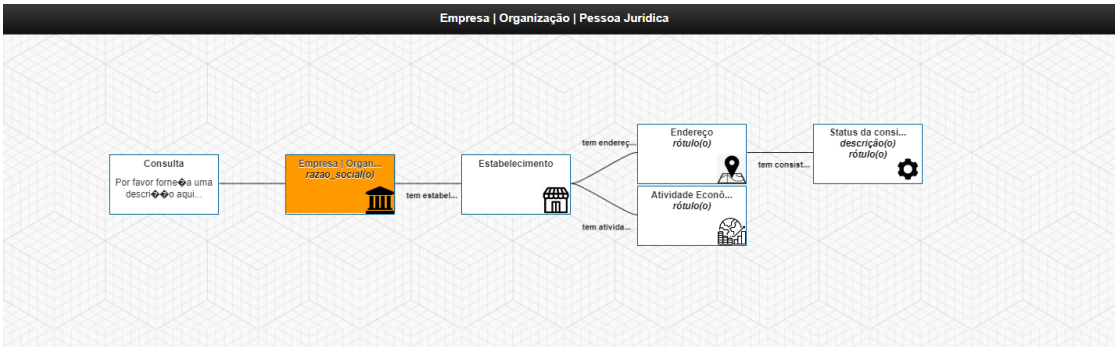

Figure 2: Exemplo de consulta utilizando o VQS

A Figura 2 apresenta um exemplo de consulta construída com a ajuda da ferramenta *Optique VQS*. Esta consulta busca as empresas contidas no grafo, além de seus estabelecimentos. Onde para cada estabelecimento são recuperadas suas atividades econômicas e seu endereço. Este exemplo apresenta uma consulta complexa, utilizando vários conceitos de fontes diferentes, tais como: *Empresa, Estabelecimento e Endereços da RFB; Atividades Econômicas do IBGE*; e *Logradouros* oficialmente cadastrados nos Correios que foram utilizados para calcular as

consistências de endereços da RFB. Deste modo, esta consulta destaca a importância do GCS construído, permitindo uma consulta integrada às múltiplas fontes de dados de maneira transparente ao usuário, além de auxiliar no processo de escrita de consultas complexas.

4. Trabalhos Relacionados

No domínio de dados de pessoas jurídicas de fontes públicas, trabalhos recentes demonstram esforços para se integrar a crescente quantidade de coleções de dados públicos. Dentre estes, alguns utilizam tecnologias da Web Semântica, tais como, ontologias, RDF, *links*, etc. Isto com o intuito de publicar e ou integrar estes conjuntos de dados com o uso da semântica visando melhorias na transparência e no processo de descoberta de conhecimento.

Em [1], os autores apresentam um modelo conceitual proposto, junto de sua arquitetura e uma ferramenta demostrativa 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. [10] apresenta uma forma de garantir um conjunto de metadados capazes de descrever *datasets* publicados por municípios, fazendo assim com que os dados sejam encontrados de uma forma mais simples além de fornecer uma linguagem comum e compreensível ao cidadão.

[11] propõe uma ontologia de domínio sobre licitações como base de conhecimento primário, visando facilitar a elicitação de requisitos para novos portais de transparência municipal.

Tendo em vista os trabalhos anteriormente citados, podemos observar que até então as pequisas 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 para a criação de um grafo de conhecimento semântico baseado em dados abertos públicos, desde sua modelagem, representação, acesso e consumo.

5. Conclusões

Esse artigo descreve um Grafo de Conhecimento Semântico resultante da integração de fontes de dados abertas relacionadas a pessoas jurídicas.

Primeiro, o artigo apresentou a arquitetura de cinco camadas usada para a construção do **SKG:CNPJ**. Depois, são descritos os metadados de cada camada do **SKG:CNPJ**, os quais estão disponíveis para consulta no portal semântico do **SKG:CNPJ**. Os dados do **SKG:CNPJ** também podem ser consumidos através do endpoint SPARQL disponibilizado no Portal.

Por último, foi apresentado um estudo de caso ilustrando o uso do **SKG:CNPJ** para realização de consultas semânticas usando a ferramenta VQS.

References

[1] M. Victorino, M. T. de Holanda, E. Ishikawa, E. C. Oliveira, S. Chhetri, 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 (2018) 443–466.

[2] L. S. de Oliveira Araújo, M. T. Santos, D. A. Silva, The brazilian federal budget ontology: a semantic web case of public open data, in: Proceedings of the 7th International Conference on Management of computational and collective intElligence in Digital EcoSystems, 2015, pp. 85–89.

[3] L. M. Nascimento, Utilizando linked data para publicação e cruzamento de dados governamentais abertos (2017).

[4] W3C, A direct mapping of relational data to rdf, 2012. URL: https://www.w3.org/TR/rdb-direct-mapping/.

[5] A. Bertails, E. G. Prud'hommeaux, Interpreting relational databases in the rdf domain, in: Proceedings of the sixth international conference on Knowledge capture, 2011, pp. 129–136.

[6] T. V. Rolim, C. V. S. Avila, R. G. Mariano, T. Calixto, P. Ivo, J. M. M. Filho, A. Brayner, V. M. P. Vidal, Uso das tecnologias da web semântica na construção de grafos de conhecimento semântico baseado no enfoque híbrido. (use of semantic web technologies in the construction of semantic knowledge graphs based on the hybrid approach), in: ONTOBRAS, 2021.

[7] D. Calvanese, B. Cogrel, S. Komla-Ebri, R. Kontchakov, D. Lanti, M. Rezk, M. Rodriguez-Muro, G. Xiao, Ontop: Answering sparql queries over relational databases, Semantic Web 8 (2017) 471–487.

[8] W3C, R2rml: Rdb to rdf mapping language, 2012. URL: https://www.w3.org/TR/r2rml/.

[9] A. Soylu, E. Kharlamov, D. Zheleznyakov, E. Jimenez-Ruiz, M. Giese, M. G. Skjæveland, D. Hovland, R. Schlatte, S. Brandt, H. Lie, et al., Optiquevqs: a visual query system over ontologies for industry, Semantic Web 9 (2018) 627–660.

[10] L. M. F. Pereira, et al., Ogdpub: uma ontologia para publicação de dados abertos governamentais (2017).

[11] T. L. Bernardi, et al., Uma ontologia sobre licitações aplicada na elicitação de requisitos de portais de transferência municipal, Ph.D. thesis, Universidade de Passo Fundo, 2017.