Aplicativo que gera enriquecimento semântico de dados abertos: Estudo de caso do metrô de São Paulo

Giovane Santos Silva

Faculdade de Tecnologia de Garça – giovanesantos1999@gmail.com

Larissa Pavarini da Luz

Universidade Estadual Paulista Júlio de Mesquita Filho - Marilia

Faculdade de Tecnologia de Garça – larissa.luz01@fatec.sp.gov.br

Resumo

O presente trabalho apresenta a proposta de um aplicativo que gera enriquecimento semântico a partir de dados abertos na Web, tendo como estudo de caso os dados do metrô da cidade de São Paulo. O metrô é um meio de transporte extremamente importante principalmente em cidades grandes, no Brasil cidades como Rio de Janeiro e São Paulo são movidos por esse meio de transporte tão importante. Verifica-se então, que ainda hoje, além da importância da locomoção há também a importância de saber as linhas e as possibilidades de funcionamento dele. Essas informações são ricas e preciosas para o cotidiano das pessoas. Baseado nesse contexto, e na linha de pesquisa que envolve dados abertos e informações do transporte público, realizou-se uma pesquisa exploratória por meio de estudo bibliográfico, e aplicado. Desta forma a abordagem do trabalho foi qualitativa, pois foi feito um levantamento e análise desses dados, trazendo como resultados parciais o desenvolvimento de um aplicativo com a base no enriquecimento semântico de dados.

Palavras-chave*:* Metrô. Semântica. Dados abertos. Enriquecimento Semântico.

Application that generates semantic enrichment: Case study of the São Paulo subway

Abstract

The present work presents the proposal of an application that generates semantic enrichment from open data on the Web, using the case data of the São Paulo subway as a case study. The subway is an extremely important means of transport, especially in big cities, in Brazil cities like Rio de Janeiro, São Paulo are moved by this very important means of transport. It appears that even today, in addition to the importance of locomotion, there is also the importance of knowing the lines and the possibilities of its functioning. This information is rich and precious for people's daily lives. Based on this context, and in the line of research that involves open data and public transport information, an exploratory research was carried out through a bibliographic study, and applied. Thus, the work approach was qualitative, as a survey and analysis of this data was carried out, bringing as partial results the development of an application based on semantic enrichment.

Keywords: Subway. Semantics. Open data. Semantic enrichment.

---

1 Introdução

A Surgida em 1991, a *World Wide Web* (WWW), ou simplesmente WEB, é atualmente tão popular e utilizada por todos no mundo e que não difícil, no imaginário dos usuários confundida facilmente pela Internet (KARINE; GABRIELLE, 2013).

Portanto a Internet é uma grande rede que conecta milhões de computadores enquanto a WEB é uma das ferramentas onde usamos para acessar essa rede utilizando navegadores tais como o Chrome da Google mundialmente conhecido.

Dentro da WEB existe um conceito bastante importante que é a WEB semântica ou WEB dos dados tendo uma solução para organizar a nossa rede de informação, tratando também do futuro dela.

É visível que há um aumento enorme de informações e dados disponíveis na WEB, dentre elas estão em redes sociais, notícias, *Application Programming Interface* (API), privadas e públicas, Portable Document Format (PDF) e Comma Separated Values (CSV) e muitos outros meios de obtê-las, qual tais informações são na sua maioria são interpretadas por humanos, porém muitas das vezes pelo excesso, má apresentação e publicação desses dados e informações não fica tão visível, e acaba quase sempre prejudicando a consulta ou nem sendo encontrado causando alguma divergência de informação, visando então um problema não só para se gerar novas informações, mas também o mal uso ou até não uso de certos domínios que possuem informações úteis e válidas mas não semanticamente apresentável nem ligadas de maneira que possa ser de fácil acesso como os dados abertos governamentais e também não governamentais por exemplo o do metrô de São Paulo.

Segundo a definição da *Open Knowledge International*, antes conhecida como *Open Knowledge Foundation,* os "dados são abertos quando qualquer pessoa pode livremente usá-los, reutilizá-los e redistribui-los, estando sujeito a, no máximo, a exigência de creditar a sua autora e compartilhar pela mesma licença" (TCU, 2015, p. 5).

Assim esse projeto consiste em desenvolver e implementar uma aplicação que possa gerar enriquecimento semântico de dados abertos de diversas fontes a fim de ligá-los e utilizá-los, tendo como estudo de caso nesse projeto o metrô de São Paulo.

2 Justificativa

Justifica-se assim a importância deste trabalho, que apresenta uma abordagem de um aplicativo que gere enriquecimento semântico com a finalidade de não apenas a atender um domínio específico, e sim a diversas ontologias tornando-se um processo que era difícil para os profissionais da computação, uma vez que hoje esses profissionais são responsáveis pelo processo de enriquecimento semântico de dados conectados. Entretanto a abordagem se estende em aplicar o mapeamento e enriquecimento de dados utilizando como estudo de caso o metrô de São Paulo.

3 Objetivo

Apresentar um aplicativo que gere enriquecimento semântico a partir de dados abertos, possibilitando um fácil acesso e reuso desses dados, tendo como estudo de caso metrô de São Paulo.

Os objetivos específicos delimitados no desenvolvimento do trabalho:

- Apresentar a Web semântica e seus objetivos;
- Processo de extração de dados abertos;
- Processo de anotação semântica;
- Realizar o mapeamento dos dados abertos;

- Realizar o enriquecimento semântico

4 Desenvolvimento

A pesquisa e desenvolvimento para realização do projeto foi para elaborar a melhor maneira de gerar enriquecimento semântico de dados aberto com uma ontologia generalizada com base na entrada do recurso ou definindo na hora de utilizar o aplicativo que no exemplo vai ser o do metro de São Paulo.

4.1 Extração de metadados

A extração de metadados ocorre quando um algoritmo minera um recurso produzido metadados estruturados para representar esse objeto (GREENBERG, 2004). Vários métodos têm sido usados para extração automática de metadados como, por exemplo, expressões regulares (REGEX), parses, baseados em regras e aprendizado de máquina são robustos e adaptáveis, para serem usados em qualquer domínio, mas exigem um conjunto de treinamento.

Outro aspecto importante é a capacidade de classificar o documento segundo categorias pré-determinadas. Está tarefas é denominada Categorização de Texto e utiliza muitas técnicas de Aprendizado e Máquina (SEBASTIANI, 2002; YANG e LIU, 1999). As tarefas de extração e mineração de metadados são mais efetivos quando guiados por uma ontologia específica.

4.2 Anotação semântica

Uma característica do conteúdo de WEB Semântica é a disponibilização dos dados de um modo que seja passível de interpretação por máquina. Um dos meios de elevar a qualidade de um dado disponível na WEB de *human-readable* para *machine-readable* é o processo chamado anotação semântica, o qual utiliza ferramenta como RDF para permitir essa mudança nas características dos dados (JULIANO; NEWTON, 2016).

Assim o processo de anotação semântica consiste na geração de metadados, textos ou trechos, e se torna semântico quando relacionado ou ligado com certos recursos e ontologias.

Segundo UREN et al. (2006) anotação semântica pode ser descrita da seguinte maneira:

> Anotação semântica identifica formalmente conceitos e relacionamentos entre conceitos em documentos e é destinada, principalmente, a ser processada por máquinas. Por exemplo, uma anotação semântica pode relacionar “Paris” em um texto, a uma ontologia que identifica o conceito abstrato” Cidade” ao mesmo tempo que o conecta à instância “França”, do conceito abstrato “País”, eliminando, assim, qualquer ambiguidade ao que o termo “Paris” se refere.

O processo de anotação semântica pode ser manual, requer intervenção do usuário em todas as etapas, a semiautomática, que requer a intervenção humana em algumas etapas e a automática no qual necessita da utilização de técnicas de IA (Inteligência Artificial), como aprendizagem de máquina para que isso ocorra.

4.3 Mapeamento de dados abertos

Consiste em descobrir links entre as combinações semânticas dos dados e metadados com outros recursos na WEB de dados (ANGELO; MÁRCIO, 2014). Segundo Sorrentino et al.

(2013) é muito utilizado para interligar recursos na nuvem LOD (*Linked Open Data*). O mapeamento dos dados abertos representa em realizar consultas no LOD através do SPARQL utilizando um recurso ou ontologia como parâmetro para busca e tendo em seu retorno informações para realizar a criação de um novo RDF utilizando um padrão de metadados específica do tema abordado, assim podendo então realizar e ligar diretamente com dados abertos, uma vez onde já se encontra definida a ontologia.

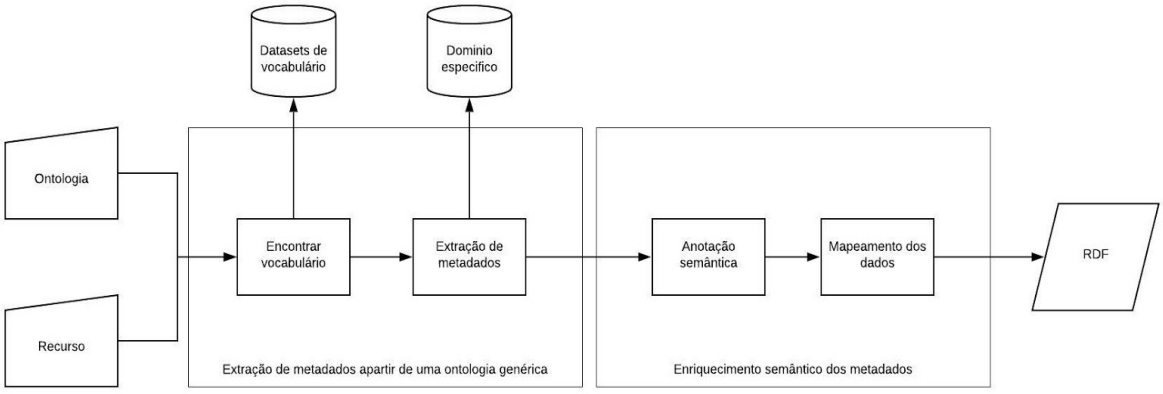
4.4 Enriquecimento semântico

Segundo Chris Clarke (2009), o enriquecimento semântico pode ser entendido como um recurso projetado para aumentar a riqueza dos dados. O enriquecimento semântico pode ser também visto como um processo de atribuir significado para os dados e metadados onde tem como objetivo facilitar a compreensão e processamento dos dados por pessoas e máquinas. Entretanto o enriquecimento semântico torna os dados e metadados mais qualificados, através do uso da semântica que pode ser vinculada a ontologias e vocabulários existentes e sinônimos.

4.5 Unindo os conceitos a nível técnico

O desenvolvimento constitui-se a partir de um modelo que representa o processo e fluxo da aplicação que irá gerar o enriquecimento semântico utilizando uma abordagem nova para ontologias dinâmicas introduzidas pelo usuário a consumir o processo que pode ser desde um desenvolvedor de aplicações a um pesquisador.

Figura 1 - Abordagem para enriquecimento semântico.



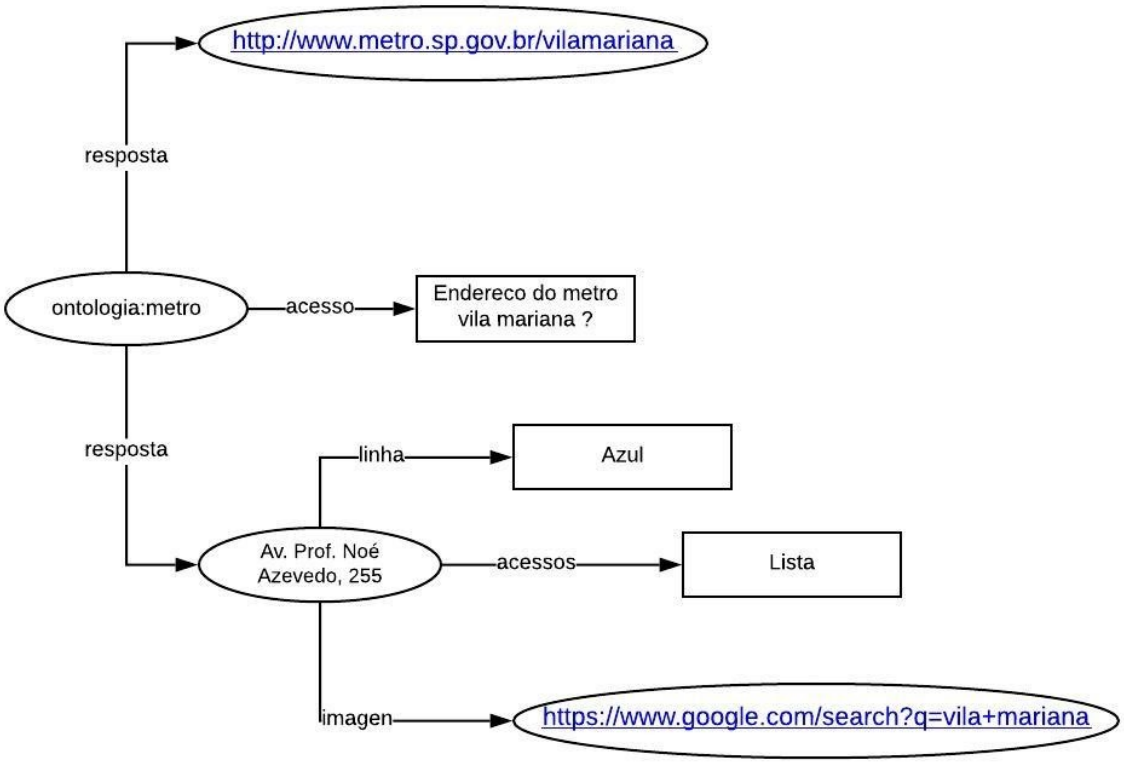
Fonte: Própria dos autores (2020).

De acordo com a figura 1 pode-se averiguar que dado a entrada inicial do usuário que seria a ontologia desejada e o recurso a ser acessado é dado o início do processo assim é realizado a classificação da ontologia buscando uma já existente ou abrindo oportunidade também da criação de uma nova visando a extração dos metadados de dados abertos que serão buscados a partir da mesma, com os dados selecionados passam para fase de anotação semântica onde é atribuído um significado aos dados e metadados encontrados facilitando a realização do mapeamento, e por fim realiza-se o processo de enriquecimento onde tendo passado por todas

etapas anteriores, tendo já um dado limpo, anotado e mapeados fica fácil gerar triplas em RDF tornando um dado rico e pronto para ser implementado.

Como resultado podemos utilizar o exemplo do metro demonstrado na figura 2 que remete um exemplo de grafo RDF gerado a partir da ontologia Metro e do recurso "Endereço do metro vila mariana", e traz como resultado uma gama de dados ligados como onde se classifica em duas respostas iniciais cujo uma é o endereço, e outros auto associados como a linha, acessos, imagens e descrições do metro.

Figura 2 - Exemplo do RDF enriquecido e mapeado.



Fonte: Própria do autor (2020).

Contudo o presente trabalho foi desenvolvido uma API (*Application Programing Interface*) desenvolvida em Python utilizando as bibliotecas RDFlib, Bottle e BeautitifulSoup que facilitaram o desenvolvimento e que ajudou para mostrar o funcionamento do enriquecimento semântico e prover os dados de maneira sutil. Para deixar o projeto mais dinâmico e aberto a possibilidade de crescimento e implementação a URI da API foi pensada e elaborada com um padrão a fim de facilitar a busca e resposta providos pelo mesmo, que seria o seguinte padrão /wiki/<ontology>/<type> onde <ontology> seria qual vocabulário deseja acessar e <type> o formato do arquivo que são em xml ou json, o final esperado da URI seria por exemplo como "/wiki/metro/json" ou "/wiki/metro/xml", a figura 3 abaixo representa em código a rota da API.

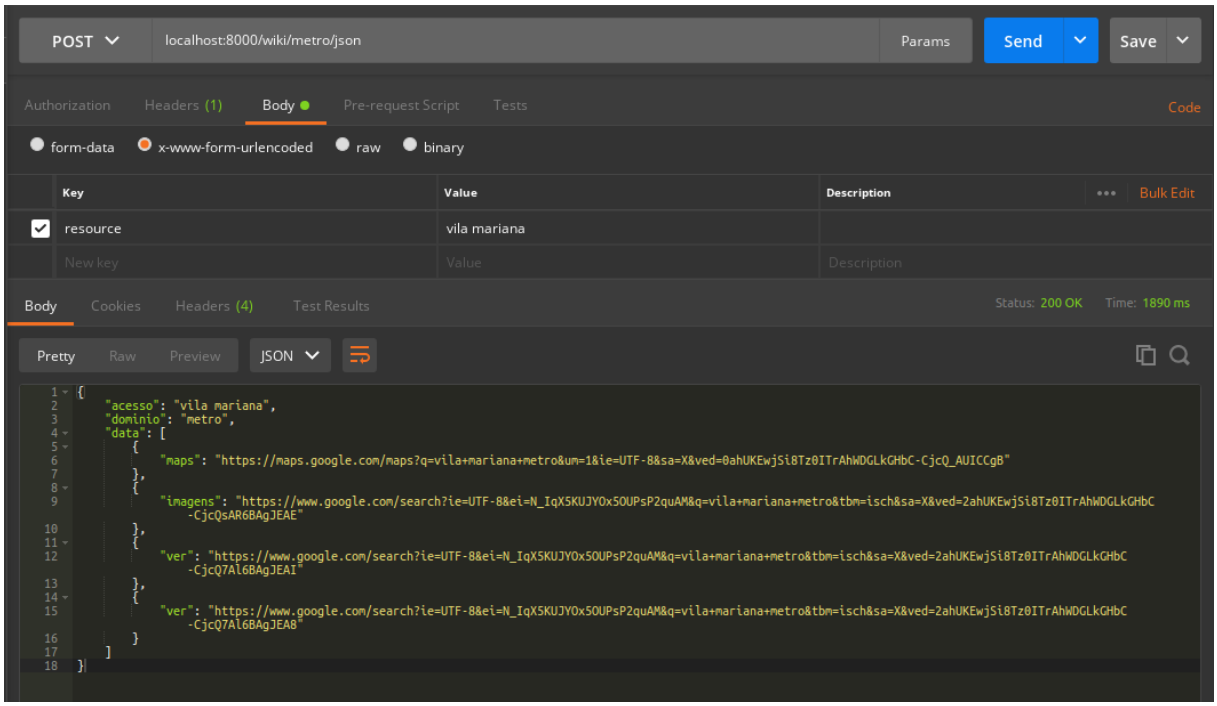
Figura 3 - Modelo de URI para API

```python
from bottle import post, run, request, response
from input_data import InputData


@post('/wiki/<ontology>/<type>')
def wiki(ontology, type):
    resource = request.forms.get('resource')
    input_data = InputData(resource, ontology)
    resp = input_data.extract_data(type)
    response.content_type = f"aplication/{type}"
    return resp


run(host='localhost', port=8000, debug=True)

```

Fonte: Própria do autor (2020).

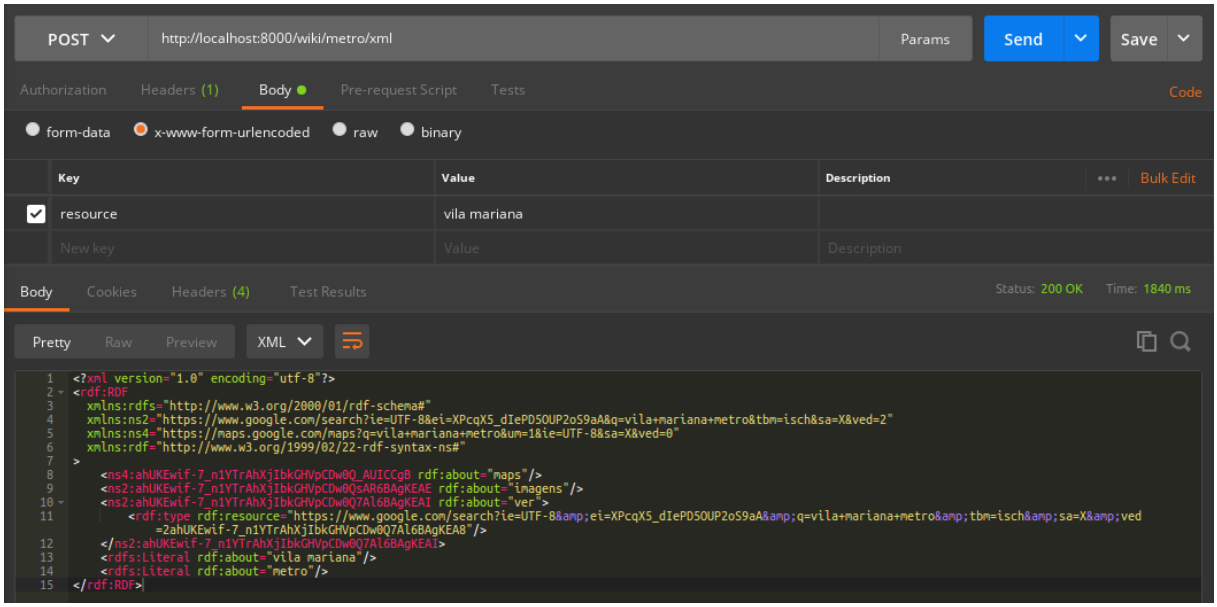
Após entrar com a URI desejada o aplicativo irá gerar uma tripla RDF contendo o máximo de informações procurada nos domínios específicos a figura 4 e figura 5 mostra um exemplo usando a ontologia metro e recurso vila mariana, utilizando um cliente HTTP para executar a requisição e tendo como resultado um JSON e XML como resposta a fim de prover os dados ligados e enriquecidos.

Figura 4 - Requisição com cliente HTTP Postman retorno em JSON.



Fonte: Própria do autor (2020).

Figura 5 - Requisição com cliente HTTP Postman retorno em XML.



Fonte: Própria do autor (2020).

5 Resultados

Como resultado temos um modelo de aplicação com grande potencial de se tornar um framework para ajudas e contribuir com o avanço e crescimento da WEB Semântica, e apresentar um fluxo onde diferente de um buscador normal ou um simples processo de busca de informação que hoje retorna milhões de resultados que muitas das vezes se torna difícil de utilizar ou de mesmo reutilizar algum dado e informação disponível como artigos, documentos, pesquisas e até do metro.

A figura 6 mostra um pouco de como o processo final está ocorrendo quando solicitada uma informação para API.

Figura 6 - Log da aplicação quando requisitado uma informação.

Fonte: Própria do autor (2020).

O log da aplicação mostrado na figura 6 é dividido de uma maneira fácil de se entender como ocorre o fluxo mostrando o recurso solicitado que no exemplo é “vila mariana” o vocabulário onde vulga ser a ontologia cujo é “metro” a junção das entradas que é representado pela busca onde ficou “vila mariana metro”, o próximo são os processos de anotação e mapeamento semântico representado por anotando e os dados que são ligados como “maps”, “imagens” e “metro” e pôr fim a tripla RDF gerada já pronta para ser consultada ou utilizada.

6 Conclusão

O projeto apresenta uma abordagem para realizar enriquecimento semântico de dados abertos onde visa utilizar ontologia genérica tornando útil em qualquer área a ser utilizado como saúde, ou simplesmente para ter um dado ligado e rico de informações trazendo uma ampla utilidade, e também em contribuir com a comunidade a fim de ajudar com trabalhos futuros, tendo como grande diferencial para um buscador comum a sua alta ligação, separação e a atribuição de significado nos dados devido à grande aumento de informações providas na WEB.

REFERÊNCIAS

ANDRADE, Jaider da Fonte. O modelo de dados resource description framework (RDF) e o seu papel na descrição de recursos Marília: 2013. Disponível em: https://periodicos.ufpb.br/ojs/index.php/ies/article/view/15436/9681.

Acesso em: 3 jun. 2020.

ANGELO, Marcio da Fonte. Uma abordagem para enriquecimento semântico de metadados para publicação de dados abertos Recife: 2014. Disponível em: https://repositorio.ufpe.br/handle/123456789/11570.

Acesso em: 18 jun. 2020.

ARAÚJO, Jáderson da Fonte. Mapeamento de banco de dados para domínios semânticos Goiana: 2009. Disponível em: https://repositorio.bc.ufg.br/tede/bitstream/tede/4639/5/Disserta%C3%A7%C3%A3o%20-%20J%C3%A1derson%20Ara%C3%BAjo%20Gon%C3%A7alves%20da%20Cruz%20-%202015.pdf.

Acesso em: 4 mai. 2020.

DARLAN, Jonathan da Fonte. Enriquecimento semântico de perfil de usuário para apoio a um modelo de aprendizagem informal no contexto da saúde Mossoró: 2015. Disponível em: https://ppgcc.ufersa.edu.br/wp-content/uploads/sites/42/2014/09/jonathan-darlan-cunegundes-moreira.pdf.

Acesso em: 14 mai. 2020.

HENRIQUE, Luiz da Fonte. Extração de metadados utilizando uma ontologia de domínio Porto Alegre: 2009. Disponível em: https://www.lume.ufrgs.br/handle/10183/22814.

Acesso em: 18 jun. 2020.

JUNIANO, Newton da Fonte. Proposta de uma ferramenta de anotação semântica para publicação de dados estruturados na Web São Paulo: 2016. Disponível em: https://sapientia.pucsp.br/bitstream/handle/18992/2/Newton%20Juniano%20Calegari.pdf.

Acesso em: 17 jun. 2020.

KARINE, Gabrielle da Fonte. Desafios e perspectivas da web semântica Medianeira: 2013. Disponível em: http://repositorio.roca.utfpr.edu.br/jspui/bitstream/1/2340/1/MD_COADS_2013_1_04.pdf.

Acesso em: 16 jun. 2020.

MEDEIROS, Magdiel da Fonte. Enriquecimento semântico da HPSG e definição de argumento como uma estrutura de traços Florianópolis: 2007. Disponível em: https://periodicos.ufpb.br/ojs/index.php/ies/article/view/15436/9681.

Acesso em: 7 mai. 2020.

SOARES, Liliane da Fonte. Enriquecimento semântico de informação geográfica voluntária utilizando linked data e tesauro Minas Gerais: 2018. Disponível em: https://www.locus.ufv.br/bitstream/handle/123456789/17963/texto%20completo.pdf?sequence=1&isAllowed=y.Acesso em: 11 mai. 2020.