Validando impacto do enriquecimento semântico: Busca com repositórios do metrô de São Paulo

GIOVANE SANTOS SILVA

Resumo

O presente trabalho apresenta a proposta de validar o uso do enriquecimento semântico com repositórios de dados abertos na Web utilizando como estudo de caso o metrô da cidade de São Paulo a fim de desenvolver e aprimorar uma aplicação capaz de prover dados ligados em grafos com as melhores rotas a partir de um ponto de referência ou palavra-chave. O metrô é um meio de transporte muito utilizado em grandes metrópoles e no Brasil, cidades como: São Paulo e Rio de Janeiro, são movidos por esse meio de locomoção tão importante. Essas informações, além de preciosas para o cotidiano das pessoas quase nem sempre é de fáceis acessos. Nesse contexto e na linha de pesquisa que envolve enriquecimento semântico de dados abertos, realizou-se uma pesquisa exploratória por meio de estudo e revisão bibliográfica, e aplicado trazendo um trabalho com método qualitativo, pois foi feito um levantamento e validação dos dados já enriquecidos, e gerando como resultados parciais o desenvolvimento aprimorado e a validação do uso de dados enriquecidos para auxiliar o desenvolvimento de futuras aplicações mais completas, afim de também poder compartilhar os dados em nuvem possibilitando a busca dos mesmos.

Palavras-chave: Metrô; Semântica; Dados abertos; Enriquecimento Semântico;

Validating the impact of semantic enrichment: Search with São Paulo metro repositories

Abstract

This work aims to validate the use of semantic enrichment with open data repositories on the Web, using the São Paulo subway as a case study, in order to develop and improve an application capable of providing data linked in graphs with the best routes from a reference point or keyword. Subway is a means of transport widely used in large metropolises and in Brazil, cities such as São Paulo and Rio de Janeiro are powered by this important means of transport. This information, as well as being invaluable to people's daily lives, is not always easy to access. In this context and in the line of research that involves semantic enrichment of open data, exploratory research was carried out by means of a study and bibliographic review, and applied, bringing a work with a qualitative method, since a survey and validation of the data already enriched was carried out, and generating as partial results the improved development and validation of the use of enriched data to assist in the development of future more complete applications, in order to also be able to share the data in the cloud making it possible to search for it.

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

1 INTRODUÇÃO

Surgida em 1991 a *World Wide Web* (www) ou simplesmente Web é popularmente tão utilizada por todos no mundo, que não é difícil no imaginário dos usuários confundida facilmente pela Internet (Karine; Gabriele, 2013). Dado esse fato, a Web é uma ferramenta usada para acessar a rede Internet através de navegadores, tais como: Chrome e Safari, exemplos mundialmente conhecidos.

Assim o desenvolvimento e crescimento da Web trouxe consigo diversos meios de comunicação em rede, ou seja, é responsável por conectar a sociedade mundial, o que acaba por permitir uma grande explosão de informações e conhecimento disponíveis e acessíveis para qualquer indivíduo. Tal advento teve como fruto a dependência de diversas ferramentas capazes de encontrar, entre inúmeras informações irrelevantes, uma que fosse precisa.

Nesse sentido, Souza e Alvarenga (2004) consideram que a dificuldade em determinar os contextos informacionais tem como consequência a impossibilidade de se identificar de forma precisa a atinência dos documentos. Além disso, as presentes tecnologias linguísticas utilizadas para desenvolvimento Web tradicional é centralizada em aspectos, como a visualização, o que, consequentemente, engloba *design* das informações, os quais agregam de maneira pobre e pouco passível de ser consumida tanto por máquinas como seres humanos.

Contudo, vale ressaltar, segundo Luz (2021) que para possibilitar uma recuperação eficiente da informação, visando suprir as necessidades do usuário em qualquer contexto, é fundamental que os dados estejam disponibilizados de forma organizada. Desse modo, surge a proposta da Web Semântica por Berners-Lee para melhorar e otimizar as pesquisas realizadas na Web, gerando uma nova versão da Web, de maneira a adicionar a semântica ao atual formato de representação de dados.

Assim, houve a integração de diversas tecnologias e aprimoramentos da Web a fim de torná-la ainda melhor. Uma dessas tecnologias foi a elaboração e utilização de ontologias atribuindo sentido e significado a determinados termos em dados, tendo por finalidade atribuir semântica aos documentos.

Partindo da dificuldade atual de prover dados enriquecidos e ligados, de fácil acesso para humanos e máquinas, surgiu o objetivo da criação de um *software* que pudesse prover essas informações de maneira eficiente seguindo os padrões necessários a partir de domínios com dados abertos.

Inicialmente tendo como caso de uso o metrô de São Paulo a fim de validar a relevância dos dados enriquecidos provando sua eficácia e motivando a contribuição para novos projetos.

2 DESENVOLVIMENTO

O processo de enriquecimento semântico, como já apresentado anteriormente, consiste em aplicar as diretrizes propostas por Luz (2021) a fim de obter um enriquecimento ainda mais limpo e eficaz na publicação de dados na Web.

Diretrizes foram criadas sobre esse processo, e vale explicar que, para cada uma delas foi descrito uma etapa a fim de mostrar como esse processo genérico ocorria quando aplicado há um domínio juntamente com a ontologia específica, gerando todo o estudo nos dados do Metrô de São Paulo.

2.1 Extração de metadados

Metadados são definidos como dados sobre dados. O termo é entendido de diferentes maneiras pelas diversas comunidades profissionais que projetam, criam, descrevem, e usam sistema de informação e recursos (Gilliland-Swetland, 2000).

Um exemplo de um objeto que possui metadados é aquilo que se pode ver em imagens, com extensões *.jpeg*, por exemplo, as quais contém pequenas informações como dispositivo que foi utilizado e localização.

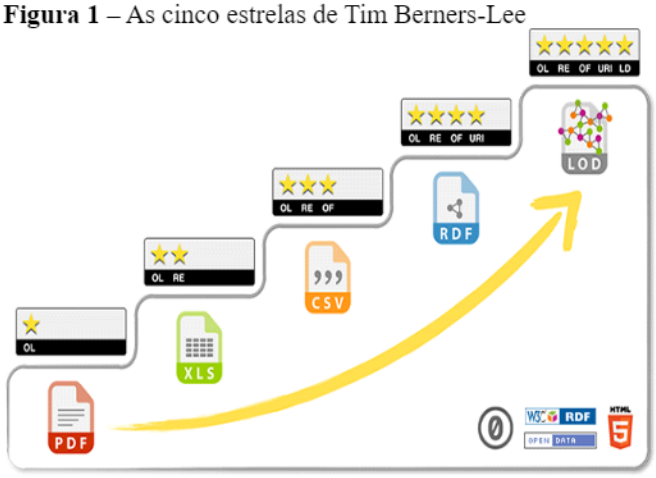
Vários têm sido os meios utilizados para extração desse metadados na Web como, por exemplo, expressões regulares (REGEX), *parses*, baseados em aprendizado de máquina (*Machine Learn*) que são robustos e adaptáveis, para serem usados em qualquer domínio, mas exigem um conjunto de treinamento.

Portanto as tarefas de extração de metadados podem ser mais eficazes quando apontadas para domínios específicos partindo de uma determinada ontologia.

2.2 Padronização dos dados

Conforme Michael (2015) Tim Berners-Lee, o inventor da Web e o primeiro a pensar em dados ligados, sugeriu um esquema de interpretação das 5 estrelas dos dados abertos como apresentado na Figura 1:

Figura 1 – As cinco estrelas de Tim Berners-Lee



Fonte: Michael; James (2015).

A primeira estrela diz para tornar seus recursos disponíveis na Web não importando seu formato sob uma licença aberta como, por exemplo aqruivos em formatos de texto e Imagens.

A segunda estrela pede para tornar os seus recursos disponíveis como dados estruturados sendo eles por exemplo uma planilha do excel.

A terceira estrela sugere para não se ultilizar formatos proprietários sendo eles por exemplo o CSV (*Comma-separated values*) ao invés do excel.

A quarta estrela diz para utilizarmos URIs fáceis e limpas para ajudar a identificar o recurso.

E por fim a quinta estrela diz para conectar os dados providos com o de outras ontologias para prover um contexto (dados ligados).

2.3 Anotação semântica

Na Web atual seus conteúdos são disponibilizados em dados de modo que humanos possam entender e, também, para elevar essa qualidade de um dado disponível na Web de *human-readable* para *machine-readable,* sendo esse um processo de anotação semântica que por sua vez utiliza a linguagem RDF, o qual tem o propósito da representação de recursos na Web (Juliano; Newton, 2016).

Assim o processo de anotação semântica consiste na geração de metadados, textos e links, e torna-se semântico quando relacionado ou ligados com demais recursos e ontologias. Segundo Uren *et al.* (2006) anotação semântica pode ser descrita da segunda maneira:

> A 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, requerendo intervenção do usuário em todas suas etapas, a semiautomática, que requer intervenção humana em algumas etapas e a automática no qual a as etapas requer a utilização de técnicas de IA (*Inteligência Artificial*) como aprendizagem de máquina para que isso ocorra.

2.4 Mapeamento dos dados abertos

Consiste em descobrir *links* entre as combinações semânticas dos dados e metadados com outros recursos na WEB de dados.

Segundo Sorrentino *et al.* (2013) é muito utilizado para interligar recursos na nuvem LOD.

Assim pode-se concluir que o processo de mapeamento de dados aberto é ligar qualquer novo dado gerado com algum outro por meio das URLs a fim de facilitar sua busca e localização.

2.4 Enriquecimento semântico

O Enriquecimento semântico, que pode ser visto como o processo de atribuir maior significado aos metadados e dados através da aplicação de recursos auxiliares, com o objetivo de facilitar a compreensão, a integração e o processamento dos dados por pessoas e máquinas. ou seja, o enriquecimento semântico torna os dados e metadados mais qualificados e descritivos, através do uso da semântica atribuída por recursos semanticamente bem-descritos, como recursos de base de conhecimento, ontologias, vocabulário existentes entre outros (Lira, 2014).

Sendo assim um processo de vai atribuir significado partindo de uma ontologia que pode ir sendo moldada de acordo com o decorrer do processo de enriquecimento já que irá aumentando sua base de conhecimento.

2.4 Diretrizes para enriquecimento semântico

Para uma boa publicação de dados na Web é preciso diversos processos capazes de atingir os princípios do *Linked Data* e baseados nas boas práticas de Publicação de dados propostos pela W3C, independente do domínio dos dados.

Assim Luz (2021) apresenta as diretrizes necessárias sobre o processo informacional para enriquecer e mapear os dados abertos semânticos, onde a solução não é apenas contextualizar um resultado com uso de um determinado vocabulário, ontologia e ferramenta adequada, mas sim abstrair e contextualizar o processo desse enriquecimento e mapeamento que se chegue ao contexto da publicação na Web.

Nesse contexto, as diretrizes são apresentadas como:

- **Diretriz 1**: Extração de Dados;
- **Diretriz 2**: Limpeza e Recuperação dos dados;

- **Diretriz 3**: Criação de colunas URIs e representação dos dados na Web de Dados.
- **Diretriz 4**: Processo de Análise de similaridade/Enriquecimento e Mapeamento Semântico; e
- **Diretriz 5**: Fusão de Dados.

2.5 Aplicando as diretrizes

A primeira fase do projeto é a coleta de dados ou como citado, extração de metadados, utilizando dá Diretrizes 1 e 2 foi preciso selecionar as maiores bases de dados abertos, dentre eles a maior base contendo esses dados foi o Portal da transparência do metrô de São Paulo contendo a maior gama de dados como horários, localização das estações, rede de funcionamento, pontos e estações sendo eles dados em qualquer uma das fases das cinco estrelas proposto por TIm Berners-Lee, estando de acordo com a ontologia de entrada sendo ela o metrô, tornando um processo automático juntamente com o seu manual correspondente.

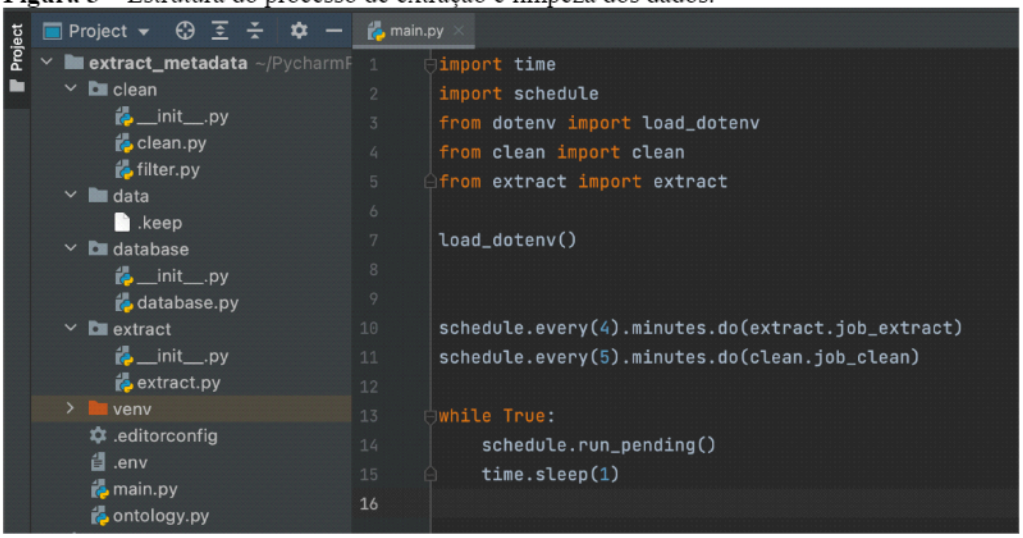
Figura 2 – Portal da transparência Metrô de São Paulo.

Fonte: Portal da transparência (2022).

Os processos de extrações foram feitos de forma automática, inteligente e agendada separada por etapas, cujo a primeira etapa foi a extração de metadados contidos na própria Web com as bibliotecas Beautifulsoup4 da linguagem Python que é utilizada para analisar documentos HTML e XML colocando e retirando informações de domínios do metrô e outros domínios públicos como Google Maps e Google Fotos a fim de futuramente ligá-los com os dados que serão enriquecidos trazendo mais informações na hora da busca.

A segunda etapa foi a padronização dos dados convertendo inicialmente todos documentos, arquivos em textos e localizações utilizando a ferramenta Tabula transformando da primeira para terceira estrela, em seguida a biblioteca Pandas responsável por manipular os novos arquivos gerados para o fim de facilitar a anotação e limpeza desses dados. Tendo já os dados eles serão limpos e separados para remover os não utilizados e o próprio poder computacional utilizando *machine learn* para classificá-los e atributos em um *database* onde será utilizado pelos próximos passos do processo. A Figura 3 demonstra o *microservice* em Python a fim de realizar a extração, limpeza e também salvar as informações para que possam ser mapeadas, enriquecidas e publicadas futuramente.

Figura 3 – Estrutura do processo de extração e limpeza dos dados.



Fonte: O próprio autor.

Tendo esses dados limpos e separados em uma *database* foi elaborado outro serviço responsável por enriquecer e prover os dados, aplicando as diretrizes 3, 4 e 5 utilizando a biblioteca rdf2go da linguagem Go Lang foi feita a ligação dos dados sendo eles informações Textuais e ou novas URIs e URLs a fim de ser convertida para modelo de Grafo utilizando em seus nós Literais de textos como descrição e outros demais tipos contidos no contexto RDF, podendo ser salvo em um outro *database* como por exemplo Neo4j, e assim provendo em modelo de API a devida informação por meio de JSON e XML para ser consumida por outras aplicações tais como uma página Web, aplicativo Mobile e até mesmo para usos estatísticos e estudos.

2.3 Resultados e discussões

Como resultado pode-se observar os logs das aplicações na Figura 4, tendo como primeiro log a parte de extração e limpeza dos dados, realizando o processo de procurar o recurso no domínio público da transparência trazendo um PDF com a linha e endereço dos metros, que por sua vez é convertido para texto afim de padronizar e limpo que de ser salvo em um banco de dados para realização da próxima etapa.

Figura 4 – Log do processo de extração e limpeza dos dados.

```
[~extract] Search in: https://transparencia.metrosp.com.br/sites/default/files/Endere%C3%A7o%20das%20Esta%C3%A7%C3%B5es.pdf
[~extract] Init extract
[~extract] Download finished. File saved in: data/address.pdf
[~clean] Init clean data
[~convert] Convert finished. File saved in: data/address.csv
[~Database] Saved register: {'line': 'LINHA 1-AZUL', 'stations': 'ESTAÇÃO', 'address': 'ENDEREÇO', 'access': 'ACESSOS'}
[~Database] Saved register: {'line': 'LINHA 1-AZUL', 'stations': 'JABAQUARA', 'address': 'R. dos Jequitibás, 80', 'access':
[~Database] Saved register: {'line': 'LINHA 1-AZUL', 'stations': 'CONCEIÇÃO', 'address': 'Av. Eng. Armando de Arruda Pereira
Armando de Arruda Pereira (lado par)\rCentro Empresarial Itaú\rR. Volkswagen'}
[~Database] Saved register: {'line': 'LINHA 1-AZUL', 'stations': 'SÃO JUDAS', 'address': 'Av. Jabaquara, 2438', 'access': 'A
[~Database] Saved register: {'line': 'LINHA 1-AZUL', 'stations': 'SAÚDE', 'address': 'Av. Jabaquara, 1634', 'access': 'Av. J
[~Database] Saved register: {'line': 'LINHA 1-AZUL', 'stations': 'PRAÇA DA ÁRVORE', 'address': 'Praça da Árvore, 39', 'acces
s'}
[~Database] Saved register: {'line': 'LINHA 1-AZUL', 'stations': 'SANTA CRUZ', 'address': 'R. Domingos de Morais, 2564', 'ac
do ímpar)'}
[~Database] Saved register: {'line': 'LINHA 1-AZUL', 'stations': 'VILA MARIANA', 'address': 'Av. Prof. Noé Azevedo,  255', '
 - Av. Lins de Vasconcelos\rAv. Lins de Vasconcelos - Terminal Urbano\rR. Domingos de Morais - R. Madre Cabrini'}
[~Database] Saved register: {'line': 'LINHA 1-AZUL', 'stations': 'ANA ROSA', 'address': 'R. Domingos de Morais, 505', 'acces
```

Fonte: O próprio autor.

3 CONSIDERAÇÕES FINAIS/CONCLUSÃO

Considerando as análises e validações dos resultados apresentados neste trabalho, conclui que os dados enriquecidos trazem uma informação que por sua vez era de difícil acesso e mal utilizada, comprovando sua eficiência e possibilitando uma nova informação mais precisa e rica de acordo com a ontologia específica, que neste trabalho foi utilizada a do Metrô de São Paulo, assim fornece resultados capazes da criação e exposição novamente da informação na Web ou em um aplicativo mobile de maneira atrativa para o ser humano e sendo possível também o seu consumo por máquinas, realizando assim também o cumprimento das cinco estrelas de Tim Berners-Lee tornando a Web cada vez mais semântica tendo dados ligados de maneira eficiente.

Dessa forma foi implementado as diretrizes propostas por Luz (2021) em seu *Framework*, assim foi construído a aplicação em uma arquitetura de *microservices* sendo eles um processo de agendamento construído na linguagem Python onde há de salvar os dados em uma base própria e uma API construída na linguagem GO a fim de completar todo ciclo e salvar os grafos provendo dados enriquecidos atingindo assim todos os objetivos propostos.

Na intenção de dar continuidade e desdobramento continuado dos estudos e pesquisa sobre a temática. Este trabalho é perceptível à futuras implementações em outros segmentos como também deixando a possibilidade de aperfeiçoamento técnico e teórico.

REFERÊNCIAS

ANDRADE, Jaider da Fonte. O modelo de dados resource description framework (RDF) e o seu papel na descrição de recursos, Marília, 2012. Disponível em: https://periodicos.ufpb.br/ojs/index.php/ies/article/view/15436/9681. Acesso em 20 abr. 2022.

CLARKE, C. da Fonte. A resource list management tool for undergraduate students based on linked open data principles, Berlin, 2009. Disponível em: https://link.springer.com/chapter/10.1007/978-3-642-02121-3_51. Acesso em 21 mai. 2022.

COSTA, L. S. da Fonte. Enriquecimento semântico de informação geográfica voluntária usando linked data e tesauro, 2018. Disponível em: https://www.locus.ufv.br/bitstream/123456789/17963/1/texto%20completo.pdf. Acesso em 21 mai. 2022.

GILLILAND-SWETLAND, da Fonte. Introduction to Metadata. Setting the stage. Introduction to Metadata: Pathways to Digital Information, Los Angeles, 2000. Disponível em: https://www.getty.edu/research/institute/standards/intrometadata/. Acesso em 14 mai. 2022.

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 07 mai. 2022.

KARINE, Gabrielle da Fonte. **Desafios e perspectivas da Web semântica**. Medianeira 2013. Disponível em: http://repositorio.utfpr.edu.br/jspui/handle/1/13453. Acesso em 09 mai. 2022.

LIRA, Márcio 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 21 mai. 2022.

LUZ, L, P. da Fonte. Framework para publicações de dados com ênfase em enriquecimento e mapeamento semântico, Marília, 2021. Disponível em: https://repositorio.unesp.br/bitstream/handle/11449/204711/luz_lp_dr_mar.pdf?sequence=3&isAllowed=y. Acesso em 26 jan. 2022.

MICHAEL; JAMES da Fonte. As 5 estrelas dos dados abertos, 2015. Disponível em: https://5stardata.info/pt-BR/. Acesso em 14 mai. 2022.

OLIVEIRA, Luis, da Fonte. Extração de Metadados utilizando uma ontologia de domínio, Porto Alegre, 2009. Disponível em: https://www.getty.edu/research/institute/standards/intrometadata/. Acesso em 21 mai. 2022.

PICKLER, Maria da Fonte. Web Semântica: Ontologias como ferramentas de representação do conhecimento, Londrina, 2007. Disponível em: https://www.scielo.br/j/pci/a/HHdw6KMPG45HxwShcwTmFSs/?format=pdf&lang=pt. Acesso em 12 jan. 2022.

SILVA, G, S.; LUZ, L, P. da Fonte. Aplicativo que gera enriquecimento semântico de dados abertos, Garça, 2020. Disponível em: https://pesquisafatec.com.br/ojs/index.php/efatec/article/view/206. Acesso em 09 abr. 2022.

SORRENTINO, S. et al. Semantic annotation and publication of linked open data. in 13th international Conference, ICCSA 2013, Ho Chi Minh City, Vietnam, 2013. p. 462-474.

SOUZA, Renato Rocha; ALVARENGA, Lídia da Fonte. A Web Semântica e suas contribuições para a ciência da informação, Brasília, 2004. Disponível em: https://doi.org/10.1590/S0100-19652004000100016. Acesso em 09 mai. 2022.

UREN, V. et al. Semantic annotation for knowledge management. Journal of Web Semantics, v. 4, p. 14-28, 2005.