IV FORPED PPGGOC

FÓRUM DE PESQUISAS DISCENTES DO PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO

INFORMAÇÃO, TECNOLOGIA E PESSOAS: TRANSFORMAÇÕES E HUMANIDADES DIGITAIS

BELO HORIZONTE - 16 A 18 DE AGOSTO DE 2023

ISSN 2965-4068

Modalidade: Trabalho Completo

MODELO DE ENRIQUECIMENTO SEMÂNTICO DE DADOS CORPORATIVOS: uma proposta de desenvolvimento

SEMANTIC ENRICHMENT MODEL OF CORPORATE DATA: a development proposal

Claudiane Emanuele Nazário [1]
Universidade Federal de Minas Gerais

Célia da Consolação Dias [2]
Universidade Federal de Minas Gerais

RESUMO

Neste trabalho é apresentado um esboço do planejamento de um estudo preliminar, cujo objetivo principal é propor um modelo de enriquecimento semântico de dados corporativo para publicação na web, utilizando os princípios do *Linked Data*. A metodologia baseou-se na pesquisa aplicada exploratória, de caráter qualitativo, compostas pelas etapas de estudo bibliométrico, levantamento e análise de dados, identificação de modelos e tecnologias de enriquecimento semântico, análise da proposta metodológica, desenvolvimento e aplicação do modelo em uma micro empresa. O modelo a ser desenvolvido baseia-se em uma proposta metodológica desenvolvida anteriormente, assim como em tecnologias e métodos disponíveis na web para a realização do enriquecimento semântico.

Palavras-Chave: Enriquecimento Semântico; Linked Data; Organização do Conhecimento.

ABSTRACT

This work presents an outline of the planning of a preliminary study, whose main objective is to propose a model for semantic enrichment of corporate data for publication on the web, using the principles of Linked Data. The methodology was based on exploratory applied research, of a qualitative nature, composed of the stages of bibliometric study, data collection and analysis, identification of models and semantic enrichment technologies, analysis of the methodological proposal, development and application of the model in a micro company. The model to be developed is based on a previously developed methodological proposal, as well as on technologies and methods available on the web for carrying out semantic enrichment.

Keywords: *Semantic Enrichment; LinkedData; Knowledge Organization.*

[1] Mestranda pelo Programa de Pós-Graduação em Gestão e Organização do Conhecimento da Escola de Ciência da Informação da Universidade Federal de Minas Gerais. Orcid: https://orcid.org/0000-0002-3746-0655. E-mail: cnazario@gamarratecnologia.com.br.

[2] Docente do Programa de Pós-Graduação em Gestão e Organização do Conhecimento da Escola de Ciência da Informação da Universidade Federal de Minas Gerais. Orcid: https://orcid.org/0000-0003-0891-6454. E-mail: celiadias@eci.ufmg.br.

1 INTRODUÇÃO

Diversas empresas, órgãos de governos e institutos de pesquisa têm empreendido esforços para disponibilizar e produzir tecnologias web voltadas para a produção e consumo de dados, com o objetivo de facilitar a descoberta de novos conhecimentos e agregar valor a qualquer informação disponibilizada na Internet. (ISOTANI e BITTENCOURT, 2015).

Devido ao excesso de informações disseminadas diariamente na web, e da ausência de padrões e princípios para a publicação destes dados, nem sempre os usuários que precisam destas informações e mesmo algoritmos de busca sofisticados conseguem localizá-los.

As necessidades provenientes da disponibilização destas informações, fizeram com que Tim Bernes-Lee em 2006, propusesse um conjunto de boas práticas para publicação de dados, denominados *Linked Data,* cujo objetivo é facilitar a integração de dados de diferentes fontes, de forma a torná-los compreensíveis tanto por homens, quanto por máquinas (BIZER; HEATH; BERNERS-LEE, 2009).

O *Linked Data* estabelece um conjunto de padrões que permite a interoperabilidade de dados de diferentes fontes, ampliando as possibilidades de busca e recuperação de informações por humanos e computadores. No entanto, para obter todas as vantagens do *Linked Data* é necessária a utilização de sistemas de organização do conhecimento tais como ontologias e vocabulários controlados durante o processo de enriquecimento e publicação destes dados.

Apesar da relevância do *Linked Data* para a disseminação e organização do conhecimento, a bibliografia brasileira acerca deste tema na área de Biblioteconomia e Ciência da Informação (BCI) parece ser restrita. Durante a etapa de levantamento bibliográfico, alguns estudos referentes a este tema foram identificados dentre os quais podemos citar:

a) Lóscio; Burle; Calegari (2017) descreveram boas práticas para a produção e disponibilização de dados abertos no contexto do Linked Open Data;
b) Martins (2019) propôs um modelo conceitual de ecossistema semântico de informações corporativas para aplicação em objetos multimídia;

c) Soergel(2018) definiu um modelo conceitual de representação de objetos informacionais dinâmicos (Universal Documento Model) que adota alguns princípios de Linked Data;
d) Nazário e Dias (2018) desenvolveram uma proposta de metodologia para avaliar o enriquecimento semântico de objetos publicados na web através do Linked Data.

O propósito deste artigo é o de apresentar, de forma resumida, o planejamento do projeto de pesquisa de doutorado, ainda em fase inicial, cujo objetivo é desenvolver um modelo de enriquecimento semântico de dados corporativos para sua publicação, baseado na proposta metodológica desenvolvida por NAZÁRIO e DIAS (2018) utilizando o *Linked Data.*

A criação de modelos de enriquecimento semântico de objetos para publicação em *Linked Data*, podem facilitar a divulgação de dados empresariais de forma precisa e padronizada, viabilizando sua recuperação e reutilização no contexto corporativo.

Para atingir seus objetivos, o presente artigo está dividido em cinco seções, a saber: A primeira seção contém os aspectos introdutórios e objetivos da pesquisa. A segunda seção é composta da exploração preliminar da literatura, contendo o referencial teórico, em suporte ao estudo apresentado. A seção três apresenta parte da proposta metodológica que será utilizada para elaboração deste estudo. Já a seção quatro apresenta a metodologia da pesquisa. Na última seção, são descritas as considerações finais e na sequência estão as referências bibliográficas usadas para a construção deste artigo.

2 REFERENCIAL TEÓRICO

Como relatado neste artigo, a pesquisa se encontra em fase inicial, portanto, foram avaliadas publicações que abordaram os temas principais que circundam o objeto deste estudo, sendo eles: **Web Semantica**, ***Linked Data*** e **Enriquecimento Semântico.** As seções a seguir descrevem estes conceitos.

2.1 Web Semântica

A Web Semântica tem como objetivo principal disponibilizar dados estruturados, relacionados semanticamente entre si, visando facilitar a reutilização de informação bem como a descoberta de novos conhecimentos (BREITMAN, 2005).

O conceito da web semântica foi proposto por Berners-Lee et al (2001) para manipular o conteúdo web, utilizando a estruturação semântica de dados, de modo a permitir o processamento automático de informações pelos agentes inteligentes.

A web semântica é uma extensão da web tradicional na qual as páginas são enriquecidas por meio de uma estrutura semântica, visando sua compreensão tanto por agentes humanos, quanto por agentes de software.

Associado ao conceito da Web Semântica, surgiu o conjunto de práticas propostas por Tim Bernes Lee (2006) denominado de “Linked Data”, cujo objetivo é o de disseminar dados de forma estruturada na web, permitindo a conexão e interoperabilidade entre sistemas e dados de diferentes fontes.

2.2 *Linked data*

O conjunto de princípios e boas práticas propostas no Linked Data que podem ser assim resumidos: a) utilizar URI para identificar e denominar os recursos da web; b) adotar URI HTTP para que os usuários possam recuperar esses recursos; c) disponibilizar informação útil, por meio dos padrões RDF e SPARQL quando alguém procurar uma URI; d) incluir links para outras URIs, para localizar novos recursos.

Grisoto (2016) afirma que o conjunto de dados publicados no *Linked Data* utilizam tecnologias da Web Semântica e padrões de metadados para descrever e representar as informações, favorecendo seu uso por homens e computadores.

O enriquecimento semântico de dados associado ao *Linked Data* fornece um conjunto de tecnologias e padrões que possibilita a publicação de dados, de forma que sejam expressivos para aplicações computacionais, permitindo sua recuperação, interoperabilidade e reutilização.

2.3 Enriquecimento Semântico

O enriquecimento semântico também pode ser entendido como um processo de atribuição de maior significado aos dados e metadados, tornando os mesmos mais qualificados, pelo uso da semântica atribuída por vocabulários pré-existentes, sinônimos e informações de proveniência, com o objetivo de facilitar a compreensão, a integração e o processamento dos dados por pessoas e máquinas. (LIRA 2014).

O enriquecimento de dados é um conjunto de processos que podem ser usados para aprimorar, refinar ou melhorar dados brutos ou processados anteriormente. Esse processo contribui para tornar os dados um ativo valioso para qualquer empresa. (LÓSCIO; BURLE; CALEGARI, 2017).

Segundo BARRÉRE, et al (2020) os dados devem ser disponibilizados, utilizando um padrão compreensível por humanos e máquinas, adequado ao domínio do conhecimento a ser publicado e em quantidade suficiente para representar seu conteúdo e contexto, visando facilitar a descoberta, uso, reutilização, compreensão e processamento dos dados.

3 MODELOS PARA ENRIQUECIMENTO SEMANTICO

3.1 Universal Documento Model

Soergel (2018) propôs um modelo conceitual de representação de objetos informacionais dinâmicos denominado Universal Documento Model (UDM), onde os dados dos documentos são decompostos em unidades documentais menores e armazenados em bases de dados hipermídia. O modelo proposto por Soergel pode ser utilizado tanto para representar a estrutura, quanto o conteúdo de um documento.

O Soergel ressalta que a estrutura proposta em seu modelo pode ser reconhecida automaticamente por agentes de softwares, criada por indexação automática ou mesmo enriquecimento semântico. O autor também reitera que seu modelo é conceitual e genérico, devendo ser aperfeiçoado por outras contribuições que possam melhorar seu funcionamento.

3.2 Modelo de Ecossistema semântico de informações corporativas

Martins (2019) propôs um modelo utilizando como base o Universal Document Model UDM, proposto por D. Soergel, com o objetivo de estabelecer um ecossistema semântico de informações corporativas.

O modelo proposto pretende ser implementável e adaptável a vários tipos de organizações empresariais, na medida em que considera os elementos gerais observados no ambiente corporativo tais como: atores, processos, áreas ou departamentos,

metadados. Esse modelo também propõe a seleção e utilização de vocabulários semânticos que podem ser adaptados, substituídos ou acrescidos de novos vocabulários para atender as demandas dos usuários.

Ainda segundo o autor, o modelo proposto não considera todos os aspectos da gestão da informação corporativa, priorizando o tratamento de objetos de informação não estruturados, como multimídia. através de decomposição e enriquecimento semântico de objetos de informação

3.3 Proposta Metodológica para avaliar o enriquecimento semântico de objetos

A proposta metodológica elaborada por NAZARIO e DIAS (2018) teve como base o desenvolvimento de uma Matriz de Técnicas e Recursos para o Enriquecimento Semântico de Objetos denominada Matriz Treso, na qual foram apresentados sete critérios para avaliação dos modelos de dados EDM e BIBFRAME. São eles:

1. Utilização de recursos de anotação semântica para o enriquecimento de objetos publicados em *Linked Data.*
2. Reuso de metadados para facilitar as atividades de publicação de dados em *Linked Data*, otimizando o trabalho do publicador.
3. *Links* entre as combinações semânticas dos dados e metadados com outros recursos da web.
4. Modelagem de dados num formato semântico estruturado, permitindo sua manipulação por aplicações que consomem esses modelos de dados, ampliando as possibilidades de conexão com outros *Datasets* do LOD.
5. Utilização de ferramentas para o enriquecimento semântico.
6. Utilização de interface gráfica para apoiar a execução do processo de enriquecimento semântico e publicação de dados.
7. Relações de sinonímia (equivalência), associação e hierarquia entre o metadado e o termo correspondente em outros vocabulários utilizados.

A proposta também contemplou um formulário de avaliação e parâmetros para pontuação de cada um dos modelos nos critérios avaliados.

Os artefatos desenvolvidos na proposta de Nazário e Dias (2018), bem como os modelos desenvolvidos por Soergel e Martins serão utilizados para o estabelecimento de um

modelo de enriquecimento semântico de dados corporativos para sua publicação na web utilizando o Linked Data.

4 METODOLOGIA

Para o desenvolvimento do modelo será adotada a pesquisa aplicada exploratória, de caráter qualitativo. (GIL,1994; LAKATOS e MARCONI, 1991). Como Procedimentos técnicos será realizada a associação de pesquisa bibliográfica e documental, além de estudo de caso a ser realizado na Empresa Gamarra Tecnologia, testes e validação. A figura 1 apresenta as etapas da pesquisa.

Figura 1 – Etapas da metodologia da pesquisa.

Fonte: Elaborado pelo Autor, 2023

1. **Estudo bibliométrico e levantamento bibliográfico -** A primeira etapa, já em andamento, se refere ao estudo bibliométrico por meio de buscas a artigos científicos, livros, teses e dissertações que versam sobre o tema desta investigação, nos idiomas português, inglês e espanhol, , Tais como Journal of Knowledge Management; Knowledge Management Research Practice; Perspectivas em Ciência da Informação, dentre outros, com objetivo de conhecer o estado da arte dos principais conceitos que nortearão essa pesquisa.

As pesquisas foram realizadas durante o período de março a junho de 2023, e consideraram principalmente publicações dos últimos dez anos. A exceção no recorte temporal se deve ao fato de que vários estudos sobre os temas *Linked Data* e *Web Semântica*, foram encontrados no período de 2001 a 2010.

O levantamento das informações iniciou-se com o estabelecimento das palavras-chave “Enriquecimento Semântico”, “Organização do Conhecimento,” “Linked Data” e “Web Semântica” em todos os idiomas já citados na pesquisa. Além disso, foram estabelecidas como domínios de conhecimento para realização das buscas as áreas: “Biblioteconomia e Ciência da Informação”, “Ciência da Computação” e Filosofia.

Como esta etapa ainda se encontra em fase de desenvolvimento, os resultados desta pesquisa serão objeto de novos artigos.

2. **Leitura e análise das informações** – essa etapa consiste na **l**eitura e fichamento das fontes selecionadas, organização lógica das informações, para embasar a discussão teórica dos diferentes pontos de vista identificados na literatura sobre o tema.
3. **Identificação de tecnologias e modelos para o enriquecimento semântico de dados** – Nessa etapa serão analisados os modelos existentes para identificar padrões, pontos positivos e negativos dos modelos em questão e sua aplicabilidade para uso no contexto corporativo. Também serão avaliadas tecnologias que podem ser utilizadas neste processo.
4. **Análise da Proposta Metodológica** - Esta etapa consiste na análise da Proposta de Metodologia para Avaliar o Enriquecimento Semântico de Objetos Publicados na Web Através do Linked Data desenvolvida por Nazário (2018), a partir da literatura recente, para identificar pontos fortes e fracos, requisitos e demandas de um modelo de dados.
5. **Elaboração do Modelo de Dados** - Construção do modelo de enriquecimento semântico de dados corporativos para publicação de objetos publicados na web através do Linked Data.
6. **Aplicação do modelo** - Teste no contexto corporativo a ser realizado na empresa Gamarra selecionada durante a execução da pesquisa.

5 CONSIDERAÇÕES FINAIS

Este artigo apresenta a análise inicial dos dados que estão sendo utilizados para o desenvolvimento de uma pesquisa de doutorado, que visa o desenvolvimento de um modelo de enriquecimento semântico de dados corporativos para sua publicação em *Linked Data*.

Entende-se que o modelo a ser desenvolvido neste projeto de pesquisa, terá relevante contribuição para o processo de enriquecimento semântico de dados corporativos, na medida em que sua construção será baseada tanto em requisitos e recursos tecnológicos disponíveis para utilização, quanto de métodos e técnicas difundidas na BCI para Organização do Conhecimento.

Espera-se que o modelo a ser desenvolvido no âmbito dessa pesquisa possa favorecer o enriquecimento semântico de informações empresariais, e que a aplicação a ser realizada na Empresa Gamarra Tecnologia, possa demonstrar sua contribuição para o processo de organização e representação do conhecimento corporativo.

REFERÊNCIAS

BARRÉRE, E. et al. Utilização de Enriquecimento Semântico para a Recomendação Automática de Videoaulas no Moodle. **Revista Brasileira de Informática na Educação**, [S.l.], v. 28, p. 319-334, maio 2020. ISSN 2317-6121. Disponível em: http://ojs.sector3.com.br/index.php/rbie/article/view/v28p319. Acesso em: 06 nov. 2022.

BERNERS-LEE, T. Linked data-design issues (2006). **URL http://www. w3. org/DesignIssues/LinkedData. html**, v. 10, n. 11, 2011.

BIZER, C.; HEATH, T.; BERNERS-LEE, Tim. Linked data-the story so far. **Semantic services, interoperability and web applications: emerging concepts**, p. 205-227, 2009. Disponível em: https://books.google.com.br/books?hl=pt-BR&lr=&id=tP8HLETgbKcC&oi=fnd&pg=PA205&dq=Linked+Data:+Design+issues&ots=-hJuToD1BB&sig=dT6uRAuVjm_XB-ZiCWa-9EWLmP4#v=onepage&q=Linked%20Data%3A%20Design%20issues&f=false. Acesso em: 06 nov. 2022.

BREITMAN, K. **Web Semântica: A Internet do Futuro**. Rio de Janeiro: LTC, 2005

GRISOTO, A. P. **Um estudo acerca dos recursos audiovisuais no contexto do Linked data**. 2016. Dissertação (Mestrado em Ciência da Informação) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2016.

ISOTANI, S.; BITTENCOURT, I. I. **Dados Abertos Conectados: Em busca da web do Conhecimento**. Novatec Editora, 2015.

LAKATOS, E. M.; MARCONI, M. A. **Fundamentos de metodologia científica**. 3ed. São Paulo: Atlas, 1991.

LÓSCIO, B. F.; BURLE, C.; CALEGARI, N. (Ed.). **Data on the web best practices.** 2017. Disponível em: https://www.w3.org/TR/dwbp/. Acesso em: 26 out. 2022.

LIRA, M. A. B. **Uma Abordagem Para Enriquecimento Semântico de Metadados Para Publicação de Dados Abertos**." (2014).

MARTINS, Sergio de Castro. **Modelo conceitual de ecossistema semântico de informações corporativas para aplicação em objetos multimídia**. 276f. Tese (Doutorado em Ciência da Informação) – Universidade Federal Fluminense, Niterói. Programa de Pós-Graduação em Ciência da Informação, Niterói, 2019.

NAZÁRIO, C. E; DIAS, C. C. **Proposta de metodologia para avaliar o enriquecimento semântico de objetos publicados na web através do Linked Data**. 2018. 138f. Dissertação (Mestrado em Gestão e Organização do Conhecimento) –Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2018.

SOERGEL, D.; ESCUDERO, F. **Toward a Universal Document Model for Active Knowledge.** Paper, November 10, 2018. Disponível em: http://digital.library.unt.ed /ark:/67531/metadc1393843/. Acesso em: 06 nov. 2022.