XIX encontro nacional de pesquisa em ciência da informação
ENANCIB
// SUJEITO INFORMACIONAL E AS PERSPECTIVAS ATUAIS EM CIÊNCIA DA INFORMAÇÃO. //
22-26 OUTUBRO 2018
LONDRINA/PR
Universidade Estadual de Londrina

XIX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2018

GT 02 - Organização e Representação do Conhecimento

PROPOSTA PARA AVALIAÇÃO DAS TÉCNICAS E DOS RECURSOS PARA O ENRIQUECIMENTO SEMÂNTICO DE OBJETOS PUBLICADOS EM LINKED DATA

Claudiane Emanuele Nazario (Universidade Federal de Minas Gerais)

Dra. Célia da Consolação Dias (Universidade Federal de Minas Gerais)

PROPOSAL FOR THE EVALUATION OF TECHNIQUES AND RESOURCES FOR THE SEMANTIC ENRICHMENT OF OBJECTS PUBLISHED ON LINKED DATA

Modalidade da Apresentação: Pôster

Resumo: O presente artigo apresenta a proposta de uma Matriz de Técnicas e Recursos para o Enriquecimento Semântico de Objetos - Matriz TRESO, desenvolvida na pesquisa cientifica do mestrado acadêmico, com o objetivo de investigar as contribuições de dois modelos de dados para o enriquecimento semântico de objetos publicados na web através do *Linked Data.* Durante aplicação da Matriz nos dois modelos de dados avaliados, foi possível verificar que que a referida matriz poderia ser utilizada para apoiar no desenvolvimento e na avaliação de outros modelos dados. Esse artigo descreve como os sete critérios propostos pela matriz podem ser utilizados no desenvolvimento, avaliação e monitoramento de modelos de dados que tenham como objetivo publicar seus dados em linked data. Como procedimentos metodológicos foram realizados o levantamento bibliográfico, visando identificar na literatura quais os requisitos necessários para realizar o enriquecimento semântico de objetos e quais as técnicas envolvidas neste processo. De posse destes critérios foi elaborada a Matriz TRESO e definidos os parâmetros para a avaliação dos modelos de dados em cada um dos critérios. Como resultado do trabalho para cada um dos critérios são sugeridas recomendações que devem ser seguidas pelos modelos de dados para potencializar o processo de enriquecimento semântico.

Palavras-Chave: Enriquecimento Semântico; *Linked Data*; Web Semântica.

Abstract: This article presents the proposal of a Matrix of Techniques and Resources for the Semantic Enrichment of Objects - Matrix TRESO, developed in the scientific research of the academic masters, with the objective of investigating the contributions of two data models for the semantic enrichment of published objects on the web through Linked Data. During the application of the Matrix in the two data models evaluated, it was possible to verify that said matrix could be used to support the development and evaluation of other models. This article describes how the seven criteria proposed

by the matrix can be used in the development, evaluation and monitoring of data models that aim to publish their data in a linked data. As methodological procedures were carried out the bibliographical survey, aiming to identify in the literature the requirements necessary to perform the semantic enrichment of objects and which techniques involved in this process. With these criteria, the TRESO Matrix was elaborated and parameters were defined for the evaluation of the data models in each of the criteria. As a result of the work for each of the criteria are suggested recommendations that should be followed by the data models to potentiate the process of semantic enrichment.

Keywords: semantic enrichment; Linked Data; Semantic Web.

1 INTRODUÇÃO

A Web Semântica tem como objetivo facilitar o compartilhamento de informações pelos usuários, atribuindo significado ao conteúdo existente. Diferente da web tradicional onde as páginas são interligadas por meio de hiperlinks, a Web Semântica emprega tecnologias como o RDF (*Resource Description Framework) e OWL* (Web Ontology Language*)* para publicar dados estruturados na web. (BATISTA; LOSCIO, 2013).

Um dos desafios para a implementação da Web semântica está em identificar as relações existentes entre os vários recursos disponibilizados na internet, para que possam ser codificados, manipulados e disseminados por computador.

Alinhado aos conceitos demandadas pela *Web Semântica*, surgiu um conjunto de práticas propostas por Tim Bernes Lee em 2006 denominado *Linked Data*, cuja proposta consiste em interligar dados por meio de *links* semânticos significativos também para programas de computador, de modo a automatizar tarefas e conexões antes possíveis somente por humanos.

Esse artigo descreve parte da pesquisa cientifica realizada no mestrado acadêmico em Ciência da Informação, cujo objetivo consistiu em propor uma metodologia para avaliar o enriquecimento semântico de objetos publicados em *Linked Data*.

Para atingir tal objetivo foi desenvolvido nesse estudo uma Matriz de Técnicas e Recursos para o Enriquecimento Semântico (Matriz TRESO). Essa matriz foi criada para verificar como os modelos de dados investigados atendem aos critérios de enriquecimento semântico encontrados na literatura.

Este artigo está organizado conforme segue: introdução, conceitos de enriquecimento semântico, *Linked Data.* Em seguida são apresentados os critérios da Matriz TRESO e as recomendações para o desenvolvimento de modelos e finalmente, a apresentação das conclusões e indicação de direções para trabalhos futuros.

2 PROPOSTA PARA AVALIAÇÃO DO ENRIQUECIMENTO SEMÂNTICO

Para avaliar como os modelos realizavam o enriquecimento semântico de objetos para a publicação em *Linked Data* foi desenvolvida a Matriz TRESO, a partir dos conceitos e requisitos propostos na literatura apresentados nas seções a seguir.

2.1 Enriquecimento Semântico

Para estruturar uma proposta metodológica foi necessário entender os conceitos e requisitos necessários para realizar o enriquecimento semântico de objetos para sua publicação em *Linked Data*.

De acordo com Lira (2014) o enriquecimento semântico é um processo de atribuição de maior significado aos dados e metadados, tornando os mesmos mais qualificados, através do uso da semântica atribuída por vocabulários pré-existentes, sinônimos e informações de proveniência.

Uren et. Al (2006) formularam sete requisitos para serem considerados pelas ferramentas de anotação semântica, sendo: 1) Formatos padronizados; 2) usuário centrado / projeto colaborativo; 3) suporte a ontologia; 4) suporte de documentos de formatos heterogêneos; 5) atualização de documentos; 6) armazenamento de anotações; 7) automação.

Silva, 2014 apresenta vários modelos de anotação semântica, incluindo *tags*, atributos, relações e ontologias, bem como as vantagens e desvantagem de cada tipo de anotação.

A seção a seguir apresenta os conceitos *Linked Data* que foram utilizados como referência para a construção da matriz TRESO.

2.2 Linked Data

As necessidades advindas da evolução da web, como manipular um grande volume de dados e facilitar o acesso à informação, fizeram com que Tim Bernes-Lee em 2006, recomendasse um conjunto de princípios e boas práticas para publicação de dados na web. Estes princípios denominados *Linked Data* têm como objetivo fundamental facilitar a integração de dados de diferentes fontes, de forma a torná-los compreensíveis também para as máquinas (BIZER; HEATH; BERNERS-LEE, 2009).

As oportunidades de integração semântica dos dados na web motivam o desenvolvimento de novos tipos de aplicação e de ferramentas, como navegadores e motores de busca (ISOTANI; BITTENCOURT, 2015).

Para permitir qualificar e enriquecer semanticamente os objetos, suas representações digitais e suas diferentes relações é necessário o desenvolvimento de um vocabulário específico – uma ontologia – de classes de objetos e processos e de relações entre estes. (MARCONDES, 2012).

2.3 Critérios para o Enriquecimento Semântico de Objeto.

A seguir são apresentados os sete critérios para o enriquecimento semântico de objetos avaliados na Matriz TRESO (figura 1)

Critério 1 - Anotação semântica

Esse critério se refere ao modo com que os modelos devem utilizar o recurso de anotação semântica para atribuir significado aos recursos. A recomendação neste caso, é que os modelos sempre que possível façam uso de vocabulários controlados e ontologias, para facilitar a compreensão do usuário e permitir a reutilização e compartilhamento destas anotações por sistemas automatizados.

Para atendimento a este critério as seguintes questões devem ser respondidas:

1. O modelo permite a anotação semântica para enriquecer seus dados e metadados?
2. As anotações são realizadas em linguagem natural, ou padronizada?
3. O tipo de anotação semântica permite o uso de vocabulários e ontologias?
4. A anotação pode ser reutilizada por usuários e agentes de software.

Ressalta-se quanto mais padronizada for tipo de anotação, maior a possibilidade de reutilização e compartilhamento no contexto do *Linked Data*.

Critério 2 - Reuso de metadados

Este critério avalia se o modelo de dados realiza a reutilização de dados e metadados no processo de publicação de dados em *Linked Data* para otimizar o trabalho do publicador e se esta reutilização ocorre em alguma classe especifica, ou em todas as classes do modelo.

O diferencial está no reuso de vocabulários existentes, consolidados e disponíveis no *Linked Open Data* (LOD), ao invés de replicá-los. Para avaliar se o modelo de dados atende a este critério devem ser respondidas as seguintes questões:

1. O modelo reutiliza dados e metadados de outros vocabulários?
2. O modelo optou por utilizar as propriedades de outros vocabulários em seu contexto original, ou duplicou estas propriedades em seu modelo?
3. A reutilização das propriedades é realizada em classes especificas ou em todas as classes do modelo.

Neste critério é importante observar a quantidade de propriedades reutilizadas de outros modelos. Deste modo, no desenvolvimento de um modelo quanto maior o número de propriedades e classes onde foram reutilizados metadados melhor a avaliação do modelo.

Critério 3 - Links entre dados e metadados do modelo com outros recursos da web

Este critério avalia a existência de links entre as combinações semânticas dos dados e metadados dos modelos com outros recursos da web. Neste critério, a quantidade de conexões com outros *datasets* está diretamente relacionada à facilidade para publicar os dados em *Linked Data* e realizar a interoperabilidade de dados.

Para avaliar se o modelo de dados atende a este critério devem ser respondidas as seguintes questões:

1. O modelo permite o link de seus dados com outros metadados do LOD?
2. O modelo apresenta conexões com outros *datasets* no LOD?

Neste critério quanto maior a conexão deste modelo com outros *datasets* da LOD, maior a facilidade para o compartilhamento de informações com dados de outras fontes. Dessa forma, ao desenvolver um modelo de dados devem ser observados os princípios do *Linked Data*, de modo a permitir que o modelo seja publicado no LOD.

Critério 4 - Modelagem de dados num formato semântico estruturado

A modelagem é considerada como uma boa prática, em virtude de definir a forma de representação dos dados e os relacionamentos entre os mesmos independente da aplicação e contexto em que os dados são utilizados.

Para avaliar o modelo neste critério, as seguintes perguntas devem ser respondidas:

1. O modelo realiza a modelagem de dados em um formato semântico estruturado?
2. A forma de modelagem adotada permite sua compreensão por sistemas automatizados?
3. A modelagem em formato estruturado é realizada em todas as classes do modelo ou em classes especificas?

4. A estruturação adotada pelo modelo de dados permite a utilização de ontologias?

No processo de avaliação ou desenvolvimento de modelos de dados a utilização de ontologias para a estruturação dos dados pode ser entendida como um diferencial importante, na medida em que as relações semânticas criadas pelas ontologias podem ser utilizadas pelo *Linked Data* para ampliar as conexões entre os recursos na web.

Critério 5 - Utilização de ferramentas para o enriquecimento semântico

A adoção de ferramentas automatizadas para a extração de conhecimento e anotação semântica é fundamental para reduzir o gargalo na aquisição de conhecimento, principalmente considerando os grandes volumes de documentos existentes na web. (URENT et al., 2005)

Neste critério, os modelos deverão ser avaliados considerando os seguintes questionamentos:

1. O modelo utiliza ferramentas para apoiar no processo de enriquecimento semântico?
2. As ferramentas adotadas pelo modelo realizam as tarefas de forma automática, ou semiautomática?
3. As ferramentas adotadas realizam todas as etapas do processo de enriquecimento semântico (extração do conhecimento, conversão, classificação, organização, anotação, publicação dentre outros) ou apenas algumas das atividades?
4. As ferramentas são proprietárias ou software livre?

A adoção de software livre é superior a de softwares proprietários, uma vez que os custos de aquisição da licença e manutenção podem inviabilizar a utilização da ferramenta por outros modelos.

Critério 6 - Utilização de interface gráfica

Este critério considera a utilização de interface gráfica para processo de enriquecimento semântico e publicação de dados. Para cada modelo deverão ser avaliadas as seguintes questões:

1. As ferramentas disponibilizadas pelo modelo possuem interface gráfica para facilitar o trabalho do publicador?
2. O *design* do sistema facilita a colaboração entre os usuários, permitindo que especialistas possam contribuir e compartilhar documentos?

3. Os usuários do modelo podem utilizar estas interfaces para compartilhar informações e contribuir para o processo de enriquecimento semântico?

O principal objetivo da utilização de interfaces gráficas é facilitar o trabalho do publicador. Neste item a interação do publicador e do usuário comum com a ferramenta é o principal foco a ser avaliado.

Critério 7 – Relacionamento entre os metadados do modelo e termos de outros vocabulários.

Este critério se refere à utilização de relações de sinonímia (equivalência), associação e hierarquia entre o metadado e o termo correspondente em outros vocabulários utilizados.

Para este critério devem ser avaliados as seguintes questões:

1. Quais os tipos de relações são utilizados pelo modelo: sinonímia (equivalência), associação e hierarquia?
2. O modelo estabelece relações em uma única classe ou em todas as classes do modelo?

Neste critério o mais relevante é avaliar quantos tipos de relações o modelo apresenta, quantas classes fazem uso destas relações e se a natureza das relações é entendida pelos usuários do modelo.

3 CONCLUSÕES

Nesse artigo, apresentou-se a proposta uma Matriz de Técnicas e Recursos de Enriquecimento Semântico - Matriz TRESO, com o objetivo de investigar as contribuições dos modelos de dados para o enriquecimento semântico como uma ferramenta para avaliação e desenvolvimento de outros modelos de dados.

Para cada um dos sete critérios propostos pela Matriz, foram elaboradas questões de competência, para ser serem utilizadas tanto para avaliação de modelos existentes, quanto para serem observados durante o processo de construção de novos modelos.

Dessa forma, acredita-se que novos modelos podem ser desenvolvidos ou aprimorados por meio da observação dos critérios para avaliação de enriquecimento semântico proposto na Matriz TRESO.

REFERÊNCIAS

BATISTA, M. G. R.; LÓSCIO, B. F. OpenSBBD: Usando Linked Data para Publicação de Dados Abertos sobre o SBBD. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS – SBBD, 2013, Recife. Anais... Recife: UFPE, 2013. Disponível em: <https://goo.gl/LN4CQ5>. Acesso em: 27 jul. 2018.

BERNERS-LEE, Tim. **Linked Data-design issues**. 2006. Disponível em: <https://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 15 mar. 2017.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. **Linked Data-the story so far**. Semantic services, interoperability and web applications: emerging concepts, United States of America, p. 205-227, 2009. Disponível em: <https://books.google.com.br/books?hl=ptBR&lr=&id=tP8HLETgbKcC&oi=fnd&pg=PA205&dq=Linked+Data:+Design+issues&ots=-hJuToD1BB&sig=dT6uRAuVjm_XBZ iCWa-9EWLmP4#v=onepage&q=Linked%20Data%3A%20Design%20issues&f=fa se>. Acesso em: 15 mar. 2017.

HALPIN, H.; LAVERENKO, V. Relevance feedback between hypertext and semantic search. In: SEMANTIC SEARCH WORKSHOP AT THE WORLD WIDE WEB CONFERENCE, 18., 2009. Proceedings… Madrid: [S.n.], 2009. Disponível em: <https://pdfs.semanticscholar.org/b1ea/4d6bb0091004f3884920c1a57031c7d4e58f.pdf>. Acesso em: 4 jun. 2013.

ISOTANI, Seiji; BITTENCOURT, Ig Ibert. **Dados Abertos Conectados: Em busca da web do Conhecimento**. São Paulo: Novatec, 2015.

LIRA, M. A. B. de. **Uma Abordagem Para Enriquecimento Semântico de Metadados Para Publicação de Dados Abertos.** 2014. 95 f. Dissertação (Mestrado em Ciência da Informação) - Centro de Informática, Universidade Federal de Pernambuco, Recife, 2014. Disponível em: <https://repositorio.ufpe.br/bitstream/handle/123456789/11570/DISSERTA%C3%87%C3%83O%20M%C3%A1rcio%20Angelo%20Bezerra%20de%20Lira.pdf?sequence=1&isAllowd=y>. Acesso em: 19 fev. 2018.

MARCONDES, CH. Linked Data - dados interligados - e interoperabilidade entre arquivos, bibliotecas e museus na web. Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 17, n. 34, p. 171-192, mai./ago. 2012.

SILVA, D. L. da. **Ontologias para representação de documentos multimídia: análise e modelagem**. 2014. 442 f. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2014. Disponível em: <https://goo.gl/FJqtkC>. Acesso em: 27 jul. 2018.

SAYÃO, L. F. Modelos teóricos em Ciência da Informação: abstração e método científico. **Ciência da Informação**, Brasília, v. 30, n. 1, p. 82-91, jan./abr. 2001. Disponível em: <http://www.scielo.br/scielo.php?pid=S010019652001000100010&script=sci_abstract&tlng=pt>. Acesso em: 15 ago. 2017.

UREN, V., CIMIANO, P., IRIA, J., HANDSCHUH, S., VARGAS-VERA, M., MOTTA, E., CIRAVEGNA, F., Semantic Annotation for Knowledge Management: Requirements and a survey of the state of the art. **Journal of Web Semantics**: Science, Services and Agents on the World Wide Web 4, 14-28, 2005. Acesso em: 12 out. 2016.