XVIII ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2017

GT-2 – Organização e Representação do Conhecimento

CONVERSÃO DE DADOS DE BIBLIOTECAS DIGITAIS DE TESES E DISSERTAÇÕES EM *LINKED DATA*

Umberto Lima Diniz - Universidade Federal de Minas Gerais (UFMG)

Gercina Ângela de Lima - Universidade Federal de Minas Gerais (UFMG)

Benildes Coura Moreira dos Santos Maculan - Universidade Federal de Minas Gerais (UFMG)

CONVERSION OF DIGITAL LIBRARY DATA OF THESES AND DISSERTATIONS IN LINKED DATE

Modalidade da Apresentação: Comunicação Oral

Resumo: Este trabalho apresenta os resultados de uma pesquisa que teve como objetivo estudar a possibilidade de uso do Linked Data para tratamento de informação em bibliotecas digitais de teses e dissertações (BDTDs), visando agregar valor às informações disponibilizadas. Linked Data oferece um conjunto de boas práticas, propostas por Tim Berners-lee, *Hendler e Lassila* (2001), para permitir a ligação de dados interoperáveis na web. A biblioteca digital avaliada foi BDTD, da Escola de Ciência da Informação (ECI) da Universidade Federal de Minas Gerais (UFMG). A metodologia aplicada incluiu o uso de parte dos resultados do trabalho de Maculan (2011), a TAFNAVEGA. A seleção da amostra foi feita de forma aleatória, sendo constituída por dez documentos. Os documentos selecionados foram recuperados na BDTD/ECI/UFMG e identificados com um Uniform Resource Identifier (URI), atribuído pela própria BDTD. A TAFNAVEGA forneceu as classes e os termos que foram utilizados para relacionar os dados da DBpedia aos documentos e para a construção das triplas em Linked Data. Uma vez que a tripla é formada por Sujeito, Predicado e Objeto, foram utilizados os URIs dos documentos para identificar o sujeito; o conjunto de elementos do Dublin Core e seus qualificadores para identificar o Predicado e, por fim, o Projeto DBpedia para identificar URIs para os termos da taxonomia. O uso do Projeto DBpedia se justifica por ter caráter multidisciplinar, ter dados interoperáveis e estar livremente disponibilizado**.**

Palavras-Chave: Linked Data; Biblioteca Digital; Conversão de Dados.

Abstract: This paper presents the results of a research that had as objective to study the possibility of using Linked Data to treat information in digital libraries of theses and dissertations (BDTDs), aiming to add value to the information made available. Linked Data offers a set of best practices, proposed by Tim Berners-lee, *Hendler e Lassila* (2001), to allow interoperable data on the web. The digital library evaluated was BDTD, of the School of Information Science (ECI) of the Federal University of Minas Gerais (UFMG). The applied methodology included the use of part of the results of the work of Maculan (2011), TAFNAVEGA. The selection of the sample was made in a random way, being

constituted by ten documents. The selected documents were retrieved in BDTD/ECI/UFMG and identified with a Uniform Resource Identifier (URI), assigned by BDTD itself. TAFNAVEGA provided the classes and terms that were used to relate DBpedia's data to the documents and to construct the triples in Linked Data. Since the triple is formed by Subject, Predicate and Object were used the URIs of the documents to identify the subject; The set of Dublin Core elements and their qualifiers to identify the Predicate, and finally the DBpedia Project to identify URIs for the terms of the taxonomy. The use of the DBpedia Project is justified because it has a multidisciplinary character, has interoperable data and is freely available.

Keywords: Linked Data; Digital Library; Data Conversion.

1 INTRODUÇÃO

Com o surgimento da Internet, na década de 1990, e, posteriormente, com o surgimento da *Word Wide Web* (WWW), os recursos informacionais passaram a ser acessados de qualquer parte do mundo. Esse fato beneficiou o usuário em suas buscas, pois facilitou o acesso a acervos diversificados, de maneira rápida e eficiente.

Nesse contexto, surge a proposta da *Web* Semântica, desenvolvida por Tim Berners-Lee, com o objetivo de fazer uma extensão da *web*, de maneira que os conteúdos das páginas, seus dados, pudessem ser semanticamente estruturados, com significação de dados, facilitando o compartilhamento e o reúso de informações entre humanos e computadores (BERNERS-LEE; HENDLER; LASSILA, 2001).

Com a proposta da *Web* Semântica, as Bibliotecas Digitais (BD), criadas na década de 1990, iniciam seu processo de adequação às novas demandas de produtos disponibilizados em ambiente digital. Para implantar a semântica nesse ambiente, foi necessário o desenvolvimento de aplicações tecnológicas para a criação de padrões, protocolos e diretrizes em formato aberto, pretendendo assegurar que os dados fossem acessíveis a todos na *Web*. Esses padrões possibilitaram a proposta do *Linked Data,* que tem como objetivo a vinculação de dados, ou seja, ligar recursos similares através de *hiperlinks* apontados para recursos armazenados nas nuvens, os quais podem complementar as informações contextualizadas sobre determinado recurso informacional. Essa tecnologia foi estabelecida por Berners-Lee (2006) e adotada pelo grupo *Semantic Web Education and Outreach* (SWEO), ligado à *World Wide Web Consortium* (W3C), permitindo a conexão de dados relacionados, melhorando o tratamento dos recursos digitais e facilitando a interoperabilidade entre diferentes sistemas.

Essa integração de dados, realizada a partir de regras, normas de formatos (extensões) e metadados (para a descrição dos aspectos semânticos dos dados), pode facilitar o

compartilhamento e o reúso de informação advinda de fontes diversas. Um dos exemplos mais comuns de fontes disponíveis na *web* são as bibliotecas digitais, em especial, as Bibliotecas Digitais de Teses e Dissertações (BDTDs) que, geralmente, estão hospedadas em diferentes instituições. Entretanto, um dos problemas verificados é que, ainda que disponíveis na internet, muitas vezes utilizam sistemas que não permitem o fácil reúso de seus dados e de seu conteúdo informacional, e, assim, não permitem interoperabilidade de dados (padrões abertos). Ademais, esse problema também impede que o conteúdo das BDTDs seja interligado a outros conteúdos disponíveis na *web*, que possam agregar informação aos dados de pesquisa.

Uma das formas de solucionar esse problema de interoperabilidade de dados é utilizar dados de diversas naturezas, que já se encontram disponíveis na web e que estão marcados semanticamente (*datasets*), tais como: os dados da Wikipédia, do *United States (US) Census*, do *GeoNames*, do *DailyMed*, do *British Broadcasting Corporation Music* (*BBC Music*) e do *Association for Computing Machinery* (ACM), por exemplo. Nesse contexto, este trabalho apresenta o resultado de uma pesquisa de mestrado, que teve como proposta estudar a contribuição do *Linked Data* como recurso incorporado à BDTD da Universidade Federal de Minas Gerais, visando agregar novos dados às informações disponibilizadas.

Partiu-se dos pressupostos de que o recurso *Linked Data* pode: (a) contribuir com as informações disponibilizadas na BDTD ao incorporar dados que permitem um maior detalhamento do conteúdo dos trabalhos científicos, para melhorar a sua compreensão; (b) agregar informações à BDTD, a partir do estudo de suas características; e (c) permitir a inclusão e a recuperação de características específicas da pesquisa constantes nas teses e dissertações que não são contempladas no estado atual do registro bibliográfico.

Depois da introdução do tema a ser tratado, o texto está assim organizado: na seção 2, apresentam-se os conceitos de repositório e de biblioteca digital, assim como uma descrição sucinta do padrão de metadados *Dublin Core*; na seção 3, descrevem-se os princípios da *Web* Semântica e do *Linked Data,* utilizados para interligar dados em ambiente digital; na seção 4, descreve-se a *DBpedia*, que é uma base de dados colaborativa multilíngue, cujo objetivo é extrair representações de informações, na forma de conjuntos de triplas em RDF; na seção 5, apresentam-se a metodologia e os procedimentos empregados; na seção 6, são expostas as análises dos resultados; e, na sequência, seguem-se as considerações finais e a lista de referências.

2 REPOSITÓRIOS, BIBLIOTECAS DIGITAIS E PADRÃO DUBLIN CORE DE METADADOS

Repositórios Digitais são coleções de informação digital, que têm formas diversas e propósitos variados. Podem tanto ser direcionadas a um público em geral ou à audiência específica. Os repositórios digitais são muito utilizados em armazenamento da produção científica de uma instituição e são conhecidos, também, na literatura da Biblioteconomia e Ciência da Informação, como Repositórios Institucionais; a Biblioteca Digital é um exemplo desses Repositórios. Os repositórios digitais surgiram no contexto da universidade e relacionaram-se com a introdução do *Open Acess* (Acesso Aberto) para a literatura científica.

As Bibliotecas Digitais são um tipo de Repositório Institucional, que se caracteriza por manter as mesmas características quanto à seleção e ao tratamento de seu acervo, se comparado à biblioteca tradicional, mantendo, assim, os mesmos princípios já alicerçados de como a informação é organizada, criados pela Biblioteconomia e Ciência da Informação. Esse é um conceito surgido na década 1990 e, de acordo com Cunha (2008), a biblioteca digital precisa ter conteúdo, que pode ser material antigo, convertido para o formato digital, ou material novo, nascido digitalmente.

Faz-se necessário o desenvolvimento de normas e padrões para a representação das informações que facilitem a identificação de sua descrição e sua localização e ainda proporcionem a interoperabilidade entre os sistemas informacionais. Nesse sentido, os metadados têm papel fundamental, pois estão estruturados em uma ambiência padronizada e facilitam os processos de busca e recuperação dos recursos informacionais em ambientes informacionais digitais. Os padrões de metadados têm como objetivo principal auxiliar a recuperação da informação, prover a interoperabilidade entre os recursos de informação na Internet e, de modo simplista, apontam para o dado que deve ser informado sobre o conteúdo de um documento (item de informação).

Existem diversos padrões de metadados desenvolvidos para diferentes finalidades (SOUZA, T. B.; CATARINO, M. E.; SANTOS, 1997). O padrão *Dublin Core* (DC) é um dos padrões de metadados utilizados para a descrição dos documentos disponíveis em uma biblioteca digital. Este padrão foi criado em 1995, por um comitê formado pela *OnLine Computer Library* e pelo *National Center for Supercomputer Applications*, na cidade de Dublin (Ohio, EUA).

O padrão DC foi desenvolvido de modo a possuir diversas características que o tornam intuitivo para qualquer pessoa que deseja representar documentos e recursos,

sobretudo na internet. O conceito inicial do *Dublin Core Element Set*, ou simplesmente DC, é permitir a descrição de material na web pelos próprios autores, sendo suficientemente simples para tanto, sem perder a profundidade de detalhes que a tarefa exige para se conseguir uma boa descrição. O DC é um formato independente de sintaxe, uma vez que não é uma linguagem de intercâmbio, mas um conjunto de 15 elementos de dados (CAPLAN, 1995), conforme mostra o Quadro 1.

Quadro 1: Elementos de metadados do Dublin Core.

Elementos	**Descrição**
Título	Nome dado aos recursos.
Criador	Entidade originalmente responsável pela criação do conteúdo do recurso.
Assunto	Tema do conteúdo do recurso. Pode ser expresso em palavras-chaves e/ou categoria. Recomenda-se o uso de vocabulários controlados.
Descrição	Relato do conteúdo do recurso. Exemplos: sumários, resumo e texto livre.
Publicador	Entidade responsável por tornar o recurso disponível.
Colaborador	Entidade responsável pela contribuição intelectual do conteúdo do recurso.
Data	Data associada a um evento ou ciclo de vida do recurso.
Tipo	Natureza ou gênero do conteúdo do recurso. Exemplos: texto, imagem, som, dados, *software.*
Formato	Manifestação física ou digital do recurso. Exemplos: html, pdf, ppt, gif.
Identificador	Referência não ambígua (localizador) para o recurso dentro do dado contexto.
Fonte	Referência a um recurso do qual o presente é derivado.
Idioma	Língua do conteúdo intelectual do recurso.
Relação	Referência para um recurso relacionado.
Cobertura	Extensão ou escopo do conteúdo do recurso; pode ser temporal e espacial.
Direitos autorais	Informação sobre os direitos assegurados dentro e sobre o recurso.

Fonte: Alves e Souza (2007, p. 3).

Para suprir necessidades especiais, o padrão DC pode ser personalizado com campos adicionais e existem duas classes de qualificadores para atender a esta demanda: os elementos de refinamento, que dão mais especificidade a um elemento, e um conjunto de

esquemas de codificação, que identificam esquemas para o valor do elemento.

Como iniciativa para atender aos princípios da *Web* Semântica, foi criado o *Dublin Core Metadate Iniciative* (DCMI), que estabeleceu um identificador URI para cada um de seus elementos de metadados, em RDF, de maneira que é possível especificar, em detalhes, o seu uso, que é fundamental para aplicações processáveis por máquina (CATARINO; SOUZA, 2012). Alguns exemplos do esquema DCMI com a URI são apresentados no Quadro 2.

Quadro 2: Exemplos do esquema DCMI com URL.

Campo	**URI**
Assunto	http://purl.org/dc/elements/1.1/subject
Título	http://purl.org/dc/elements/1.1/title
Tipo	http://purl.org/dc/elements/1.1/type

Fonte: Elaborado pelos autores (2017).

Às vezes, a especificação também aponta para o tipo de valor que deve ser empregado na descrição do recurso, tal como para o metadado DATE que orienta o uso da Norma ISO 8601 (CATARINO; SOUZA, 2012).

Dessa forma, espera-se que a descrição de dados por meio de metadados proporcione qualidade na representação de um recurso e permita a sua interoperabilidade entre aplicações, com outros recursos, para facilitar o compartilhamento de informações disponíveis na *Web*. Sabe-se que o padrão DC para a descrição de seus recursos não é o único utilizado na web, mas é uma comunidade internacional forte, cujo desenvolvimento é um trabalho colaborativo de profissionais de diversas áreas, que pode atender a alguns princípios da *Web Semântica*, em especial aos do *Linked Data*.

3 *WEB* SEMÂNTICA

A *Web* é considerada um ambiente que possibilita que as pessoas criem repositórios de dados, permitindo a interoperabilidade desses dados, por meio de *links*, utilizando tecnologias específicas para este compartilhamento de conhecimento entre as pessoas e as máquinas. Souza e Alvarenga (2004, p. 133) afirmam que:

> Embora tenha sido projetada para possibilitar o fácil acesso, intercâmbio e a recuperação de informações, a *Web* foi implementada de forma descentralizada e quase anárquica; cresceu de maneira exponencial e caótica e se apresenta hoje como um imenso repositório de documentos que deixa muito a desejar quando precisamos recuperar aquilo de que

temos necessidade.

Nota-se, pois, que a facilidade de publicação na *Web* fez com que as informações fossem disponibilizadas desordenadamente, sem nenhuma estrutura, o que causou deficiências na recuperação da informação. Dessa forma, alguns problemas na recuperação de informações foram evidenciados com o passar do tempo, o que promoveu uma maior preocupação dos pesquisadores com o tema. Uma das tentativas propostas por Berners-Lee, Lassila e Hendler (2001) para solucionar tais deficiências foi o surgimento da *Web Semântica, Web* 3.0 ou *Web* de Dados, que aplica tecnologias para criar um ambiente com dados integrados, facilitando a recuperação de informações disponíveis na *Web*.

Berners-Lee, Hendler e Lassila (2001, n.p.) afirmam que "a *Web* Semântica não é uma *Web* separada, mas uma extensão da atual. Nela, a informação é dada com um significado bem definido, permitindo melhor interação entre os computadores e as pessoas". Portanto, o objetivo da *Web* Semântica é estruturar o conteúdo que está disponível na Internet, e, para isso, os computadores precisam entender este conteúdo por meio de dados estruturados, sendo necessário um conjunto de regras que ajudem a interpretação de dados por máquinas, de forma que eles possam ter significado e, principalmente, que se tornem passíveis de serem recuperados. Para isso, a *Web* Semântica engloba conceitos e tecnologias, entre as quais se destacam, conforme Dias e Santos (2003) e Catarino, Cervantes e Andrade (2015):

(1) *Extensible Markup Language* (XML): permite criar *tags* (campos de texto) em documentos, que podem ser usados por programas ou *scripts*.

(2) *Resource Description Framework* (RDF): determina um significado às estruturas, codificando as *tags* e construindo triplas: Sujeito+Predicado+Objeto, que podem ser representadas com a tecnologia XML e por um *Universal Resource Identifier* (URI), para intercâmbio de dados.

(3) Aplicações verticais (*vertical applications*): recomendações, tecnologias e padrões de comunidades específicas que utilizam as tecnologias do W3C.

(4) Inferência: possibilita a descoberta de conhecimentos a partir dos dados estruturados disponíveis na web e de informações adicionais advindas de um vocabulário ou de um conjunto de regras, representadas de maneira formal.

(5) *Web Ontology Language* (OWL): tecnologia para definição e instanciação de conceitos, indivíduos, classes e propriedades em vocabulários, utilizando uma semântica formal.

(6) *Simple Knowledge Organisation System* (SKOS): tecnologia utilizada para representar a estrutura básica de diferentes tipos de vocabulários ou sistemas de organização do conhecimento (SOCs).

A partir dessas e de outras tecnologias, a busca e consulta (*query*) é realizada utilizando-se uma tecnologia própria, a *Simple Protocol and RDF Query Language* (SPARQL), disponível desde 2008. O SPARQL possibilita recuperar conteúdos de dados estruturados e semiestruturados na *web*, explorando também outras relações e recursos, por meio de conexões entre conjuntos de dados heterogêneos, em uma única consulta.

Assim, o desenvolvimento da *Web* Semântica apresenta diversas tecnologias que incorporam outros elementos a este conjunto de regras e normas, possibilitando a explicitação do significado desses dados para que a sua proposta inicial seja alcançada. Entre elas, há o *Linked Data*.

3.1 LINKED DATA

O W3C apresentou a proposta denominada de *Linked Data* (Dados Ligados), conjunto de dados inter-relacionados na Web, que permite ligar dados interoperáveis entre sistemas de informação, reduzindo a complexidade e aumentando a compatibilidade entre os recursos na internet. De acordo com Heath e Bizer (2011, p. 8, tradução nossa):

> A ideia básica do *Linked Data* é aplicar a arquitetura geral da *World Wilde Web* à tarefa de compartilhar dados estruturados em escala Global. Para entender os princípios do *Linked Data*, é importante entender a arquitetura de um documento *Web* clássico[1].

Para criar aplicativos com esses dados, é necessário utilizar as tecnologias da *Web* Semântica que permitem a organização, a gestão e o acesso aos dados, tais como o formato *Resource Description Framework* (RDF), para fazer a conversão; as linguagens *Web Ontology Language* (OWL), *Simple Knowledge Organization System* (SKOS), para a representação semântica; as linguagens *Extensible Markup Language* (XML), *Hypertext Markup Language* (HTML) *e eXtensible Hypertext Markup Language* **(**XHTML), para a marcação; e o *Protocol And RDF Query Language* (SPARQL), para obter acesso aos dados (busca).

A estrutura central para interligar dados é a criação de um conjunto de triplas, cada uma consistindo de um Sujeito (nó), um Predicado (relação entre os nós) e um Objeto (nó).

[1] The basic idea of Linked Data is to apply the general architecture of the World Wide Web to the task of sharing structured data on global scale. In order to understand these Linked Data principles, it is important to understand the architecture of the classic document Web.

O conjunto de triplas é chamado de Grafo RDF (RDF *Graph*). Um Grafo RDF pode ser representado como um diagrama de nós direcionado (Figura 1).

Figura 1: Representação gráfica de uma tripla.

Fonte: Elaborado pelos autores (2017).

Os dados são formados por triplas que especificam os relacionamentos (Predicado) entre as entidades (Sujeito e Objeto). Cada um desses itens é especificado através de um URI. Assim, o RDF fornece, também, uma sintaxe baseada em XML para apresentação desses grafos que, utilizando URIs, podem buscar informações em outras fontes da Web. O Quadro 3 é um extrato de RDF na notação RDF/XML, que corresponde ao grafo da Figura 1.

Quadro 3: RDF/XML descrevendo Eric Miller.

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:contact="http://www.w3.org/2000/10/swap/pim/contact#">
  <contact:Person rdf:about="http://www.w3.org/People/EM/contact#me">
  <contact:fullName>Eric Miller</contact:fullName>
   <contact:mailbox rdf:resource="mailto:em@w3.org"/>
   <contact:personalTitle>Dr.</contact:personalTitle>
 </contact:Person>
</rdf:RDF>
```

Fonte: https://www.w3.org/TR/2004/REC-rdf-primer-20040210/#figure1 – 2017.

Para Berners-Lee, Hendler e Lassila (2001), ao contrário da estrutura hipertextual, em que os *links* são relações âncoras em documentos de hipertexto escritos em HTML ou XHTML, quando as ligações são descritas em RDF, torna-se possível acessar, aleatoriamente, dados descritos por RDF, porque os URIs possibilitam a identificação de qualquer tipo de objeto ou conceito**.** Para isso, Berners-Lee, Hendler e Lassila (2001) propõem quatro regras:

> - Use URIs como nomes para as coisas; - Use HTTP URIs para que as pessoas possam procurar esses nomes; - Quando alguém procura um URI, forneça informações úteis, usando os padrões (RDF, SPARQL); - Inclua links para outros URIs para que eles possam descobrir mais coisas (BERNERS-LEE;

HENDLER; LASSILA, 2001, n.p.).

Apesar de o uso dessas regras garantir que os dados estejam interconectados, tais regras limitam a reutilização de maneira imprevisível, que também é uma característica da Web, que agrega valor à informação. Porém, pela sua característica versátil, o *Linked Data* permite a publicação dos dados e a criação de fontes de dados de diversas origens. Aplicativos criados com *Linked Data* nem sempre precisam ser abertos, pois podem ser criados para interligar dados em uma mesma base de dados internamente.

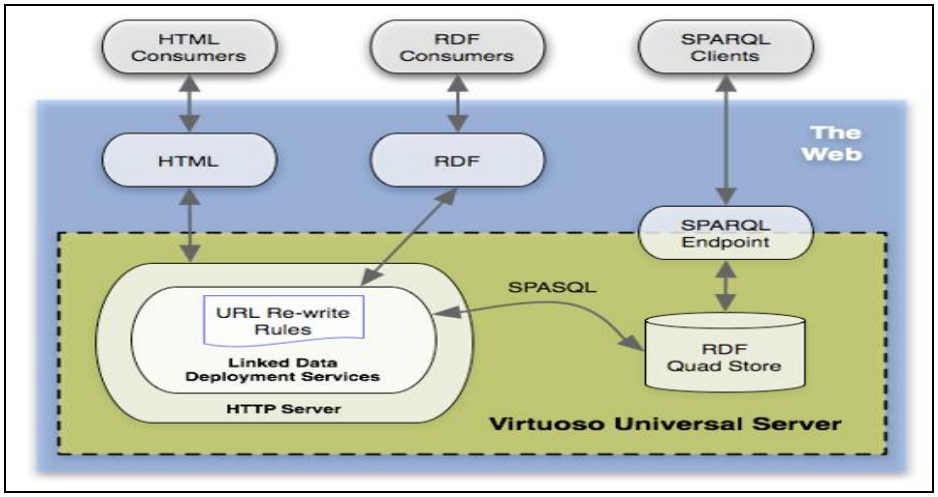
Para medir a qualidade dessas fontes de dados disponíveis na internet, Berners-Lee (2006) criou um sistema de avaliação em cinco níveis, que o autor chamou de "Cinco Estrelas". Nesse método, utiliza-se o conceito de *Namespace*, que são esquemas que descrevem as entidades semânticas, que são sempre relacionadas a algum *Namespace* que descreve sua estrutura, e ditará suas regras, dentro de um contexto específico.

Outro conceito importante é o de *Triple Stores* que, de acordo com Sequeda (2013), são sistemas de gerenciamento de bases de dados (DBMS – *Database management Systems*) para dados modelados usando RDF. Diferentemente dos sistemas de gerenciamento de bases de dados relacionais (RDBMS – *Relational Database Management Systems*), que armazenam dados em relações ou tabelas e cujas consultas são feitas em SQL, *triplestores* armazenam triplas RDF e são consultados usando-se SPARQL.

4 BASE DE DADOS *DBpedia*

O *DBpedia* é composto de tópicos referenciados na Wikipédia, sendo uma versão no formato *Resource Description Framework* (RDF) da mesma, que é uma enciclopédia colaborativa multilíngue, disponibilizada na *Web* e de uso livre. O objetivo do *DBpedia* é extrair, na forma de conjuntos de triplas RDF, o conhecimento acumulado nos diferentes artigos da Wikipédia, que é um recurso semiestruturado de informações, conforme afirmam Hovy, Navigli e Ponzetto (2013). Esses dados estruturados são disponibilizados em uma base de dados, usando o *software OpenLink Virtuoso* (desenvolvido pela *OpenLink Software*), que é um mecanismo de *middleware* e um banco de dados híbrido, para consultas por meio de *endpoints* SPARQL, utilizando tecnologias da *Web* Semântica e do *Linked Data* (HELLMANN et al., 2014). Esses dados possuem Identificadores de Recursos Uniforme (URI), o que os tornam uma fonte rica para os dados vinculados. A Figura 2 mostra a arquitetura do *DBpedia*, com o *OpenLink Virtuoso*.

Figura 2: Arquitetura atual do DBpedia.



Fonte: (<Dbpedia.org>) – 2017.

Ressalta-se que o Virtuoso é um "servidor universal", que não necessita ter servidores dedicados para cada um dos domínios de funcionalidades, precisando apenas de um único processo de servidor para vários protocolos diferentes.

Os recursos informacionais existentes na Wikipédia consistem, principalmente, de texto livre e de diversos outros tipos de recurso informacional já estruturados, a exemplo das caixas e categorização de informações, as referências geográficas, as imagens e, também, os *links* que levam a outras páginas na Web (LAUFER, 2015).

Atualmente, a Wikipédia possui informações em 266 idiomas, conforme estatística[2] de fevereiro de 2017, e o *DBpedia* extrai as suas informações estruturadas e as combina em uma grande base de conhecimento, na qual:

> Cada entidade (recurso) no conjunto de dados do *DBpedia* é denotado por uma URI diferenciável, na forma "http:DBpedia.org/resource/{nome}", onde "{nome}" é derivado da URL do artigo origem da Wikipédia, que tem a forma "http://en.wikipedia.org/wiki/{nome}". Assim, cada entidade da DBpedia está conectada diretamente a um artigo da Wikipédia. Cada {nome} de entidade DBpedia retorna uma descrição de um recurso na forma de um documento Web (LAUFER, 2015, n.p.).

Essa ação é realizada a partir de extratores, que são programas específicos, criados com essa finalidade, "para converter partes específicas de artigos da Wikipédia em sentenças RDF" (WEBER, 2015, p. 27). O autor afirma que esse processo dá origem a um conjunto de entidades que são classificadas de acordo com a *DBpedia Ontology*, que

> [...] é uma ontologia multi-domínio que em sua versão (release 2014 [http://wiki.DBpedia.org/Ontology]) cobre um conjunto de 685 diferentes

[2] Informação disponível em: <http://stats.wikimedia.org/PT/Sitemap.htm>. Acesso em: 11 abr. 2017.

> classes composto por uma hierarquia de subsunção de até oito níveis de profundidade. Em seus níveis superiores estão conceitos como Pessoa, Organização, Localização e Eventos – conceitos que correspondem às categorias semânticas comumente requisitadas em tarefas de Classificação de Entidades Nomeadas. Nos níveis mais baixos estão conceitos mais específicos tais como Banda, uma especialização de Organização, ou Escritor, especialização de Pessoa (WEBER, 2015, p. 27).

O projeto *DBpedia* é hospedado e publicado usando-se o *software OpenLink Virtuoso* (desenvolvido pela *OpenLink Software*), que é um mecanismo de *middleware* e um banco de dados híbrido, que combina diferentes funcionalidades em um único sistema.

Contudo, conforme afirma Kontokostas et al. (2014 *apud* SOUZA; SANTARÉM SEGUNDO, 2014), testes aplicados na ontologia do *DBpedia* apontam para um grande número de erros nos recursos disponibilizados, tais como: código postal em formato errado, dados de datas de nascimento/morte incompletos, locais sem coordenadas geográficas ou com dados errados ou duplicados, classificação de tipologia de dados incorreta, entre outros.

Considera-se que esses problemas podem ser minimizados ou, até mesmo eliminados, caso haja um esforço comum para a criação de uma infraestrutura de colaboração para pesquisadores e especialistas em domínios, conforme aponta Morsey (2012). Nesse tipo de colaboração, a possibilidade de ocorrer erros de representação é bem menor.

5 PROCEDIMENTOS METODOLÓGICOS

Esta pesquisa caracteriza-se como empírica, exploratória e aplicada, com uma abordagem qualitativa, que teve como objetivo explorar o tema do *Linked Data,* verificando a sua contribuição para agregar valor às informações disponibilizadas por BDTDs. O objeto da pesquisa foram os *links* criados para ligar conteúdo na *web*, sendo que o *corpus* de aplicação para teste foi uma amostra com dez documentos que são parte do trabalho de Maculan (2011), que indexou 41 documentos, entre teses e dissertações oriundas da BDTD/UFMG/ECI, que foi o ambiente da pesquisa.

A metodologia para criação do *Linked Data* se orientou:

(1) nas dez classes básicas da taxonomia facetada denominada TAFNAVEGA (MACULAN, 2011): C1. Tema; C2. Objeto Empírico; C3. Escopo; C4. Ambientação; C5. Tipo de Pesquisa; C6. Coleta de Dados; C7. Métodos; C8. Fundamento Teórico; C9. Fundamento

Histórico/Contextual; C10. Resultados; (2) nos elementos de metadados do *Dublin Core* e seus qualificadores; e (3) no *dataset DBpedia*.

Os procedimentos metodológicos para a construção das Triplas RDF (*Linked Data*) foram os seguintes:

(1) Identificação das classes, termos e relacionamentos: buscaram-se pelos documentos na BDTD/UFMG/ECI para verificar se possuíam URI, utilizando o elemento *Handle*[3], como identificador único para esse tipo de obra na internet.

(2) Identificação dos indicadores de classes no *Dublin Core*: verificou-se a correspondência das classes que compõem a TAFNAVEGA (MACULAN, 2011) para determinar os *namespaces* de cada uma delas, conforme mostra o Quadro 4.

Quadro 4: Termos da Taxonomia e URLs correspondentes.

Classe	**URI Correspondente**
Ambientação	http://purl.org/dc/terms/spatialhttp://purl.org/dc/terms/spatial
Causa e efeito	Classe não disponível e equivalente não localizado.
Coleta de dados	Classe não disponível e equivalente não localizado.
Escopo	http://purl.org/dc/terms/coveragehttp://purl.org/dc/terms/coverage
Fundamento histórico/conceitual	Classe não disponível e equivalente não localizado.
Fundamento teórico	Classe não disponível e equivalente não localizado.
Método	http://purl.org/dc/terms/instructionalMethod
Objeto	http://purl.org/dc/dcmitype/PhysicalObjecthttp://purl.org/dc/dcmitype/PhysicalObject
Resultados	Classe não disponível e equivalente não localizado.
Tema	http://purl.org/dc/elements/1.1/subjecthttp://purl.org/dc/elements/1.1/subject
Tipo de pesquisa	Classe não disponível e equivalente não localizado.

Fonte: Elaborado pelos autores (2017).

Nessa verificação, percebeu-se que nem todas as classes puderam ser equiparadas a uma URI do DCMI, e, portanto, não foram utilizadas; são elas: fundamento histórico/conceitual; causa e efeito; resultados, fundamentos teóricos; e tipo de pesquisa.

[3] *Handle* é um tipo de serviço de Proxy que fornece um *link* persistente para determinados conteúdos disponíveis na web. (HDL.handle.net)

Como resultado, foram obtidos 77 termos da TAFNAVEGA, que preenchiam as classes dos dez documentos da amostra.

(3) Identificação de URI para os termos da taxonomia: verificação de correspondência entre os 77 termos que compõem a TAFNAVEGA e as URIs disponíveis no *DBpedia*, que, para o exemplo do termo "indexação", foi feita da seguinte maneira: a) buscou-se pelo termo "indexação" na Wikipédia; b) encontrado o artigo, o termo foi traduzido para o idioma inglês, pois, apesar de algumas vezes existirem artigos com URIs para termos em português, em geral, os conteúdos do *DBpedia* são encontrados, em sua grande maioria, no idioma inglês; c) repetiu-se esses procedimentos para todos os 77 termos.

(4) Identificação de relacionamentos – Construção das triplas: já tendo sido identificadas, nas etapas anteriores, as URIs dos documentos (sujeito), das classes (predicado) e dos termos (objeto), esses dados permitiram criar as triplas que representam os relacionamentos entre os três elementos, conforme mostra o Quadro 5.

Quadro 5: Exemplo dos elementos que formam a tripla.

Sujeito	Predicado	Objeto
http://hdl.handle.net/1843/RRSA-6GGGUF	http://purl.org/dc/elements/1.1/subject	http://pt.DBpedia.org/page/Subject_indexing

Fonte: Elaborado pelos autores (2017).

Pode-se observar que foi possível a identificação das URIs para o Sujeito (documento), o Predicado (classe) e o Objeto (termo). Essa mesma representação da relação, com os três elementos, formada pelas triplas, foi realizada para cada um dos dez documentos da amostra desta pesquisa.

Com o conjunto de triplas pronto, realizou-se a sua transcrição para um editor de textos, utilizando o Bloco de Notas do Windows, e empregando o formato *Ntriple,* considerado mais intuitivo para trabalhar em formato RDF, que seguiu o seguinte padrão:

<URI do sujeito> <URI do predicado> <URI objeto ou um valor>

O exemplo a seguir mostra as triplas para a tese intitulada "Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais".

```
        <http://hdl.handle.net/1843/RRSA-6GGGUF>
    <http://purl.org/dc/elements/1.1/subject>
    <http://pt.DBpedia.org/page/Subject_indexing> .
        <http://hdl.handle.net/1843/RRSA-6GGGUF>
    <http://purl.org/dc/elements/1.1/PhysicalObject>
    <http://pt.DBpedia.org/page/Noun_phrase> .
        <http://hdl.handle.net/1843/RRSA-6GGGUF>
    <http://purl.org/dc/elements/1.1/coverage>
    <http://pt.DBpedia.org/page/Subject_indexing> .
        <http://hdl.handle.net/1843/RRSA-6GGGUF>
    <http://purl.org/dc/elements/1.1/spatial>
    <http://pt.DBpedia.org/page/Information_retrieval> .
```

O mesmo processo foi repetido para todos os dez documentos, separadamente, e, depois, as triplas foram convertidas para o formato RDF, utilizando-se a ferramenta *EasyRdf – Converter*[4].

```
xml version="1.0" encoding="utf-8" ?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:dc11="http://purl.org/dc/elements/1.1/">
 <rdf:Description rdf:about="http://hdl.handle.net/1843/RRSA-6GGGUF">
   <dc11:subject rdf:resource="http://pt.DBpedia.org/page/Subject_indexing"/>
   <dc11:PhysicalObject rdf:resource="http://pt.DBpedia.org/page/Noun_phrase"/>
   <dc11:coverage rdf:resource="http://pt.DBpedia.org/page/Subject_indexing"/>
   <dc11:spatial rdf:resource="http://pt.DBpedia.org/page/Information_retrieval"/>
 </rdf:Description>
        </rdf:RDF>
```

Com isso, todas as triplas, de todos os documentos, foram transformadas e declaradas em RDF.

6 ANÁLISE DOS RESULTADOS

Neste trabalho, estão apresentados os resultados e as análises de dois documentos da amostra trabalhada, visando determinar se há contribuição do *Linked Data* como recurso para ser incorporado à BDTD/ECI/UFMG, com o objetivo de agregar novos dados às informações disponibilizadas. A seleção dos dois documentos levou em consideração aqueles que pudessem demonstrar resultados representativos para criação da tripla em RDF, do uso dos elementos do DC e do potencial da *DBpedia* para interligar os dados com a BDTD.

Exemplo 1: Documento: Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais (http://hdl.handle.net/1843/RRSA-6GGGUF)**.**

[4] http://www.easyrdf.org/converter

Primeiramente, analisou-se o documento para a identificação dos termos que se referiam a cada uma das classes (conjunto que compõe a TAFNAVEGA), identificando-se a URI, por meio do identificador *Handle* já determinado na BDTD, cujos resultados estão no Quadro 6.

Quadro 6: Representação do documento pelas classes da TAFNAVEGA.

Classes	Termos
Tema	Indexação (automática e manual)
Objeto	Sintagmas nominais
Escopo	Indexação (automática e manual)
Ambientação	Sistemas de recuperação da informação
Métodos	Análise documentária

Fonte: Elaborado pelos autores (2017).

Verificou-se que foram preenchidas cinco classes da TAFNAVEGA, com a repetição de um termo nas classes Tema e Escopo. Em seguida, aplicaram-se os procedimentos descritos na seção de metodologia para a representação das triplas em RDF desse documento, quando foram identificadas as URIs do documento, dos termos e a correspondência no DC. As relações estão no Quadro 7.

Quadro 7: Relação de Classes e Termos da Taxonomia e DCMI, do Exemplo 1.

Classe	Termo
Tema	Indexação
Objeto	Sintagmas Nominais
Escopo	Indexação
Ambientação	Sistemas de Recuperação da Informação

Fonte: Elaborado pelos autores (2017).

Observa-se que somente foi possível criar as triplas em RDF para quatro classes e seus termos correspondentes. O que ocorreu foi que não foi encontrado, na base de dados *DBpedia*, o termo Análise Documentária, da classe “Métodos”. Esse fato evidenciou uma limitação da base, pois nem sempre o recurso era recuperado no *DBpedia*. Assim, sem o URI desse termo, não foi possível criar uma tripla para descrever tal relação. Ao final, foi gerado

o seguinte arquivo RDF:

```
<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:dc11="http://purl.org/dc/elements/1.1/">
  <rdf:Description  rdf:about="http://hdl.handle.net/1843/RRSA-6GGGUF">
   <dc11:subject rdf:resource="http://pt.dbpedia.org/page/Subject_indexing"/>
   <dc11:PhysicalObject rdf:resource="http://pt.dbpedia.org/page/Noun_phrase"/>
         <dc11:coverage rdf:resource="http://pt.dbpedia.org/page/Subject_indexing"/>
   <dc11:spatial rdf:resource="http://pt.dbpedia.org/page/Information_retrieval"/>
 </rdf:Description>
 </rdf:RDF>
```

Exemplo 2: Documento: As relações interdisciplinares refletidas na Ciência da Informação (http://hdl.handle.net/1843/ECID-7UUQ69).

Nesse documento, também foram aplicados os mesmos procedimentos; primeiramente, identificando-se o URI, por meio do identificador *Handle* já determinado na BDTD, cujos resultados estão no Quadro 8.

Quadro 8: Representação do documento pelas classes da Taxonomia TAFNAVEGA

CLASSE	**TERMO**
Tema	Interdisciplinaridade
Objeto	Avaliação de Periódicos
Escopo	Estudos da Produção e da Produtividade Científica
Ambientação	Base Qualis
Fundamento Histórico Conceitual	Interdisciplinaridade

Fonte: Elaborado pelos autores (2017).

Nota-se que a estrutura do documento foi representada por cinco classes da TAFNAVEGA. Contudo, ao verificar-se as correspondências com os elementos e qualificadores do DCMI e as URIs na base de dados *DBpedia*, constatou-se a inexistência de correspondência para muitos deles, conforme mostra o Quadro 9.

Quadro 9: Relação de Classes e Termos da Taxonomia e DCMI, do Exemplo 2

CLASSE	**TERMO**
Tema	Interdisciplinaridade

Fonte: Elaborado pelos autores (2017).

Percebe-se que não houve correspondência para os termos: avaliação de periódicos; estudos da produção e da produtividade científica; e Base *Qualis*. Ressalta-se que, apesar de se ter encontrado correspondência para o termo "interdisciplinaridade", para a classe Tema, na base de dados do Projeto *DBpedia*, não foi possível representá-lo na classe "Fundamento Histórico Conceitual", uma vez que não foi encontrada, para esta classe, correspondência nos elementos e qualificadores do DC. Assim, houve uma limitação, uma vez que não havendo um elemento DC apropriado, fica impossível explicitar a relação entre a classe e o termo.

Dessa forma, somente foi encontrada URI correspondente no *DBpedia* para o termo "interdisciplinaridade", para a classe Tema, tendo sido criada a tripla RDF para ela. Ao final, foi gerado o seguinte arquivo RDF:

```
<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:dc11="http://purl.org/dc/elements/1.1/"
      xmlns:ns0="http://purl.org/dc/dcmitype/"
      xmlns:dc="http://purl.org/dc/terms/">
  <rdf:Description rdf:about="http://hdl.handle.net/1843/ECID-7UUQ69">
         <dc11:subject rdf:resource="http://pt.dbpedia.org/page/Interdisciplinarity"/>
         <ns0:PhysicalObject>Avaliação de periódicos</ns0:PhysicalObject>
         <dc:coverage>Estudos da produção científica e produtividade científica</dc:coverage>
         <dc:spatial>Base Qualis</dc:spatial>
  </rdf:Description>
</rdf:RDF>
```

Nos dois exemplos apresentados, ficou evidenciado que, realmente, nem todos os termos e classes da TAFNAVEGA puderam encontrar correspondência nos elementos e qualificadores do DC e/ou na base de dados do projeto *DBpedia*, o que impediu a construção das triplas em RDF. Isso se deve, sobretudo, ao fato de o projeto *DBpedia* não possuir conteúdos específicos sobre temas (métodos, teorias, etc.) de pesquisas científicas.

Nos dois exemplos demonstrados, escolhidos entre os 10 documentos analisados na pesquisa, ficou evidente que nem todos os termos foram passíveis de representatividade, que permitisse a interligação aos conteúdos informacionais com o *DBpedia*. Ademais, apesar de a TAFNAVEGA possuir todas as classes que contemplam a representatividade de documentos acadêmicos, nem todas foram contempladas, uma vez que ela foi criada a partir da análise dos resumos dos documentos, elaborados pelos próprios autores, e, muitas vezes, com informações incompletas. Assim, para se atingir a plenitude do uso da tecnologia *Linked Data*, considera-se que é preciso a construção de novas *Triple Stores* para suprir a

demanda de URIs em uso pelos diversos sistemas. Apesar de as aplicações neste trabalho terem sido realizadas com URIs já disponíveis no *DBpedia*, avalia-se que seria melhor se houvesse um maior número de *datasets* e *triple stores* disponíveis.

Assim, pode-se dizer que houve limitações para a análise das contribuições do *Linked Data* como recurso para agregar valor às informações disponibilizadas na BDTD, mas considera-se que ficou demonstrado que, havendo bases de dados interoperáveis sobre temas de pesquisa, uma BDTD poderá ligar os dados dos documentos com, pelo menos, esta base específica.

7 CONSIDERAÇÕES FINAIS

Este trabalho apresentou os resultados de uma pesquisa que teve como objetivo explorar o tema e verificar a contribuição do *Linked Data* como recurso incorporado às BDTDs. O *Linked Data* é um esforço comunitário que visa prover um conjunto de dados (informações) estruturados e interoperáveis na *web*. É esperado que isso seja feito a partir de princípios e boas práticas, no sentido de identificar cada recurso de informação por um URI (identificador de dados) único e descrito em RDF (pequeno fragmento disponível na web que descreve o dado que faz referência a uma URI). Assim, visa determinar ligações de dados, entre diferentes fontes de dados na web.

Os resultados demonstraram que é possível utilizar o *Linked Data* para agregar os dados do *DBpedia*, um projeto colaborativo que está atrelado aos princípios da *Web* Semântica. Apesar das limitações verificadas, sobretudo em relação ao conteúdo temático da BDTD, que é específico sobre pesquisas científicas, foi possível criar uma teia de dados, identificados por URIs e descritos em RDF, que apontam para um tipo de objeto ou conceito da BDTD/ECI/UFMG para os dados do *DBpedia*. Considera-se que, para atingir o potencial máximo que essa tecnologia possui, é preciso a construção de novas *Triple Stores* para suprir a demanda de URIs em uso pelos diversos sistemas e nas diferentes temáticas.

Embora tenham sido percebidos os problemas e as limitações, se as BDTDs fossem capazes de fazer a gestão do conhecimento por meio do *Linked Data*, poderiam se tornar centros de excelência e ampliar suas fronteiras, trazendo benefícios aos usuários. Assim, a contribuição desta pesquisa para a Ciência da Informação foi trazer à discussão o tema do *Linked Data*, e apontar as suas potencialidades e limitações, uma vez que a integração de dados vinculados na *web*, seja em pequena seja em grande escala, e em diferentes níveis de

complexidade, é o foco das pesquisas que tratam de questões da *Web* Semântica.

AGRADECIMENTOS

Agradecimento às agências de fomento Fundação de Amparo à Pesquisa em Minas Gerais (FAPEMIG), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e à Pró-Reitoria de Pesquisa (PRPq) da Universidade Federal de Minas Gerais, pelo apoio financeiro com bolsas e ajudas de custo para participação em eventos.

REFERÊNCIAS

ALVES, M. D. D.; SOUZA, M. I. F. Estudo de correspondência de elementos metadados: Dublin Core e MARC 21. **Revista Digital de Biblioteconomia e Ciência da Informação**, v. 4, n. 2, p. 20-38, 2007.

BAKER, T. Dublin Core in multiple languages: Esperanto, Interlingua, or Pidgin. In: International Symposium on Research, Development and Practice in Digital Libraries, November 18-21, 1997, Tsukuba Science City, Japan. **Proceedings...** Tsukuba Science City, Japan: ISDL, 1997. p. 8-15.

BERNERS-LEE, T. Linked Data. **W3C Publishing Summit**, San Francisco, California, jul. 2006. Disponível em: <https://goo.gl/jK6GwE>. Acesso em: 22 jul. 2017.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. **Scientific American,** v. 284, n. 5, p. 34-43, 2001.

CAPLAN, P. You call it corn, we call it syntax-independent metadata for document-like objects. **Public Access-Computer Systems Review**, v. 6, n. 4, 1995.

CATARINO, M. E.; SOUZA, T. B. A representação descritiva no contexto da web semântica. **Transinformação**, Campinas, v. 24, n. 2, p. 77-90, ago. 2012.

CATARINO, M. E.; CERVANTES, B. M. N.; ANDRADE, I. A. A representação temática no contexto da web semântica. **Informação & Sociedade: Estudos**, João Pessoa, v. 25, n. 3, p. 105-116, set./dez. 2015.

CUNHA, M. B. Das bibliotecas convencionais às digitais: diferenças e convergências. **Perspectivas em Ciência da Informação**, v. 13, n. 1, p. 2-17, 2008.

DIAS, T. D; SANTOS, N. Web Semântica: conceitos básicos e tecnologias associadas. **Cadernos do IME Série Informática**, v. 14, p. 25-38, 2003.

HEATH, T.; BIZER, C. Linked data: evolving the web into a global data space. **Synthesis lectures on the semantic web:** theory and technology, v. 1, n. 1, p. 1-136, 2011.

HELLMANN, S. et al. DBpedia: a large-scale, multilingual knowledge base extracted from Wikipedia. **Semantic Web Journal**, v. 1, p. 1-29, 2014.

HOVY, E.; NAVIGLI, R.; PONZETTO, S. P. Collaboratively built semi-structured content and Artificial Intelligence: the story so far. **Artificial Intelligence**, v. 194, p. 2-27, jan. 2013.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 8601:** Elementos de dados e formatos de intercâmbio para representação e manipulação de datas e horas. Genebra, Suíça: Comitê Técnico ISO, 1988.

KONTOKOSTAS, D. et al. Test-driven evaluation of linked data quality. In: XXIII INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 23., April 7-11, Seoul, Republic of Korea, 2014. **Proceedings…** Seoul, Republic of Korea, ACM, 2014. p. 747-758.

LAUFER, C. **Guia da Web Semântica**. [Online]. São Paulo: Centro de Estudos sobre Tecnologia Web – CeWeb.br, 2015. Disponível em: <http://ceweb.br/guias/web-semantica//>. Acesso em: 9 abr. 2017.

MACULAN, B. C. M. S. **Taxonomia Facetada Navegacional**: construção a partir de uma matriz categorial para trabalhos acadêmicos. 2011. 191 f. Dissertação (Mestrado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2011.

MORSEY, M. DBpedia and the live extraction of structured data from Wikipedia. **Program:** Electronic Library and Information Systems, v. 46, n. 2, p. 157–181, 2012.

SEQUEDA, J. Introduction to: triplestores. **Data Education for Business and IT Professionals**, **Datadiversity**, 31 jan. 2013. Disponível em: <http://www.dataversity.net/introduction-to-triplestores/>. Acesso em: 27 jul. 2017.

SOUZA, R. R.; ALVARENGA, L. A Web Semântica e suas contribuições para a ciência da informação. **Ciência da Informação**, Brasília, v. 33, n. 1, p. 132-141, jan./abr. 2004.

SOUZA, T. B.; CATARINO, M. E.; SANTOS, P. C. Metadados: catalogando dados na internet. **Transinformação**. v. 9, n. 2, p. 93-105, maio/ago. 1997.

SOUZA, J. O.; SANTARÉM SEGUNDO, J. E. S. Mapeamento de problemas de qualidade no Linked Data. **JADI**, Marília, v. 1, p. 38-45, 2015.

WEBER, C. **Construção de um corpus anotado para classificação de entidades nomeadas utilizando a Wikipédia e a DBpedia.** 2015. 84 f. Dissertação (Mestrado), Faculdade de Informática, PUCRS, Porto Alegre, 2015.