UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE CIÊNCIA DA INFORMAÇÃO

Umberto Lima Diniz

ESTUDO SOBRE A INCORPORAÇÃO DE DADOS EM BIBLIOTECAS DIGITAIS DE TESES E DISSERTAÇÕES UTILIZANDO *LINKED DATA*

Belo Horizonte
2017

Umberto Lima Diniz

ESTUDO SOBRE A INCORPORAÇÃO DE DADOS EM BIBLIOTECAS DIGITAIS DE TESES E DISSERTAÇÕES UTILIZANDO *LINKED DATA*

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Informação, da Escola de Ciência da Informação da Universidade Federal de Minas Gerais, para obtenção do grau de Mestre em Ciência da Informação.

Linha de Pesquisa: Organização e Uso da Informação.

Orientadora: Gercina Ângela de Lima

Co-orientadora: Benildes Coura Moreira dos Santos Maculan

Belo Horizonte

2017

D585e Diniz, Umberto Lima.

Estudo sobre a incorporação de dados em bibliotecas digitais de teses e dissertações utilizando linked data [manuscrito] / Umberto Lima Diniz. – 2017.
85 f., enc : il.

Orientadora: Gercina Ângela de Lima.
Coorientadora: Benildes Coura Moreira dos Santos Maculan.
Dissertação (Mestrado) – Universidade Federal de Minas Gerais, Escola de Ciência da Informação.
Referências: f. 80-85.

1. Ciência da informação – Teses. 2. Bibliotecas digitais – Teses. 3. Linked data – Teses. 4. Recuperação da informação – Teses. I. Título. II. Lima, Gercina Ângela Borém de Oliveira. de. III. Maculan, Benildes Coura Moreira dos Santos. IV. Universidade Federal de Minas Gerais, Escola de Ciência da Informação.

CDU: 02:004

Ficha catalográfica: Biblioteca Profª Etelvina Lima, Escola de Ciência da Informação da UFMG.

UFMG

Universidade Federal de Minas Gerais
Escola de Ciência da Informação
Programa de Pós-Graduação em Ciência da Informação

FOLHA DE APROVAÇÃO

"ESTUDO SOBRE A INCORPORAÇÃO DE DADOS EM BIBLIOTECAS DIGITAIS DE TESES E DISSERTAÇÕES UTILIZANDO LINKED DATA"

Umberto Lima Diniz

Dissertação submetida à Banca Examinadora, designada pelo Colegiado do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Minas Gerais, como parte dos requisitos à obtenção do título de "**mestre em Ciência da Informação**", linha de pesquisa "**Organização e Uso da Informação**".

Dissertação aprovada em: 20 de junho de 2017.

Por:

Profa. Dra. Gercina Ângela Borém de Oliveira Lima - ECI/UFMG (Orientadora)

Profa. Dra. Benildes Coura M do S Maculan - ECI/UFMG (Co-orientadora)

Prof. Dr. Luiz Cláudio Gomes Maia - FUMEC

Prof. Dr. Elisangela Cristina Aganette - ECI/UFMG

Aprovada pelo Colegiado do PPGCI

Profa. Alcenir Soares dos Reis
Coordenadora

Versão final aprovada em 29/12/2017.

Profa. Gercina Ângela Borém de Oliveira Lima
Orientadora

UFMG

Universidade Federal de Minas Gerais
Escola de Ciência da Informação
Programa de Pós-Graduação em Ciência da Informação

ATA DA DEFESA DE DISSERTAÇÃO DE **UMBERTO LIMA DINIZ**, matrícula: 2014655264

Às 14:00 horas do dia 20 de junho de 2017, reuniu-se na Escola de Ciência da Informação da UFMG a Comissão Examinadora aprovada *ad referendum* pelo Colegiado do Programa de Pós-Graduação em Ciência da Informação em 17/05/2017, para julgar, em exame final, o trabalho intitulado ***Estudo sobre a incorporação de dados em bibliotecas digitais de teses e dissertações utilizando Linked Data***, requisito final para obtenção do Grau de MESTRE em CIÊNCIA DA INFORMAÇÃO, área de concentração: Produção, Organização e Utilização da Informação, Linha de Pesquisa: Organização e Uso da Informação. Abrindo a sessão, a Presidente da Comissão, Profa. Dra. Gercina Ângela Borém de Oliveira Lima, após dar conhecimento aos presentes do teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato para apresentação de seu trabalho. Seguiu-se a argüição pelos examinadores com a respectiva defesa do candidato. Logo após, a Comissão se reuniu sem a presença do candidato e do público, para julgamento e expedição do resultado final. Foram atribuídas as seguintes indicações:

Profa. Dra. Gercina Ângela Borém de Oliveira Lima - Orientadora	APROVADO
Profa. Dra. Benildes Coura M do S Maculan (co-orientadora)	APROVADO
Prof. Dr. Luiz Cláudio Gomes Maia	APROVADO
Prof. Dr. Elisangela Cristina Aganette	APROVADO

Pelas indicações, o candidato foi considerado APROVADO.

O resultado final foi comunicado publicamente ao candidato pela Presidente da Comissão. Nada mais havendo a tratar, a Presidente encerrou a sessão, da qual foi lavrada a presente ATA que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 20 de junho de 2017.

Profa. Dra. Gercina Ângela Borém de Oliveira Lima (ECI/UFMG)

Profa. Dra. Benildes C M do S Maculan (co-orientadora) (ECI/UFMG)

Prof. Dr. Luiz Cláudio Gomes Maia (FUMEC)

Prof. Dr. Elisangela Cristina Aganette (ECI/UFMG)

Obs: Este documento não terá validade sem a assinatura e carimbo da Coordenadora.

Profa. Alcenir Soares dos Reis
Coordenadora do Programa de Pós-Graduação em Ciência da Informação da UFMG

DEDICATÓRIA

Dedico este trabalho à minha família, obrigado pelo apoio. Em especial ao meu pai (*in memorian*) obrigado por tudo, fique com Deus, sentiremos saudade.

Dedico à minha noiva Walkiria, obrigado por toda cumplicidade e companheirismo, sem você esta estrada seria impossível.

Dedico também à minha orientadora Profa. Dra. Gercina Lima, sou muito grato pela confiança, paciência, incentivo e excelente orientação.

Dedico ainda à minha Co-orientadora Profa. Dra. Benildes Maculan, obrigado por toda a ajuda e apoio.

“Faça elevar o cosmo no seu coração.”

Saint Seyia

RESUMO

Esta pesquisa teve como objetivo norteador apresentar um estudo sobre a possibilidade de uso do Linked Data para tratamento de informação de bibliotecas digitais de teses e dissertações. Linked Data é um conjunto de boas práticas propostas por Tim Berners-lee *et a*l (2001). A biblioteca digital avaliada foi a biblioteca digital de teses e dissertações (BDTD) da UFMG. Para tanto, fez-se uso do trabalho de Maculan (2011), a partir da seleção feita pela autora foi feita uma nova seleção aleatória de um conjunto de 10 documentos os quais foram utilizados para este estudo. Os documentos selecionados para esta pesquisa foram localizados na base de dados da Biblioteca Digital de Teses e dissertações e identificados com uma URI fornecida pela própria BDTD. A ontologia criada por Maculan forneceu as classes e termos para relacionar aos documentos e estas foram utilizadas como elementos para 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 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 deu por seu caráter multidisciplinar e ampla disponibilidade.

Palavras Chave: Linked Data. Biblioteca Digital.

ABSTRACT

The aim of this research was to present a study on the possibility of using Linked Data to process information from digital libraries of theses and dissertations. Linked Data is a set of good practices proposed by Tim Berners-lee et al (2001). The digital library evaluated was the digitial library of theses and dissertations (BDTD) of UFMG. In order to do so, the work of Maculan (2011) was used, from the selection made by the author was made a new random selection of a set of 10 documents which were used for this study. The documents selected for this research were located in the Biblioiteca Digital database of theses and dissertations and identified with a URI provided by BDTD itself. The ontology created by Maculan provided the classes and terms to relate to the documents and these were used as elements 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 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 was due to its multidisciplinary character and wide availability.

Keywords: Linked Data, Digital Library

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE SIGLAS

TIC	Tecnologias da Informação e Comunicação
CI	Ciência da Informação
BD	Biblioteca Digital
BDTD	Biblioteca Digital de Teses e Dissertações
RDF	Resource Description Framework
XML	eXtensible Markup Language
OWL	Web Ontology language
PDF	Portable Document Format
SWEO	Semantic Web Education and Outreach
IBICT	Instituto Brasileiro de Informação em Ciência e Tecnologia
UFMG	Universidade Federal de Minas Gerais
NDLTD	*Networked Digital Library of Thesis and Dissertations*
BCI	Biblioteconomia e Ciência da Informação
DC	Dublin Core
DCMI	Dublin Core Metadata Initiative
OPAC	Online Public Access Catalog
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
LD	Linked Data
LOD	Linked Open Data
HTTP	Hypertext Transfer Protocol
HTML	*HyperText Markup Language*
SKOS	*Simple Knowledge Organization System*
SPARQL	*Simple Protocol and RDF Query Language*
XHTML	*Extensible Markup Language*
DBMS	*Database management Systems*
RDBMS	*Relational Database Management Systems*
OUI	Organização e Uso da Informação
PPGCI	Programa de Pós-Graduação em Ciência da Informação

SUMÁRIO

1 INTRODUÇÃO

A popularização do uso dos computadores auxiliou diversas áreas do conhecimento no registro informacional de sua produção científica. A evolução das tecnologias de informação e comunicação (TICs) possibilitou às bibliotecas armazenar, organizar, representar e recuperar informações de seus acervos em suportes digitais, enfim, permitiu a elas desenvolver novas formas de tratar a informação para melhor atender às suas funções. De acordo com Russo (2010, p. 71): “As bibliotecas […] têm, basicamente, duas finalidades principais: a) atender às necessidades de seus usuários e b) procurar facilitar o acesso, de forma rápida e ótima, à informação por eles solicitada”. O impacto das mudanças tecnológicas, no âmbito da Biblioteconomia e da Ciência da Informação (CI), atingiu aos processos de tratamento da informação, catalogação, indexação e classificação, tendo início já na década de 50, por meio de experimentos de indexação automática e, mais adiante, com a implantação de bibliotecas digitais (BDs).

As Bibliotecas Digitais (BD), criadas na década de 90, podem conter diversos tipos de suportes informacionais (mídias informacionais), sobre os mais variados assuntos, sem haver a separação pelo tipo de suporte em que está o conteúdo. As BDs podem ser construídas com diversos propósitos, abrangendo coleções muito diversificadas, dentre as quais destacamos as Bibliotecas Digitais de Teses e Dissertações (BDTDs), utilizadas para divulgação, acesso e recuperação da produção científica dos cursos de pós-graduação *Stricto Sensu*. Assim, as BDTDs são tipos de repositórios que permitem o acesso a esse material acadêmico, e, para tanto, utilizam bases de dados, padrões de metadados e formulários de busca. D mesma forma que as bibliotecas físicas fizeram com o universo informacional, organizando-os e agrupando-os entre suas paredes, de modo a permitir o acesso fácil aos interessados, as BDTDs fazem com esse novo universo documental, agora em formato digital, tornando o material acessível a qualquer interessado por meio da Internet.

A rede mundial de computadores, como também é chamada a internet, permite que alguém acesse uma biblioteca digital localizada em qualquer lugar do planeta. Porém, as informações disponíveis na Internet muitas vezes não possuem um padrão claro e específico que permita a interoperabilidade dos dados. Na maior parte das BDTDs, por exemplo, a obra é incluída em formato *Portable Document*

Format (PDF), desenvolvido pela *Adobe Systems*, em 1993 (ADOBE, 2008). Esse tipo de formato (extensão do documento), muitas vezes torna o conteúdo (dados) do item informacional inacessível aos buscadores da *web*. Nos últimos anos, esse tem sido um dos itens mais críticos para quem pensa no desenvolvimento e na operação de sistemas de repositórios e de bibliotecas digitais distribuídos funcionando na rede (SAYÃO; MARCONDES, 2008, p. 1).

Pensando em uma solução para a recuperação eficiente de informação que pudesse minimizar a perda de dados disponibilizados na web, Tim Berners-Lee desenvolveu a proposta da *Web* Semântica (ou *Web* de Dados). A ideia foi fazer uma extensão da *web*, de maneira que os conteúdos das páginas e seus dados pudessem ser semanticamente estruturados, com significação de dados, facilitando o compartilhamento e o reuso de informações entre humanos e computadores (BERNERS-LEE et al., 2001). Nessa perspectiva, a comunidade internacional *World Wide Web Consortium* (W3C) desenvolveu aplicações tecnológicas visando à criação padrões, protocolos e diretrizes em formato aberto, tais como o *Resource Description Framework* (RDF), o *eXtensible Markup Language* (XML) e o *Web Ontology Language* (OWL).

Conforme afirma Santarém Segundo (2014, p.3864), essas tecnologias “estão diretamente relacionadas ao processo de construção da informação e armazenamento das mesmas, constituindo assim ambientes que possam ter conjunto de dados ligados semanticamente”

Para alcançar tal objetivo, de conectividade e interoperabilidade de dados na web, Berners-Lee (2006) estabeleceu o *Linked Data*, que foi adotado pelo grupo *Semantic Web Education and Outreach* (SWEO), ligado à W3C. Resumidamente, o *Linked Data* é um conjunto de boas práticas a serem seguidas para a implementação e a publicação de dados na *web*, de modo que permita o compartilhamento e o acesso de conteúdos entre pessoas e máquinas (BERNERS-LEE, 2006). Para tanto, a integração de dados é feita a partir de regras, normas de formatos (extensões) e metadados (para a descrição dos aspectos semânticos dos dados), facilitando o compartilhamento e o reuso de informações advindas de fontes diversas.

1.2 Problema e justificativa

Hoje, existem diversas coleções digitais disponíveis na internet e acessíveis a quaisquer pessoas. Entretanto, muitas delas não são interoperáveis (padrões abertos), e são, geralmente, hospedadas em diferentes instituições. A falta de interoperabilidade entre as diversas coleções cria uma limitação na reutilização dos seus dados, mesmo quando estão disponíveis na internet. Como exemplo, temos as BDTDs, que, atualmente, ainda utilizam sistemas que não permitem o fácil reuso de seus dados e de seu conteúdo informacional.

Com o aumento dos projetos de implantação de Bibliotecas Digitais de Teses e Dissertações (BDTD), principalmente com a Portaria 13/2006-CAPES, pode-se notar a necessidade e a importância das BDTDs. Essa portaria teve como objetivo a obrigatoriedade da disponibilização do material em formato digital, porém, não especifica a forma de organização desse recurso informacional, a fim de garantir a "divulgação da ciência em geral e o saber produzido pelos programas de pós-graduação" (MACULAN, 2011, p. 19).

O Portal do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT):

> [..] tem por objetivo integrar em um único portal, os sistemas de informação de teses e dissertações existentes no país e disponibilidade para os usuários um catálogo nacional de teses e dissertações em texto integral, possibilitando uma forma única de acesso a esses documentos (IBICT, 2015).

Seguindo a mesma linha, mas com um escopo limitado às suas publicações, a Universidade Federal de Minas Gerais possui uma Biblioteca Digital de Teses e Dissertações que tem como objetivo disponibilizar:

> [...] as produções científicas defendidas nos programas de pós-graduação *stricto sensu* da UFMG. Seu acervo inclui registros correspondentes a textos completos, capítulos, referências e resumos de teses e dissertações, em formato eletrônicos, devidamente autorizados pelos autores (UFMG, 2015).

O repositório da BDTD UFMG permite que essas produções estejam "disponíveis para consulta e download" (UFMG, 2015). Existe, desde 2004, um convênio da UFMG com o IBICT para associar a Biblioteca Digital de Teses e

Dissertações à *Networked Digital Library of Thesis and Dissertations* (NDLTD, www.ndltd.org).

A organização desses recursos eletrônicos se dá com a utilização de padrões de metadados, que são dados codificados e estruturados que descrevem a característica de recursos de informação. O uso desses padrões possibilita aos sistemas de informação e de gestão do conhecimento, a integração e o compartilhamento de recursos e aplicações.

Essa integração de metadados ocorre com o uso do padrão Dublin Core. O Dublin Core possui, em seus 15 elementos, um conjunto básico de metadados que permite o *harversting* (técnica utilizada para varrer as páginas na web em busca de dados). Entretanto, somente esse conjunto de campos não consegue contemplar uma recuperação semântica apropriada para estruturas mais complexas. Para tanto, alguns estruturalistas propõem qualificadores que têm como propósito estender a semântica dos elementos do Dublin Core (BAKER, 1997, p. 2).

Com o advento da web semântica, criou-se a expectativa da possibilidade de criação de uma teia de dados relacionados semanticamente na internet. Com as tecnologias semânticas, constituídas pelo conjunto de famílias que buscam o significado dos dados e da informação (ALLEMANG; HENDLER, 2011), houve condições para promover a ligação entre sistemas, permitindo a localização e reuso de conteúdo. Nesse sentido, o que há são dados apontando para dados, e não páginas levando a outras páginas: é a web de dados. Assim, nessa web, que deu origem à metodologia do *Linked Data*, torna-se possível conectar dados relacionados e melhorar o tratamento desses recursos digitais, permitindo, assim, uma interoperabilidade legível por humanos e por máquinas.

Usado em conjunto com outros padrões, como o Dublin Core, o Linked Data permite ligar dados interoperáveis entre sistemas de informação e, principalmente, a recuperação de dados por máquinas. O *Linked Data* traz consigo uma redução da complexidade e uma padronização entre procedimentos, de modo a aumentar a compatibilidade entre recursos na Internet.

Na web, já se encontram domínios com dados de diversas naturezas, que estão marcados semanticamente (*datasets*), tais como: os dados da Wikipedia, 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.

Com base nessas considerações, entende-se a importância e a relevância deste estudo na BCI (Biblioteconomia e Ciência da Informação) e na organização da informação em ambientes digitais, principalmente no que se refere à melhoria da recuperação da informação, que, no caso desta pesquisa, pode vir a ser de grande valia para a comunicação científica.

Dessa forma, considera-se que esta pesquisa se justifica pela possibilidade de uso desses *datasets* para agregar valor às informações disponibilizadas em BDTDs. Assim, tem como proposta a utilização do Linked Data, implementado em parte deste repositório, na forma de uma conversão de registros bibliográficos, para verificar a possibilidade de ampliar a forma de recuperação de informação nessas bibliotecas digitais de teses e dissertações.

Para este estudo, foram formuladas as seguintes questões: Qual a efetiva contribuição que a metodologia do *Linked Data* poderá trazer para os repositórios de teses e dissertações?

Qual é a potencialidade do *Linked Data* na recuperação de informação em uma BDTD? Como estruturar os dados relacionados aos recursos de uma BDTD, de modo a recuperar a informação esperada? Como reduzir a complexidade na busca de informações realizada de diferentes fontes de dados?

Tendo em vista a proposta e a expectativa, a seguir são apresentados o objetivo geral e os objetivos específicos desta pesquisa.

1.2.1 Objetivos

A seguir são apresentados os objetivos que nortearam a pesquisa.

1.2.1.1 Objetivo Geral

Estudar a contribuição do Linked Data como recurso incorporado à Biblioteca Digital de Teses e Dissertações da Universidade Federal de Minas Gerais, visando agregar novos dados às informações disponibilizadas.

1.2.1.2 Objetivos Específicos

1. Verificar o uso do Linked Data como mecanismo para agregar informações à

BDTD, a partir do estudo de suas características.

2. Determinar as possíveis contribuições que os novos dados, incorporados à BDTD, podem oferecer aos seus usuários.
3. 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.

1.2.2 Estrutura da Dissertação

Esta dissertação está estruturada em 5 capítulos, que foram organizados como segue abaixo:

Capítulo 1 - Introdução: introduz o tema de pesquisa, destaca as mudanças tecnológicas ocorridas com o advento da web e da Internet, passando pelos princípios da Web Semântica e, finalmente, chegando ao tópico em foco, do Linked Data. Apresenta, também, o problema e as justificativas, assim como a proposta de pesquisa apontada no objetivo geral e, mais especialmente, nos objetivos específicos, incluindo também a estrutura desta dissertação.

Capítulo 2 - Referencial teórico e metodológico: expõe os fundamentos teórico-metodológicos no âmbito dos repositórios e bibliotecas digitais, uma vez que a ambientação da pesquisa ocorre na Biblioteca Digital de Teses e Dissertações (BDTD), da Universidade Federal de Minas Gerais, passando, mais especificamente, pelo padrão de metadado Dublin Core (DC), que é o padrão utilizado na BDTD/UFMG. Em seguida, aborda e descreve as características da Web de Dados, verticaliza no tema do *Linked Data* suas características, princípios e o método de avaliação “Cinco Estrelas”, que foi criado por Bernes Lee, em 2010, para, finalmente, descrever o projeto do DBpedia, cujo banco de dados foi utilizado para a criação das triplas de dados.

Capitulo 3 – Metodologia: apresenta as características gerais da pesquisa, com os insumos utilizados, a definição do recorte e o objeto empírico manuseado. Em seguida, descreve, detalhadamente, os procedimentos metodológicos empregados na construção do objeto investigado juntamente com as análises dos resultados.

Capítulo 4 - Descrição dos procedimentos e análise dos resultados:

Detalha os procedimentos da análise da amostra, composta por 10 documentos, identificando as classes, termos e relacionamentos. Apresenta os identificadores de classes e os retrata através de quadros comparativos entre os elementos da taxonomia utilizada e os elementos do Dublin Core, além da elaboração das triplas que foram possíveis de serem criadas com apontamento para a URIs da BDPedia. Apresenta, também, os resultados dos links data criados, fazendo uma análise dos resultados alcançados.

Capitulo 5 – Considerações Finais: as considerações finais revisitam os objetivos da pesquisa, buscando responder a todas as questões levantadas e apontar os principais resultados, as limitações percebidas, os trabalhos futuros e a contribuição para o avanço dentro do campo de estudo da Ciência da Informação.

2 REFERENCIAL TEÓRICO E METODOLÓGICO

Aqui estão apresentados os fundamentos teórico-metodológicos em que se apoiam esta pesquisa. A estrutura deste capítulo se apresenta da seguinte forma: A subseção 2.1 cita os conceitos básicos de repositórios digitais e, também, de Bibliotecas Digitais; a subseção 2.2 aborda os Metadados, apresentando seu histórico, os padrões e apresenta o modelo que será utilizado nesta pesquisa. A subseção 2.3 descreve a Web de Dados. A subseção 2.4 discorre sobre o *Linked Data*, nosso objeto de estudo, seu histórico, utilização e função.

2.1 Repositórios e Bibliotecas Digitais

O sonho da criação de ambientes com acesso e alcance mundiais, com capacidade para armazenar todo o conhecimento humano, há tempos alimenta leitores em todo o mundo. No final do século XIX, foi criado, em Bruxelas, o Instituto Internacional de Bibliografia, que tinha o objetivo de registrar em fichas catalográficas a produção mundial de impressos, chamado Repertório Bibliográfico Universal. Criado pelos advogados belgas Henri La Fontaine e Paul Otlet, tivera também a contribuição do brasileiro Manuel Cícero Peregrino da Silva que enviou dados sobre a produção bibliográfica brasileira (FONSECA, 2007). Ao se falar da biblioteca universal, proposta por Otlet e La Fontaine em 1908, pouco se fala sobre os motivos de seu fracasso na época em que foi proposta. As dificuldades de manutenção, falta de padronização e homogeneização de informações vindas de países diversos e de culturas variadas se tornaram a causa deste fracasso.

No final da Segunda Guerra Mundial, Vannevar Bush, então assessor científico do governo norte-americano, publicou, em seu artigo intitulado “As we may think” de 1945, o que seria sua visão de uma máquina capaz de articular ideias e criar respostas a perguntas. O Memex, nome dado a essa máquina imaginária, facilitaria a recuperação e a disseminação da informação científica, além de armazená-la para uso posterior. Essa foi, possivelmente, a primeira ideia de um repositório digital ou de uma biblioteca digital.

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. O grau de controle do conteúdo disponível normalmente

depende de seu público, no caso de um controle mais geral, a exemplo da Wikipédia, têm um público mais geral, mas, para o caso de um controle mais específico, espera-se que seu público seja mais característico, como estudantes. Os repositórios digitais são muito utilizados em armazenamento da produção científica de uma instituição e são conhecidos também, na literatura da Biblioteconomia e da Ciência da Informação, como Repositórios Institucionais.

Existem diversos programas específicos para gerenciar os repositórios digitais, os mais utilizados são o *E*-Prints, desenvolvido pela Universidade Southampton, e o Dspace, (...) sendo o segundo o mais usado em todo o mundo (FICHE *et al*, 2017, p. 61). O "Dspace é um sistema gerenciador de bibliotecas digitais desenvolvido pelo grupo de bibliotecas do MIT (*Massachusetts Institute of Technology*) em parceria com a empresa *Hewlett Packard* (HP), em novembro de 2002" (PONTES e LIMA, 2012, p. 32). Esse sistema permite gerenciar os mais variados formatos digitais como, TIFF, AIFF, XML ou publicados em especificações PDF ou RIFF1, no final do mesmo ano de seu lançamento, em novembro de 2002, já haviam mais de 1500 instituições que utilizavam esse recurso. No Brasil, vários órgãos públicos são usuários desse sistema, dentre eles: a Biblioteca Digital Jurídica (BDJUr), Biblioteca do Senado Federal (BLATTMANN, 2008) e a BDTD da Universidade Federal de Minas Gerais, que é parte da aplicação deste estudo.

A Biblioteca Digital (BD), também conhecida como biblioteca eletrônica, biblioteca sem paredes ou, ainda, biblioteca cibernética, tem, então, sua origem na ideia visionária do "Memex", porém, ao longo dos anos e com o desenvolvimento das tecnologias da informação, a criação de bibliotecas digitais se tornou possível e sua função foi se tornando mais consolidada. De acordo com Cunha (1999, p. 258) "o conceito de biblioteca digital aparenta algo revolucionário, mas, na verdade, ele é resultado de um processo gradual e evolutivo." O que se pode traduzir como uma mudança constante, que acompanha a biblioteca em sua evolução e aprimoramento de seus serviços e acervos.

A biblioteca digital, ainda que em um contexto novo, mantém as mesmas características quanto à seleção e tratamento de seu acervo, se comparado à biblioteca tradicional. Sendo necessário: seleção, organização e o tratamento, análise de consultas, desenvolvimento de estratégias de busca, realização de

[1] RIFF (Resource Interchange File Format) é um arquivo estruturado para recursos multimídia. Especificamente falando, RIFF não é um formato de arquivo, mas uma estrutura de arquivo.

buscas e disseminação. Dentre as várias etapas do tratamento da informação, Dias (2001) destaca a seleção como um dos pilares de qualquer sistema de informação, e afirma que é uma função fundamental à própria concepção desses sistemas. Atualmente “as bibliotecas digitais se impõem como um fenômeno que pode vir a minorar alguns dos problemas enfrentados pelos que pretendem resolver suas necessidades de informação por meio do contexto digital” (DIAS, 2001, p. 1). As bibliotecas Digitais possuem diversas características que as diferem e outras que as aproximam das Bibliotecas Tradicionais. Cunha (2008), por sua vez, nos apresenta quatro aspectos principais quando compara essas unidades informacionais: quanto à organização da informação; quanto ao acesso à informação; quanto ao aspecto econômico; e quanto às ações cooperativas.

Como já sabemos, a biblioteca convencional é aquela em que a maioria dos itens é constituída de documentos em papel, ela existe desde a criação da escrita e antes da invenção dos tipos móveis, no século XV, entretanto, até esse período era comum encontrar, nos acervos de bibliotecas, também outros materiais, como tabletes de argila, pergaminhos e papiros. A biblioteca convencional, normalmente, utiliza papel tanto para o catálogo quanto para os documentos (CUNHA, 2008, p. 5). Esse tipo de biblioteca passou por evoluções ao longo dos tempos. O catálogo que, inicialmente, era em formato de livro, foi substituído por fichas de papel, e posteriormente por registro em uma base de dados de computador. Com o advento e melhoria das linhas telefônicas, utilizadas para transmissão de dados em baixa velocidade, o catálogo foi novamente atualizado para sua versão online, conhecido como OPAC (*On line Public Access Catalog*).

Como uma evolução da biblioteca convencional, a biblioteca digital combina estrutura e coleta de informação com o uso da representação digital possibilitado pela informática. Assim como a biblioteca convencional, a biblioteca digital também mantém os mesmos princípios, já alicerçados em como a informação é organizada, criados pela Biblioteconomia e Ciência da Informação.

Biblioteca Digital é um conceito recente, e, de acordo com Cunha (2008), ela precisa ter conteúdo, que pode ser material antigo, convertido do para o formato digital, ou material novo, nascido digitalmente. Seu acervo pode ser formado por itens comprados, doados, trocados ou digitalizados localmente, a partir de documentos que não mais estão presos aos princípios legais do direito autoral. Semelhantes às bibliotecas tradicionais, já consagradas, as bibliotecas digitais:

> [...] têm características únicas que diferem das bibliotecas tradicionais e suas abordagens para o fornecimento de informações. A visão evolucionária das bibliotecas digitais tem sido abordada por profissionais na Biblioteconomia e Ciência da Informação [...]. Do ponto de vista de um bibliotecário tradicional, bibliotecas digitais apresentam um modelo de transformação de uma grande escala organização, centrada no usuário que está se movendo em direção a uma forma integrada com vários componentes. No entanto, o principal objetivo das bibliotecas digitais permanece consistente com o das bibliotecas tradicionais em que a finalidade das bibliotecas digitais é organizar, distribuir e preservar os recursos de informação assim como é para as bibliotecas tradicionais (CHOI; RASMUSSEM, 2006, p. 1).

As semelhanças terminam quando se compara a uma funcionalidade na qual a biblioteca tradicional jamais será capaz de superar a biblioteca digital e, é este seu principal diferencial, a capacidade de entregar o documento na mesa do usuário (CUNHA, 2008). Através das linhas de comunicação existentes atualmente, como a Internet, a biblioteca digital pode ser acessada virtualmente de qualquer lugar do planeta, limitada apenas pela infraestrutura de conectividade disponível no local.

Dentre as várias possibilidades de tecnologias para as bibliotecas digitais, os mais comuns são arquivos em formato PDF, contendo texto, organizados em bases de dados. As bases de dados fornecem um link para o texto completo dos documentos, permitindo seu download através da internet.

Dos tipos de bibliotecas digitais existentes, destacam-se as Bibliotecas Digitais de Teses e Dissertações (BDTD), que são um tipo de repositório digital, e fornecem acesso às publicações (Teses e Dissertações) produzidas durante cursos de Pós-graduação. Esses repositórios são, normalmente, organizados utilizando normas e padrões de metadados, tal como a BDTD da UFMG, que usa o Dublin Core (DC). Tendo em vista que não há a obrigatoriedade de utilização de todos os metadados, cada sistema pode escolher os padrões que forem mais apropriados para sua própria coleção.

2.2 Normas e Padrões de Metadados

O objetivo principal dos padrões de metadados é auxiliar a recuperação da informação, prover a interoperabilidade entre os recursos de informação na Internet. De modo simplista, metadados são dados sobre dados. Em outras palavras, o metadado aponta para o dado que deve ser informado sobre o conteúdo de um

documento (item de informação).

Por exemplo, um título em um livro é um metadado, o nome do intérprete em um arquivo de música em MP3 é um metadado. Mais especificamente os metadados são conjuntos de atributos, ou dados referenciais, que representam o conteúdo informacional de um recurso que pode estar em meio eletrônico ou não. Já os formatos de metadados, também chamados de padrões de metadados, são estruturas padronizadas para a representação do conteúdo informacional que será representado pelo conjunto de dados-atributos (metadados) (ALVES, 2005, p. 115)

Padrões de metadados são ferramentas adjacentes à Web Semântica, que permitem a melhor organização e recuperação de documentos em bibliotecas digitais (CASTRO, SANTOS 2007), em uma linguagem mais popular, seriam as normas de organização. Como afirma Castro e Santos (2007, p. 15), “os bibliotecários produzem e padronizam metadados há séculos, desde as primeiras tentativas de organização da informação a partir da descrição de documentos”.

De fato, possuímos um incontável número de padrões de definição para conjuntos de elementos metadados. No entanto, vale destacar que a maioria dos recursos disponíveis na Web não possui metadados associados a eles de maneira alguma. Também se pode destacar, como uma importante atribuição dos metadados, suprir uma necessidade de organizações conhecerem melhor os dados que elas mantêm além dos dados de outras organizações (SOUZA et al, 1997).

O metadado é estruturado com elementos de descrição de conteúdo dos dados. Cada parte da informação deve conter, por exemplo, autor, título, data de publicação, etc. Existem diversos padrões de metadados desenvolvidos para diferentes finalidades (SOUZA, 1997, p. 95). O padrão *Dublin Core* (DC) é um dos padrões de metadados utilizados para a descrição dos documentos disponíveis em uma biblioteca digital, cujas características estão descritas na próxima subseção.

2.2.1 Dublin Core (DC)

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 Dublin Core foi desenvolvido de modo a possuir diversas características que o tornam atraente 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 Dublin Core 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 a seguir.

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)

Além desses 15 elementos, pela sua característica flexível, o DC também permite a inclusão de metadados adicionais, caso seja necessária a inserção de outros elementos que atendam às particularidades requeridas pela biblioteca digital a ser criada.

Para suprir necessidades especiais, o Dublin Core pode ser personalizado com campos adicionais e existem duas classes de qualificadores para atender a essa demanda, os elementos de refinamento (tabela abaixo) e um conjunto de esquemas de codificação (Tabela 1).

Tabela 1: Lista de qualificadores

Campo	Qualificador	Nome	Etiqueta	Definição
Title	Alternativo	Alternativo	Alternativo	Qualquer forma do título ou uma alternativa para o título formal do recurso. Pode incluir abreviações assim como traduções.
Assunto	LCSH	LCSH	LCSH	*Library of Congress Subject Headings*.
Assunto	MESH	MESH	MeSH	*Medical Subject Headings*. http://www.nlm.nih.gov/mesh/meshhome.html
Assunto	DDC	DDC	DDC	*Dewey Decimal Classification*. http://www.oclc.org/dewey/index.htm
Assunto	LCC	LCC	LCC	*Library of Congress Classi*fication. http://lcweb.loc.gov/catdir/cpso/lcco/lcco.html
Assunto	UDC	UDC	UDC	*Universal Decimal classification* http://www.udcc.org/
Descrição	Table of Contents	tableOfContents	Table Of Contenes	Uma lista de subunidades do conteúdo do recurso.
Descrição	Abstract	Abstract	Abstract	Um resumo do conteúdo do recurso.
Data	Created	Created	Created	Data da criação do recurso.
Data	Valid	Valid	Valid	Data (normalmente um intervalo) da validade de um recurso.
Data	Available	Available	Available	Data (normalmente um intervalo) que um recurso se tornou ou se tornará disponível.
Data	Issued	Issued	Issued	Data da emissão formal (ex. Publicação de um recurso).
Data	Modified	Modified	Modified	Data que o recurso foi alterado.
Formato	Extent	Extent	Extente	O tamanho ou duração do recurso.
Formato	Médium	Médium	Médium	O suporte físico do material.
Linguagem	Não possui qualificadores			
Relação	Is	isVersionO	Is	O recurso descrito é uma versão,

	Version Of	f	Version Of	edição ou adaptação do referido recurso. Mudanças na versão implicam em mudanças substantivas no conteúdo em vez de diferenças no formato.
Relação	Has Version	hasVersion	HasVersion	O recurso descrito tem uma versão, edição ou adaptação, ou seja, o recurso referenciado.
Relação	Is Replaced By	isReplaced By	Is Replace By	O recurso descrito é suplantado, deslocado ou substituído pelo recurso referenciado.
Relação	Replaces	Replaces	Replaces	O recurso descrito substitui o recurso referenciado.
Relação	Is required by	isRequired By	Is Required by	O referido recurso é requerido pelo recurso referenciado, fisicamente ou logicamente.
Relação	Requires	Requires	Requires	O recurso descrito requer o recurso referenciado para suportar sua função ou coerênia no conteúdo.
Relação	Is part of	is part of	Is part of	O recurso descrito é fisica ou logicamente parte do recurso referenciado.
Relação	Has part	Hás part	Hás part	O recurso descrito inclui o recurso referenciado fisica ou logicamente.
Relação	Is Referenced By	Is Referenced by	Is referenced by	O recurso descrito é referenciado, citado ou mesmo apontado para o recurso referenciado.
Relação	References			As referências de recursos descritas, citam ou apontam para o recurso referenciado.
Relação	Is format of			O recurso descrito é o mesmo conteúdo intelectual do recurso referenciado, mas apresentado em outro formato.
Relação	Has format			O recurso descrito preexistiu o recurso referenciado, que é essencialmente o mesmo conteúdo intelectual apresentado em outro formato.
Cobertura	Spatial			Características espaciais do conteúdo intelectual do recurso.
Cobertura	Temporal			Característias temporais do conteúdo intelectual do recurso.

Fonte: http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/

Tabela 2: Esquemas de codificação dos qualificadores do DCMI

Camp	Esquema	Nome	Etiqueta	Definição	Mais informações

o DCMI					
Data	DCMI Period	Period	DCMI Period	Uma especificação dos limites de um intervalo de tempo.	<http://dublincore.org/documents/dcmi-period/>
	W3C-DTF	W3C DTF	W3C-DTF	Regras de codificação do W3C para datas e horas – um perfil baseado no ISO 8601.	<http://www.w3.org/TR/NOTE-datetime>
Tipo	**DCMI Type Vocabulary**	DCMI Type	CDMI Type Vocabulary	Uma lista de tipo usada para categorizar a natureza ou gênero do conteúdo do recurso.	<http://dublincore.org/documents/dcmi-type-vocabulary/>
Formato	**IMT**	IMT	IMT	O tipo de mídia do recurso na internet.	<http://www.isi.edu/in-notes/iana/assignments/media-types/media-types>
Identificador do Recurso	**URI**	URI	URI	Um identificador de recurso Uniforme.	<http://www.ietf.org/rfc/rfc2396.txt>
Idioma	**ISO639-2**	ISO639-2	ISSO 639-2	Código para representação dos nomes de idiomas	<http://lcweb.loc.gov/standards/iso639-2/langhome.html>
	RFC 1766	RFC1766	RFC 1766	Etiquetas RFC1766 de 2 letras para a identificação de idiomas.	http://www.ietf.org/rfc/rfc1766.txt
Relação	**URI**	URI	URI	Um identificador de recurso Uniforme.	<http://www.ietf.org/rfc/rfc2396.txt>
“Spatial”	**DCMI Point**	Point	DCMI Point	O “DCMI Point” identifica um ponto no espaço suas coordenadas geográficas.	<http://dublincore.org/documents/dcmi-point/>
	ISO3166	ISO3166	ISO3166	Códigos usados para representação de nomes de	<http://www.din.de/gremien/nas/nabd/iso3166ma/codlstp1/index.html>

				países.	
	DCMI Box	Box	DCMI Box	Identifica uma região no espaço usando seus limites geográficos.	<http://dublincore.org/documents/dcmi-box/>
	TGN	TGN	TGN	Nomes do Tesauro Getty	<http://www.getty.edu/research/tools/vocabularies/tgn/>
“Temporal”	**DCMI Period**	Period	DCMI Period	A especificação dos limites de um intervalo de tempo.	<http://dublincore.org/documents/dcmi-period/>
	W3C-DTF	W3C DTF	W3C-DTF	Regras de codificação de data e hora da W3C baseadas na norma ISSO 8601.	<http://www.w3.org/TR/NOTE-datetime>

Fonte: http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/

Como iniciativa para atender aos princípios da Web Semantica, foi criado o Dublin Core Metada 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 abaixo:

Quadro 2: Esquema DCMI com a URI

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 pelo Autor

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).

Esse padrão de metadado é bastante utilizado na Web de Dados, cujas características estão descritas na próxima subseção.

2.3 A *Web* de Dados

A evolução contínua dos sistemas nos trouxe a *World Wide Web* (WWW) ou, simplesmente, *Web*, uma criação de Tim Berners-Lee para acesso hipertextual à documentos disponíveis na internet. A Web inicial, também conhecida como “Web de Documentos”, permite uma navegação por Hiperlinks embutidos nos documentos escritos em linguagem de marcação de *HyperText Markup Language* (HTML). Entretanto, essa navegação é mais apropriada para pessoas, que podem entender uma descrição de um link textual ou um ícone. Um computador ou outro dispositivo não tem a mesma facilidade de se conectar e utilizar esses sistemas por falta de uma descrição específica do alvo para o qual o *link* aponta.

Como um processo contínuo, a estrutura da *Web* está em permanente mudança, seguindo sempre um processo evolutivo e, muitas vezes, revolucionário. A forma de organização da informação na *Web* é um exemplo de como essas mudanças ocorrem. Inicialmente, a informação é organizada de modo a permitir o compartilhamento, para pessoas, em formato textual e de imagem. Posteriormente, foi alcançado um novo nível em que é possível acessar uma grande gama de conteúdo multimídia, como vídeo e áudio. De acordo com Dias e Santos (2003), um dos objetivos da sua versão inicial, conhecida como Web de documentos, era a troca de informação entre pessoas. Esta é utilizada para publicar um documento (por publicadores) e para consumí-los (por usuários). Nessa perspectiva, o acesso a um documento é realizado por meio da especificação de uma URL, sendo possível ao usuário editar a informação contida no documento, assim como gravá-lo em dispositivos de armazenamento de dados (USB, DVD) (LAUFER, 2015).

Souza e Alvarenga (2004, p. 133) afirmam que a Web:

> 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, assim, que, com o passar do tempo, o aumento do fluxo de

informações na web evidenciou alguns problemas na recuperação eficiente e relevante de informações, o que exigiu esforços no sentido de encontrar alternativas de solução. Nesse sentido, uma grande contribuição para a Web foi a idealização da *Web* de Dados, que também é denominada como *Web* 3.0 e *Web* Semântica, que desenvolve e usa tecnologias para criar um ambiente com dados integrados que facilita a recuperação da grande quantidade de dados disponíveis na Web. Essas características foram apresentadas por Berners-Lee *et al (2001)* no artigo intitulado *The Semantic Web: a new form of Web content that is meaningful to computer will unleash a revolution of new possibilities,* na revista *Scientific American*, no qual ele destaca uma série de possibilidades de uma internet semântica, baseada em dados, agentes de software e padrões de intercâmbio de informações.

A Web de Dados, para funcionar, utiliza uma série de componentes já existentes na Web e incorpora outros elementos a esse conjunto, expressando a significação desses dados, a partir de regras e normas. Para isso, utiliza conceitos e tecnologias entre as quais se destacam, conforme Dias e Santos (2003) e Catarino, Cervantes e Andrade (2015):

1) *Extensible Markup Language* (XML): utilizada para criar *tags* em documentos, que são campos de texto que podem ser usados por programas ou *scripts*, desde que o programador saiba o seu significado, pois a tecnologia XML não permite que se represente o significado de cada estrutura.

2) *Resource Description Framework* (RDF): tem como papel dar significado às estruturas, codificando as *tags* a partir da construção de triplas, representadas por sentenças compostas de: Sujeito+Verbo+Objeto, que podem ser representadas utilizando a tecnologia XML, sendo que cada elemento da tripla é representado por um *Universal Resource Identifier* (URI). Essa tecnologia é utilizada para intercâmbio de dados.

3) Aplicações verticais (*vertical applications*): são recursos (recomendações, tecnologias e padrões) de comunidades específicas que estão disponíveis na web e utilizam as tecnologias do W3C.

4) Inferência: descoberta de novos conhecimentos a partir de dados

disponíveis na web e de informações adicionais advindas de um vocabulário (representando conceitos por meio de classes, subclasses e recursos associados a partir de diferentes relacionamentos e instâncias) ou de um conjunto de regras, representadas de maneira formal.

5) *Web Ontology Language* (OWL): tecnologia utilizada para definição e instanciação de conceitos e indivíduos, assim como classes e propriedades em vocabulários, utilizando uma semântica formal. Possui três versões, consideradas como sub-linguagens derivadas: OWL Lite, OWL DL e OWL Full.

6) *Simple Knowledge Organization System* (SKOS): tecnologia utilizada para representar a estrutura básica de diversos tipos de vocabulários ou sistemas de organização do conhecimento (SOCs), tais como tesauros, sistemas de classificação e taxonomias.

Na Web de Dados, a busca e consulta (*query*) é realizada de forma distinta das existentes em bases de dados tradicionais, tais como SQL (bases de dados relacionais) e o *XQuery* (bases de dados em XML), pois utiliza a tecnologia própria, a *Simple Protocol and RDF Query Language* (SPARQL). O SPARQL é uma tecnologia que está disponível desde 2008, e possibilita recuperar valores de dados estruturados e semiestruturados, explorando dados disponibilizados na web e consultar outras relações e recursos, fazendo conexões complexas a partir desses conjuntos de dados heterogêneos, em uma única consulta.

A partir dos dados obtidos no site da W3C, uma consulta SPARQL é composta pelos seguintes passos: (1) abreviação de URIs, para declarações de prefixos; (2) definição dos dados que serão consultados, determinando os seus grafos em RDF; (3) determinação das informações que devem ser recuperadas na busca (cláusula de resultado); (4) especificação do tipo de informação que deve ser consultada na busca (padrão de consulta); e (5) inserção de recortes para a consulta, tais como limites e ordenação dos resultados recuperados (modificadores de consulta), que irão modificar o resultado final.

Para executar a busca (série de tarefas de análise de dados e aplicativos), é preciso agregar funções tais como: (a) determinar o número de recursos distintos que irão satisfazer a critérios pré-estabelecidos; (b) calcular os valores agregados

(resulta em conjuntos de respostas menores).

Assim sendo, para que seja alcançada a proposta inicial da Web Semântica é necessário utilizar linguagens e formatos comuns para a descrição de informações, para que assim possa haver o compartilhamento de dados.

Para conseguir esse intento, desenvolveu o *Linked Data*, que é um conjunto de práticas que permitem conectar dados na web que, anteriormente, não estavam conectados. As suas características estão apresentadas na próxima subseção.

2.4 Linked Data

De acordo com W3C, os conjuntos de dados inter-relacionados na Web são denominados de *Linked Data* (Dados Ligados).

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 criar um conjunto de triplas, cada uma consistindo de um sujeito, um predicado e um objeto. 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), sendo o Sujeito e o Objeto nós, e o Predicado a relação entre eles.

Figura 1: Representação gráfica de uma tripla.

Fonte: Elaborado pelo autor

Os dados são formados por triplas que especificam os relacionamentos entre as entidades. Essas triplas são formadas por Sujeito, predicado e objeto. Cada um

destes itens é especificado através de uma URI.

A Figura 2, do site da W3C Recommendation[2] ilustra o uso dos links RDF, em um exemplo através de uma a representação gráfica, onde existe uma Pessoa identificada por http://www.w3.org/People/EM/contact#me, cujo nome é “Eric Miller”, apontando para o seu e-mail, que é “em@w3.org”, e para a sua titulação é “Dr.”:

Figura 2: Uso dos links RDF

Fonte: http://www.w3.org/TR/2004/REC-rdf-primer-20040210/

Figura 2 apresenta um grafo que descreve “Eric Miller” e ilustra que o RDF utiliza URIs para identificar:

- Indivíduos ou objetos: tal como ocorreu com “Eric Miller”, que foi identificado pela URI <http://www.w3.org/People/EM/contact#me>;
- Atributos de indivíduos: mostra um atributo, característica ou a relação que foi utilizada para descrever o recurso, tal como ocorreu com “Dr.”, atributo de “Eric Miller”, que foi identificado pela URI <http://www.w3.org/2000/10/swap/pim/contact#personalTitle>;
- Propriedade dos indivíduos ou objetos: tal como ocorreu com “em@w3.org”

[2] Fonte: https://www.w3.org/TR/2004/REC-rdf-primer-20040210/#figure1

que foi identificado pela URI <http://www.w3.org/2000/10/swap/pim/contact#mailbox>;

- Valores de propriedades: tal como ocorreu com “mailto:em@w3.org”, que mostra o recurso de uma propriedade específica “em@w3.org”, juntamente com o valor do correio eletrônico “mailto:”, podendo, também, usar como valor sequências de caracteres, tal como "Eric Miller", além de usar valores de outros tipos de dados (por exemplo, números).

Assim, o RDF fornece, também, uma sintaxe baseada em XML para apresentação desses grafos. A Figura 3 é um extrato de RDF na notação RDF/XML que corresponde ao grafo da Figura 2.

Figura 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

Nota-se, portanto, com esse exemplo, que o uso do RDF/XML, utilizando URIs, pode buscar informações em outras fontes da Web.

De acordo com Heath e Bizer (2011, p. 8, tradução livre):

> A ideia básica do *Linked Data* é aplicar a arquitetura geral da World Wide 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.

Sobre a construção dos documentos Web, eles apontam que:

> O documento Web é construído em um pequeno conjunto de padrões

> simples: Identificadores de Recursos Uniforme (URI) como um mecanismo global e único de identificação, o Protocolo de Transferência de Hipertexto (HTTP) como um mecanismo de acesso universal, e a Linguagem de Marcação de Hipertexto (HTML) como um formato de conteúdo largamente usado. Como adição, a Web é construída na ideia de que hiperlinks são usados entre documentos Web que podem estar em diferentes servidores.

Pode-se considerar que o desenvolvimento e a utilização de padrões permitem à Web transcender diferentes arquiteturas técnicas, além disso, a navegação entre diferentes servidores é possível com o uso de Hiperlinks. Estes, também, permitem a indexação da Web por motores de busca e fornecem recursos de pesquisa sofisticados no conteúdo recuperado. Neste caso, as hiperligações – como são também chamados os hiperlinks – são, portanto, decisivas para conectar e agrupar o conteúdo de diferentes servidores em um único espaço de informação global, combinando simplicidade com descentralização e abertura.

Para publicar dados na Web, os itens em uma área de interesse precisam ser identificados (HEATH e BIZER, 2011). Esses itens têm propriedades e relacionamentos que serão descritos nos dados e podem incluir tanto documentos Web quanto entidades do mundo real e conceitos abstratos

Para Berners-Lee, T., Hendler, J., Lassila O. (2001), ao contrário da estrutura hipertextual, onde 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 ror RDF, porque os URIs possibilitam a identificação de qualquer tipo de objeto ou conceito**.** Para isso, Berners-Lee (2001) propõe 4 regras:

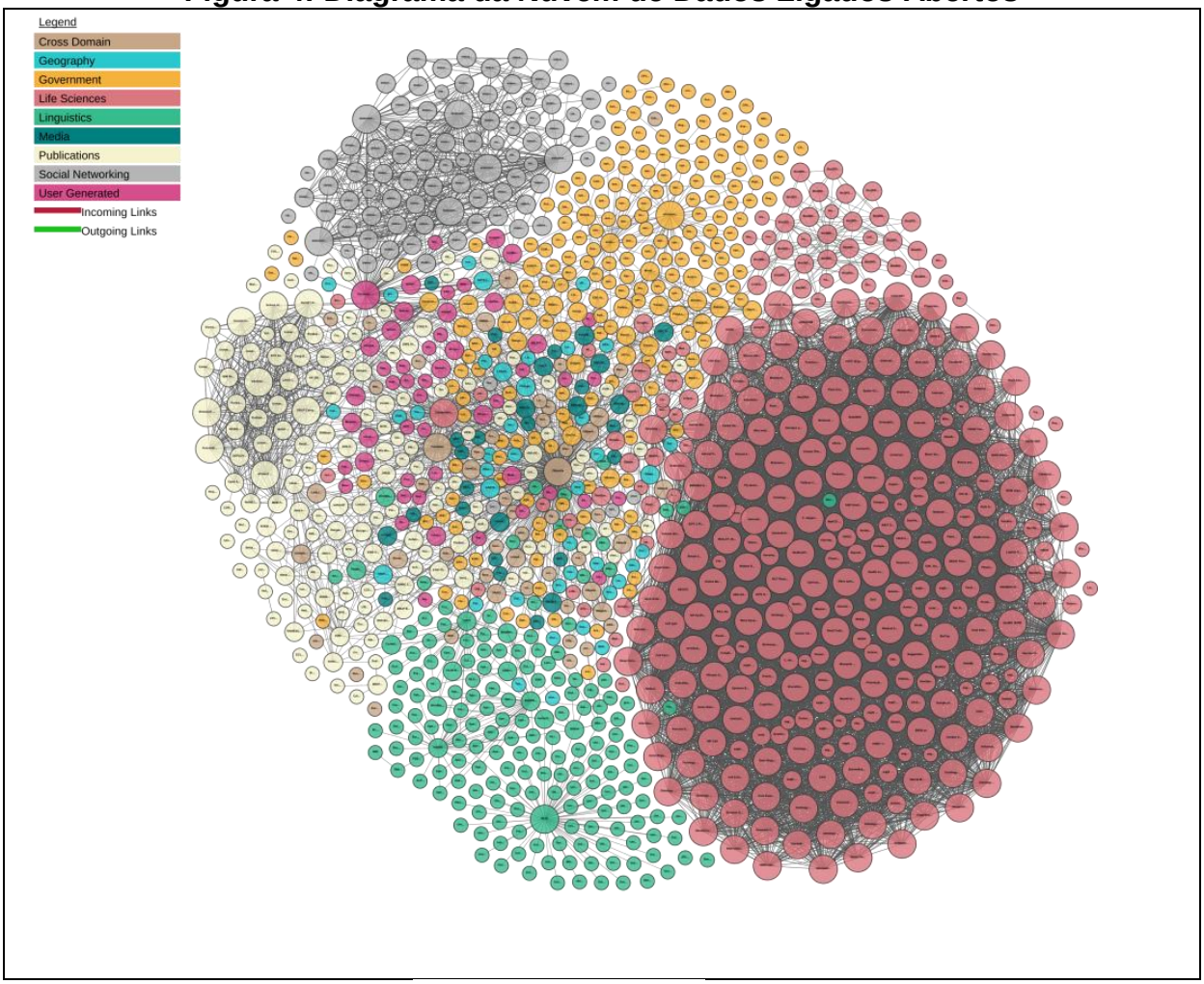
> -Use URIs como nomes para as coisas
> -Use HTTP URIs para que as pessoas possam procurar esses nomes.
> -Quando alguém procura um URI, fornecer informações úteis, usando os padrões (RDF *, SPARQL)
> -Incluir links para outros URIs. Para que eles possam descobrir mais coisas. (BERNERS-LEE, 2001)

Apesar do uso dessas regras garantir que os dados estejam interconectados, elas limitam a reutilização de maneira imprevisível, que também é uma característica da Web, que agrega valor à informação.

Um exemplo clássico de *linked data* (ou dados linkados) é a imagem que mostra um conjunto de dados que são publicados e como estão interligados com

outros conjuntos de dados, conforme Figura 4, do The Linking Open (LOD) Data Project Cloud Diagram.

Figura 4: Diagrama da Nuvem de Dados Ligados Abertos



Fonte http://lod-cloud.net/, 2016

A grande versatilidade do *Linked Data* proporciona grandes possibilidades na publicação dos dados e criação de fontes de dados de diversas origens. Com sua tecnologia aberta e facilmente disponível, surge uma variada gama de tipos de dados e de termos oriundos do conceito inicial (BIZER; CYGANIAK; GAUß, 2007).

De acordo com Pizzol (2014), diversos domínios podem ser publicados na forma de *Linked Data*, diferentes nomenclaturas são utilizadas para classificá-los. A Figura 5 destaca algumas dessas nomenclaturas, levando em consideração a natureza, origem, iniciativas e utilização dos dados.

Figura 5: Tabela de exemplos dos tipos de dados em *Linked Data*

Tipo	Estrutura			Iniciativas	
	Conteúdo	Domínio dos dados	Participantes	Propósito	Referências
Linked Government Data	Dados brutos de Agencias Governa-mentais	Público	Organizações Governamentais	- Transparência e abertura de dados. -Interligação, exploração e análise do dados.	Meskell (2007); TAYLOR (2010); Hyland e Wood (2011); Villazón-Terrazas et al., (2011)
Linked Enterprise Data	Mídias, finanças, produção, negócios, etc.	Geralmente internos a organização	Organizações Privadas	- Criação da informação está intimamente ligado com o ato de compartilhamento. - Ligação de dados internos com fontes externas, ferramenta, e novas formas de visualização.	Servant (2008); HU (2010); Allemang (2011)
Statistical Linked Data	Dados com séries estatísticas dos mais variados domínios	Público e Fechado	Pesquisadores, Organizações governamentais e privadas	- Sistemas OLAP por meio da integração estruturada de *Linked Data*, utilizando consultas em SPARQL. - Analisar grandes quantidades de dados numéricos de forma exploratória.	Cyganiak, Reynolds e Tennison (2010); Kämpgen, O'Rain e Harth (2012)
Geo Linked Data	Dados Geográficos	Público e Fechado	Pesquisadores, Organizações governamentais e privadas	- Enriquecer a Web de dados com dados ligados geoespaciais.	Auer et al., (2009); Vilches-Blázquez et al., (2010) Stadler et al., 2012
Linked Sensor Data	Proveniente de sensores	Público e Fechado	Pesquisadores, Organizações governamentais e privadas	- Dados de sensores e metadados de acesso público armazenados na nuvem de Linked Data.	Patni, Henson e Sheth (2010); Janowicz et al., (2010); Pschorr et al., (2010).

Fonte: Pizzol (2014, p. 90)

Visando clarificar sobre a natureza de cada um dos tipos de *Linked Data*, apresentam-se explanações sobre eles:

a) Government Linked Data

É importante a abertura dos dados por agentes governamentais, porém a simples publicação destes dados não resolve o problema. É necessário que estejam estruturados e legíveis, de modo a permitir a visualização, leitura, utilização e manipulação destes dados por terceiros. A aplicação dos princípios do *Linked Data* à dados governamentais traz diversos benefícios, dentre os quais destaca-se a reutilização dos dados e sua combinação com outras fontes, o que é conhecido como *mashups*[3]. Porém, a dificuldade em transformar grandes quantidades de dados em *linked data* prejudica o potencial desta aplicação.

As atividades de geração de *Linked Data* Governamental (LOGD) se assemelham ao ciclo de vida de um processo de engenharia de software, e

[3] Em software é um tipo de sistema que mistura diversas fontes de modo a gerar uma saída com um objetivo específico.

consistem em atividades principais: “(1) Especificação, (2) modelagem, (3) geração, (4) publicação e (5) exploração. Esas atividades podem ser decompostas em uma ou mais tarefas e a ordem pode ser alterada conforme a necessidade dos órgãos governamentais” (PIZZOL, 2014, p. 91).

b) Linked Enterprise Data

Inspirados na realização de projetos acadêmicos e de dados públicos, grandes empresas têm sua atenção focada na utilidade dos padrões abertos para a publicação de dados (HU; SVENSSON, 2010). A utilização dos padrões abertos pode aumentar a utilidade dos dados empresariais permitindo a criação de *mashups*.

c) Statistical Linked Data

Uma das características dos dados em *Linked Data* é serem heterogêneos e não possuírem tratamento estatístico, além disso, os navegadores e outros mecanismos de interação *linked data* não permitem que os usuários possam analisar grandes quantidades de dados numéricos de forma exploratória (PIZZOL, 2014, p. 93). Dados estatísticos apresentam as características de grandes massas de dados e para solucionar o problema é utilizada uma abordagem chamada OLAP *(Online Analytical Processing*), descrita em 1993 por W. H. Inmon, R. Kimball e E. F. Codd. O uso destas operações OLAP em grandes volumes de dados permite a visualização de grandes quantidades de dados estatísticos a partir de diferentes ângulos, granularidades, permitindo a filtragem e comparação de medidas, representando em uma interface de apoio à tomada de decisão (CHAUDHURI; DAYAL 1997; TRUJILLO, 2008 apud PIZZOL, 2010)

Duas desvantagens apontadas para o uso de OLAP são: (1) Ele requerer um modelo de cubo de dados, dimensões e medidas e (2) criar um esquema multidimensional de dados ligados genéricos é difícil. (PIZZOL, 2014, p. 94). Além disso, consultas OLAP são complexas e requerem modelos de dados especializados (GRAY et al., 1997). Pizzol (2014) ainda cita a abordagem em estrela em uso com bases relacionais que é demonstrada em Kampgem e Harte (2011), que utilizam dados de várias fontes na web em um sistema OLAP, apresentam uma forma de interagir com os dados estatísticos em um cubo modelado em RDF e permitem

consulta via SPARQL.

d) Geo *Linked Data*

Marcas em um mapa podem ser uma poderosa maneira de transmitir informação espacial, mas eles são apenas ferramentas de apresentação de aplicações da *Web* Semântica.

e) Linked Sensor Data

Um conjunto de dados RDF contendo descrições expressivas sobre dados, tal como os sobre dados metereológicos (temperatura, visibilidade, precipitação, pressão, velocidade do vento, umidade, etc.), que são disponibilizados como modelo de dados comuns compartilhados, para que os diferentes fabricantes de dispositivos e/ou sistemas possam utilizar, de forma padronizada, os mesmos sensores, num formato processável por máquina. Com isso, é possível construir algoritmos, análises e visualizações reutilizáveis. Há algumas iniciativas (SensorThings API; IOTDB.org; Ontologia de Rede de Sensores Semânticos; SensorML) que são projetos que visam descrever as capacidades e medições de uma maneira padrão, para criar modelos de referência para cada um dos diferentes tipos de sensores.

Os tipos de dados descritos anteriormente são apenas alguns exemplos de dados que podem estar disponíveis no Linked Open Data Cloud. A qualidade dos dados disponibilizados na Web podem ser avaliados a partir de alguns métodos, dentre os quais destaca-se o método de avaliação de 5 estrelas.

2.4.1 Método de avaliação Cinco estrelas

Berners Lee, em 2010, criou um sistema de cinco níveis para medir a qualidade de fontes de dados, utilizando um sistema de Cinco Estrelas, analogamente a uma avaliação de um hotel, essas estrelas são um guia para detectar o grau de reutilização dos Linked Open Data. (BARRIENTOS, S/D):

Os dados são avaliados como: uma estrela, os dados em qualquer formato, ainda que sejam difíceis de manipular como uma imagem digitalizada; duas estrelas, os dados de maneira estruturada, como um arquivo excel; três estrelas, os dados

em um formato não proprietário, como um arquivo csv; quatro estrelas, usa URI para identificar coisas e propriedades de modo que se possa apontar para os dados, usando o padrão RDF; cinco Estrelas, vincula seus dados com os de outras pessoas, colocando-os em contexto. Na prática, permite também que outras informações fornecidas apontem para outras fontes de dados, como por exemplo, a DBPedia. Esse método de avaliação foi criado para avaliar a eficiência, do ponto de vista do Linked Data, de fontes de dados disponíveis na internet.

Dentro desse método utiliza-se o conceito de Namespace. Namespaces são esquemas que descrevem as entidades semânticas. Cada entidade semântica, ao ser construída, precisa ser relacionada a algum namespace que descreve sua estrutura, e ditará suas regras. Os namespaces são um modo para amarrar um uso específico de uma palavra dentro de um contexto para o esquema onde a definição pretendida será encontrada (ZANETE, 2001, p. 27).

Outro conceito importante são os 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 as consultas são feitas em SQL, *triplestores* armazenam triplas RDF e são consultados usando SPARQL.

2.5 O DBpedia

O DBPedia é um projeto que tem como objetivo extrair o conhecimento acumulado nos diferentes artigos da Wikipédia, que é um recurso semiestruturado de informações, conforme afirmam Hovy, Navigli e Ponzetto (2013). A Wikipédia é uma enciclopédia, disponibilizada na Internet e de uso livre, onde todos os leitores podem utilizar e também atualizar seu conteúdo. Por meio da inclusão e edição de seus artigos (LASLIE, 2003). Segundo o autor, os artigos da Wikipédia estão abertos às correções, atualizações e refinamentos de conteúdos por parte dos seus próprios leitores/utilizadores.

Existem na Wikipédia diversos tipos de recurso informacional já estruturado, a exemplo das caixas (infoboxes, que são tabelas em formato fixo, orientados por um modelo pré-definido, utilizados como metadados, e apresentando atributos comuns

entre assuntos diferentes) e categorização (coleções que indicam tópicos na enciclopédia, a partir de hierarquias) de informações, as referências geográficas, imagens e, também, os links que levam a outras páginas na Web (LAUFER, 2015). A Figura 6 exemplifica os dados estruturados de um infoboxe e o modelo pré-definido para artistas:

Figura 6: Infobox do artista Devin Garret Townsend

Informação geral	
Nome completo	Devin Garrett Townsend
Nascimento	5 de Maio de 1972
Origem	New Westminster, Colúmbia Britânica
País	Canadá
Gênero(s)	Metal progressivo, metal extremo, metal industrial, música ambiente, thrash metal, death metal, grindcore, rock progressivo, punk rock, new age, música eletrônica, música clássica, country
Instrumento(s)	Guitarra, baixo, teclado, vocal, banjo
Modelos de instrumentos	Peavey, Framus, Sadowsky, ESP, Fender e Gibson
Período em atividade	1993–atualmente
Gravadora(s)	Hevy Devy Records, InsideOut Music, Century Media
Afiliação(ões)	The Devin Townsend Band, Strapping Young Lad, Steve Vai, Punky Brüster, IR8, Grey Skies, Caustic Thought, Noisescapes, Devin Townsend Project, Casualties Of Cool
Página oficial	www.hevydevy.com www.ziltoid.com

(a) Infobox

```
{{Info/Música/artista
 |nome                =
 |imagem              =
 |imagem_tamanho      =
 |imagem_legenda      =
 |fundo               =
 |nome completo       =
 |apelido             =
 |nascimento_data     =
 |nascimento_cidade   =
 |nascimento_país     =
 |origem              =
 |país                =
 |morte_data          =
 |morte_local         =
 |nacionalidade       =
 |gênero              =
 |ocupação            =
 |instrumento         =
 |instrumentos notáveis =
 |modelos             =
 |tipo vocal          =
 |período             =
 |outras ocupações    =
 |gravadora           =
 |afiliações          =
 |influências         =
 |influenciados       =
 |website             =
 |assinatura          =
}}
```

(b) Predefinição

Fonte: WEBER (2015, p.26).

Nota-se que, do lado esquerdo, há os dados estruturados do infobox do cantor, e, do lado direito, a predefinição do modelo de infobox para artistas.

O recurso de dados estruturados da Wikipédia é extraído na forma de conjuntos de triplas RDF, e disponibilizado em uma base de dados para consultas por meio de *endpoints* SPARQL, utilizando tecnologias da Web Semântica e do *Linked Data* (HELLMANN et al., 2014). A base de dados foi criada como um esforço da comunidade de *crowd-source*[4], de modo a gerar uma rede informacional de ligações, sendo um dos projetos mais famosos de disseminação de dados ligados

[4] *Crowd-source* é um termo criado em 2005, remete a projetos financiados por um grande número de pessoas, especialmente em uma comunidade *OnLine*.

(BERNERS-LEE, 2008). Com isso, a DBpedia atende, sobretudo, a dois usuários: (1) aos proprietários de dados, pois estabelece novas formas de acessar e aumentar o valor de seus dados; (2) aos desenvolvedores de programas e sistemas, uma vez que fornece dados para alimentar aplicativos e atender às necessidades dos usuários.

Atualmente, a Wikipédia possui informações em 266 idiomas, conforme estatística[5] de fevereiro de 2017, e o DBpedia extrai as suas informações estruturadas e as combina em uma grande base de conhecimento, onde:

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

Essa ação é realizada a partir de extratores, que são programas específicos, criados com essa finalidade, "para converter partes específicas de artigos da Wikipedia em sentenças RDF" (WEBER, 2015, p. 27). Afirma Weber (2015) 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 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).

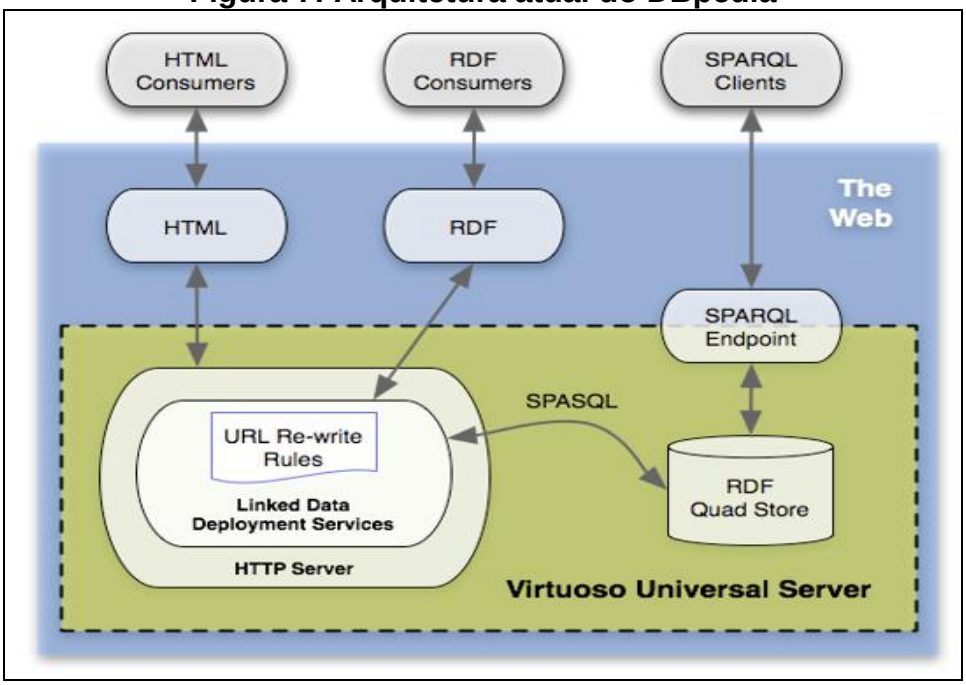
Com a ontologia é possível que outras aplicações façam consultas refinadas sobre o conteúdo disponibilizado pela Wikipédia, indo além da busca textual por palavra-chave (MOLON, 2013).

O projeto DBpedia é hospedado e publicado usando o *software OpenLink*

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

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. A Figura 7 apresenta a arquitetura do DBpedia e o uso do *OpenLink Virtuoso*.

Figura 7: Arquitetura atual do DBpedia



Fonte: (<Dbpedia.org>)

Nota-se que o Virtuoso é um “servidor universal” que não precisa possuir servidores dedicados para cada um dos domínios de funcionalidades, habilitando um único processo de servidor para vários protocolos diferentes.

Conforme informações obtidas no site do projeto do DBpedia (<Dbpedia.org>), com dados de 2017, atualmente a base de conhecimento Dbpedia, versão em inglês, tem mais de 4,58 milhões de coisas (*things*). Desse total de coisas, 1,83 milhões são classificados em uma Ontologia consistente (da qual podem ser derivados frameworks e modelos, disponível em <http://wiki.dbpedia.org/services-resources/ontology>), incluindo 1.445.000 pessoas, 735.000 localidades geográficas, 123.000 álbuns de música, 87.000 filmes, 19.000 jogos de vídeo, 241.000 organizações (incluindo 58.000 empresas e 49.000 instituições educacionais), 251.000 espécies de animais e 6.000 doenças. A DBpedia completa possui 38 milhões de rótulos e resumos em 125 idiomas diferentes, 25,2 milhões de *links* para imagens, 29,8 milhões de *links* para páginas

web externas, 80,9 milhões de *links* para categorias da Wikipédia. Esses dados compõem mais de 3 bilhões de itens de informações (conjuntos de dados RDF triplos), dos quais 550 milhões foram extraídos da versão inglesa da Wikipédia e o restante foi extraído de outras versões linguísticas.

Os dados da ontologia do DBpedia podem ser muito benéficos para usuários diversos, inclusive para bibliotecários, uma vez que incluem informações estruturadas da Wikipédia e abrangem uma variedade de domínios diferentes (MORSEY, 2012). Entre as possíveis vantagens, destacam-se: informações de contexto (lugares de nascimento, nacionalidade e influências de autores, por exemplo) para os tradicionais dados bibliográficos, adição de fatos sobre uma entidade específica visando aumentar a cobertura e a qualidade dos dados disponibilizados e fornecer um provedor de infraestrutura para armazenamento de dados lincados para diversas especialidades, aprimorando a sua qualidade, entre outros (MORSEY, 2012).

Contudo, conforme afirma Kontokostas et al. (2014 *apud* SOUZA; 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. Para os autores, os problemas encontrados no DBpedia podem ser classificados em três dimensões básicas: (1) timeliness ou atualidade, que estão relacionados com dados sobre o tempo e com a frequência da atualização desses dados, cuja falta pode provocar erro na representação de tais informações; (2) completude, que se refere ao conjunto de dados utilizados para representar um objeto ou fenômeno, de maneira que as informações sobre eles sejam completas o suficiente para descrever todos os seus atributos (de esquema, população e propriedade); (3) verificabilidade, que se refere à capacidade de se verificar a validade de uma informação, de forma fácil, para correção do dado.

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). O autor cita exemplos de algumas iniciativas colaborativas para a construção de bases de conhecimento para dados lincados, tais como o *OntoWiki* (AUER et al., 2007) e *Semantic MediaWiki* (KRÖTZSCH et al., 2006), bases essas

que podem ser utilizadas para coletar e integrar dados, tal como a aplicação do Catálogo de Professores (RIECHERT et al., 2010), onde historiadores criaram uma base de conhecimentos semânticos e prosopográficos (da carreira acadêmica) sobre professores que trabalharam na Universidade Leipzig durante seus 600 anos de história (MORSEY, 2012). Nesse tipo de colaboração, a possibilidade de ocorrer erros de representação é bem menor.

3 METODOLOGIA

Com o propósito de alcançar os objetivos de uma pesquisa, o método científico é formado por uma gama de procedimentos sistemáticos e lógicos, que traçam um caminho mais seguro e econômico, que permite detectar erros e facilitar as decisões sobre uma postura do investigador (MARCONI; LAKATOS, 2010, P. 65). De modo a seguir esse caminho, este Capítulo presenta as características da pesquisa e os procedimentos metodológicos que foram utilizados.

A pesquisa caracteriza-se como empírica, exploratória e aplicada, com uma abordagem qualitativa, tendo em vista aprofundar o conhecimento sobre o tema Linked Data, no que diz respeito à sua aplicação em dados bibliográficos de teses e dissertações disponibilizados em bibliotecas digitais.

Objeto da pesquisa são os links que serão criados para ligar conteúdo na web, sendo que os sujeitos para teste serão os resultados do trabalho de Maculan (2011), descritos na próxima seção e o ambiente da pesquisa é a BDTD da UFMG.

3.1 Materiais e insumos de pesquisa

Durante a pesquisa, para a atribuição de links para os conteúdos temáticos dos documentos, foram utilizados os resultados do trabalho de Maculan (2011), mantendo-se o mesmo escopo do material escolhido, teses e dissertações da linha OUI (Organização e Uso da Informação), defendidos no Programa de Pós-Graduação em Ciência da Informação (PPGCI) e disponíveis na Biblioteca Digital de Teses e Dissertações da UFMG (BDTD/UFMG).

Primeiramente, Maculan (2011) fez um recorte temporal no período entre 1998 e 2009, tendo sido recuperados 290 documentos, sendo 62 teses e 228 dissertações, referentes às linhas de pesquisa Gestão da Informação e do Conhecimento (GIC), Informação, Cultura e Sociedade (ICS) e Organização e Uso da Informação (OUI). Desse total, cerca de 50% estava disponível no banco de dados da BDTD/UFMG, ou seja, 146 documentos. Como Maculan (2011) trabalhou somente com documentos defendidos na linha de pesquisa OUI, foram computados 41 trabalhos, equivalendo a 66% do total de documentos disponíveis na BDTD/UFMG, que representaram o *corpus* da pesquisa da autora.

Maculan (2011) criou uma taxonomia facetada, denominada TAFNAVEGA,

para a representação do conteúdo de documentos do tipo teses e dissertações, visando facilitar a busca e a recuperação das informações disponibilizadas na BDTD. A TAFNAVEGA foi composta por um conjunto de dez classes básicas, a saber: Tema; Objeto empírico; Escopo; Ambientação; Coleta de dados; Tipo de pesquisa; Métodos; Fundamento teórico; Fundamento histórico/contextual; Resultados, conforme Tabela 3 abaixo. A partir de uma matriz categorial, contendo as dez classes básicas, a autora indexou os 41 documentos relativos à linha de pesquisa OUI. A estrutura facetada taxonômica agrupou os termos indexadores (extraídos dos documentos através do algoritmo) que compartilham características semelhantes, transformando-os em facetas e subfacetas navegáveis.

Tabela 3: Matriz categorial para trabalhos acadêmicos (teses e dissertações)

TERMOS CAFTE	QUESTIONAMENTOS (NORMA 12.676) e PRECIS	PARTE DA ESTRUTURA TEXTUAL
C1. TEMA	Qual o assunto de que trata o documento?	RESUMO / INTRODUÇÃO (problema, justificativa, objetivos)
C2. OBJETO EMPÍRICO	Qual o objeto empírico do estudo em questão? Qual objeto foi utilizado e/ou analisado na pesquisa?	RESUMO / INTRODUÇÃO (problema, justificativa, objetivos) / METODOLOGIA
C3. ESCOPO	O que pretende a pesquisa, de forma geral e específica, que seja relevante determinar? A que objetos a pesquisa tem intenção de atender (aprimorar, avaliar, analisar, identificar, contribuir, etc.)?	RESUMO / INTRODUÇÃO (problema, justificativa, objetivos)
C4. AMBIENTAÇÃO	O tema, objeto empírico e/ou ação são considerados no contexto de um lugar específico ou ambiente?	RESUMO / INTRODUÇÃO (problema, justificativa, objetivos) / METODOLOGIA
C5. TIPO DE PESQUISA	Tendo em vista a natureza (básica, aplicada), a abordagem (quantitativa, qualitativa), os objetivos (exploratória, descritiva, explicativa) ou os procedimentos (bibliográfica, experimental, documental, estudo de caso, pesquisa-ação, levantamento, *expost-facto*, participante), quais as classificações podem tipificar a pesquisa realizada?	RESUMO / METODOLOGIA / RESULTADOS / DISCUSSÃO DE RESULTADOS

TERMOS CAFTE	QUESTIONAMENTOS (NORMA 12.676) e PRECIS	PARTE DA ESTRUTURA TEXTUAL
C6. COLETA DE DADOS	Quais instrumentos específicos (questionários, entrevistas, registros áudios-visuais, coleta de documentos, etc.) foram utilizados para realizar a ação?	RESUMO / METODOLOGIA / RESULTADOS / DISCUSSÃO DE RESULTADOS
C7. MÉTODOS	Quais modos específicos foram utilizados para realizar a ação (por exemplo, técnicas ou métodos para tratamento dos dados, que podem ser do tipo: modelagem estatística, análise estrutural, codificação, análise de conteúdo, indexação, análise semiótica, retórica ou de discurso)?	RESUMO / METODOLOGIA / RESULTADOS / DISCUSSÃO DE RESULTADOS
C8. FUNDAMENTO TEÓRICO	Houve alguma corrente ou abordagem teórica específica (teorias, hipóteses, pressupostos, etc.) utilizada em função da natureza do objeto a ser pesquisado e dos objetivos pretendidos, que foram descritos na pesquisa?	RESUMO / REVISÃO DE LITERATURA / FUNDAMENTAÇÃO
C9. FUNDAMENTO HISTÓRICO/ CONTEXTUAL	Quais temas foram tratados e revisados, a partir de pesquisa bibliográfica, para contextualizar o tema pesquisado de forma profunda e consistente?	RESUMO / REVISÃO DE LITERATURA / FUNDAMENTAÇÃO
C10. RESULTADOS	Quais pontos a pesquisa alcançou, levando em consideração os objetivos propostos? Houve formulação ou reformulação de teoria, criação de um método ou de um produto?	RESUMO / RESULTADOS / DISCUSSÃO DE RESULTADOS / CONCLUSÕES

Fonte: Maculan (2011, p. 124) Adaptado do MLD de Fujita e Rubi (2006).

Segundo Maculan (2011, p. 124):

> O objetivo do modelo MCTCA [matriz] é orientar e sistematizar a análise conceitual nos documentos do tipo teses e dissertações, que, ao mesmo tempo, facilita a extração de conceitos e alimenta a estrutura da TAFNAVEGA, tornando-a um mecanismo para a navegação facetada.

Essa matriz pode ser considerada o algoritmo que permite a análise dos documentos.

Também foram utilizados os dados da DBPedia, cujos insumos estão descritos na seção 2.5.

4 DESCRIÇÃO DOS PROCEDIMENTOS E ANÁLISE DE RESULTADOS

Como descrito anteriormente, o insumo da pesquisa se baseia na taxonomia facetada elaborada por Maculan (2011), a qual indexou 41 documentos, entre teses e dissertações, da Biblioteca Digital de Teses e Dissertações (BDTD) da Universidade Federal de Minas Gerais (UFMG). Segundo a autora, a "representatividade da amostra foi garantida porque houve a análise de todos os documentos disponíveis na base de dados da BDTD da ECI-UFMG, dentro do recorte temporal e da limitação do objeto empírico" da pesquisa (MACULAN, 2011, p. 108). Para a autora, as partes da estrutura textual dos documentos serviram como parâmetro para a criação das "categorias finais", pois refletem as partes mais estáveis do documento do tipo tese e dissertação, definidas com a análise de literatura sobre esse tipo de documento.

A partir da análise dos 41 documentos "foi coletado um total de 407 termos indexadores. Esses termos foram refinados (...) Após esse refinamento, totalizaram 168 termos indexadores" (MACULAN, 2011, p. 126). Dentre esse conjunto de termos indexadores, foram selecionados os termos referentes a 10 documentos, de modo aleatório, para compor o *corpus* desta pesquisa.

A seguir, estão apresentados os procedimentos utilizados para formalizar o *Linked Data* para esses 10 documentos.

4.1 Identificação das Classes, termos e relacionamentos

Tendo em vista que para a construção de triplas RDF é necessário que alguns elementos possuam URI, como é o caso do "sujeito" da tripla, foi escolhido para tal propósito o elemento *Handle*[6], que aponta para cada tese ou dissertação na internet e fornece um Identificador único para esse tipo de obra.

A BDTD da UFMG já fornece o identificador *Handle*, bastando identificá-lo por meio de uma busca no catálogo online público. Essa busca foi realizada para cada um dos dez documentos, a partir do título e autor da tese ou dissertação, relacionados na Taxonomia Facetada de Maculan (2011).

O *handle* foi escolhido por ser uma URI (*Uniform Resource Identifier*) com acesso mundial por meio da internet, além de funcionar como um identificador

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

permanente do recurso. A Figura 8 mostra a localização do Handle na página da BDTD.

Figura 8: Localização do URI Handle no registro da página da BDTD

Fonte: https://goo.gl/r66G0N, 2016

Os documentos selecionados aparecem na Tabela 4 com seus respectivos indicadores (URI):

Tabela 4: Documentos selecionados e seus identificadores

	Título	URI
1	Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais	http://hdl.handle.net/1843/RRSA-6GGGUF
2	Análise de citações da produção científica de uma comunidade: a construção de uma ferramenta e sua aplicação em um acervo de teses e dissertações do PPGCI-UFMG	http://hdl.handle.net/1843/VALA-6KFNN9
3	Educação física no PPGMH/UFRGS: uma visão a partir da análise de citações e perfil dos pesquisadores.	http://hdl.handle.net/1843/VALA-6T7RH6
4	Metadados para descrição de documentos remanescentes de fundo eclesiástico: uma proposta de metadados e biblioteca digital para os assentos de batismo, casamento e óbito da Matriz do Pilar de Outro Preto	http://hdl.handle.net/1843/VALA-6T7R3Q
5	Organização e uso das bases de informação para o atendimento a clientes em call centers	http://hdl.handle.net/1843/VALA-6T7QQE
6	Prontuário eletrônico do paciente: estudo de uso pela equipe de saúde do Centro de Saúde Vista Alegre	http://hdl.handle.net/1843/VALA-6K5LVK
7	A análise de assunto na literatura ficcional infantil: categorias para lero o que você tem	http://hdl.handle.net/1843/VALA-6T7RN
8	MAPA HIPERTEXTUAL (MHTX): um modelo para organização hipertextual de documentos	http://hdl.handle.net/1843/LHLS-6BUPG9
9	Padrões de disciplinaridade no campo de pesquisa sobre a AIDS: uma prospecção a partir de publicações periódicas e pesquisadores	http://hdl.handle.net/1843/EARM-7HBQJU
10	As relações interdisciplinares refletidas na Ciência da Informação	http://hdl.handle.net/1843/ECID-7UUQ69

Fonte: elaborado pelo autor

Após a seleção dos identificadores dos documentos (URI/Handle), foram selecionadas as classes que identificavam os relacionamentos, representando o papel do “predicado” na tripla.

4.1.1 Identificação dos indicadores de classes

Para a identificação dos indicadores de classes foram utilizadas as classes que compõem a taxonomia de Maculan (2011) (tabela 2): a) tema; b) objeto; c) escopo; d) ambientação; e) métodos; f) tipo de pesquisa; g) coleta de dados; h) fundamento teórico; i) fundamento histórico/conceitual; j) causa e efeito; k) resultados. A partir deste conjunto de classes verificou-se se havia correspondência nos elementos do *Dublin Core Metadata Initiative* (DCMI), cujas características estão descritas na seção 2.2.1. Nessa verificação percebeu-se que nem todas as classes puderam ser equiparadas a uma URI do DCMI, que são elas: fundamento histórico/conceitual; causa e efeito; resultados, fundamentos teóricos; e tipo de pesquisa, não possuem um termo do *Dublin core* equivalente, e, portanto, não foram utilizadas.

Tabela 5:Classes da Taxonomia e URIs correspondentes

Classe	URI Correspondente
Ambientação	http://purl.org/dc/terms/spatial
Causa e efeito	Classe não disponível e equivalente não localizado.
Coleta de dados	
Escopo	http://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/PhysicalObject
Resultados	Classe não disponível e equivalente não localizado.
Tema	http://purl.org/dc/elements/1.1/subject
Tipo de pesquisa	Classe não disponível e equivalente não localizado.

Fonte:Elaborado pelo autor

As classes não localizadas foram ignoradas neste trabalho e podem vir a ser utilizadas para buscas em outros *Namespaces,* em trabalhos futuros.

Assim, o conjunto de classes, com correspondência com os elementos e qualificadores do DCMI é demonstrado na Tabela 6.

Tabela 6: Relação de classes da Taxonomia e URIs do DCMI

Classe	URI Correspondente
Tema	http://purl.org/dc/elements/1.1/subject
Metodologia	http://purl.org/dc/terms/instructionalMethod
Objeto	http://purl.org/dc/dcmitype/PhysicalObject
Escopo	http://purl.org/dc/terms/coverage
Ambientação	http://purl.org/dc/terms/spatial

Fonte: elaborado pelo autor, a partir dos resultados de Maculan (2011).

A escolha dos URIs do DCMI se deu por aproximação, considerando que esses não são específicos para descrever este tipo de documento (teses e dissertações). Conforme se nota na tabela, para a classe (1) Tema foi utilizado do elemento “subject”, por ser o elemento que descreve a temática do recurso representado. Na classe Metodologia, foi feita a escolha do Qualificador “InstructionalMethod”, por ser um qualificador que descreve métodos para gerar conhecimento. Na classe objeto foi utilizado o qualificador “PhysicalObject” por sua capacidade para descrever coisas da realidade. Para a classe escopo utilizou-se o elemento “coverage” por tal elemento descrever características temporais e espaciais do recurso. E, por fim, para a classe Ambientação usou-se o qualificador “Spatial” por descrever as características espaciais do recurso de qualquer natureza.

Após a identificação das URIs para cada classe, foi necessário identificar as URIs também para os termos que compõem cada uma destas classes, a partir da taxonomia de Maculan (2011).

4.1.2 Identificação de URI para os termos da taxonomia

Tido como base a indexação realizada por Maculan (2011), utilizou-se a sua matriz categorial, nos 10 documentos que foram utilizados como recorte nesta pesquisa, foi levantado o conjunto de termos que foram empregados na representação dos conteúdos das teses e dissertações, encontrado-se um total de

77 termos.

A exemplo do que foi feito para as classes, também para os termos foi necessário realizar a verificação de correspondência entre os 77 termos e as URIs disponíveis no DBPedia. A verificação foi feita da seguinte maneira:

1. Buscou-se o termo na Wikipédia. A exemplo do que foi com o termo “indexação”.
2. Uma vez encontrado um artigo para o termo “indexação” na Wikipédia, foi necessário alterar para o artigo em inglês; isso foi necessário porque apesar de algumas vezes existirem artigos com URIs para termos em português, em geral, os conteúdos da DBpedia são encontrados, em sua grande maioria, no idioma inglês; dessa forma, esse princípio foi empregado para todos os termos buscados e recuperados.
3. Dentro do artigo em inglês, foi copiada a última parte do seu endereço HTML, cujo conteúdo estava após a última barra; no caso do exemplo para o termo “indexação”, foi o trecho “Subject_indexing”.
4. Em seguida, foi preciso acessar o DBPedia, no endereço <http://pt.dbpedia.org/page/>, acrescido do trecho HTML que foi copiado anteriormente, ficando a busca da seguinte forma: <http://pt.dbpedia.org/page/Subject_indexing>; com isso, encontrou-se a URI para o termo buscado.
5. Nos casos nos quais não foi recuperado qualquer resultado, significou que não há uma URI para aquele termo específico; quando isso ocorreu, foi necessário reiniciar a busca por um termo equivalente.

Após essa verificação, constatou-se que 22 termos não possuíam URIs correspondentes, e que 55 termos obtiveram uma URI compatível ou semelhante. Entretanto, o total URIs consideradas foi de 51, uma vez que houve URIs apontadas para mais de um termo. Isso ocorreu para os seguintes casos: (a) uma única URI para os três termos - Biblioteca Digital, Biblioteca Eletrônica e Biblioteca Virtual; (b) uma única URI para os dois termos - Indexação e Análise de Assunto; (c) uma única URI para os dois termos - Sistemas de Recuperação da Informação e Recuperação da Informação. A Tabela 77 apresenta os termos recuperados e suas respectivas

URIs.

Tabela 7: Relação dos termos e URIs correspondentes

	Termo		URI Correspondente
1	Bibliometria	1	http://pt.dbpedia.org/page/Bibliometrics
2	Campo científico	2	http://pt.dbpedia.org/page/Branches_of_science
3	Atendimento em call centers	3	http://pt.dbpedia.org/page/Call_centre
4	Gráficos	4	http://pt.dbpedia.org/page/chart
5	Literatura infantil	5	http://pt.dbpedia.org/page/Children's_literature
6	Análise de citações	6	http://pt.dbpedia.org/page/Citation_analysis
7	Classificação	7	http://pt.dbpedia.org/page/Classification
8	Mapa conceitual	8	http://pt.dbpedia.org/page/Concept_map
9	Bases de dados	9	http://pt.dbpedia.org/page/Database
10	Biblioteca virtual	10	http://pt.dbpedia.org/page/Digital_library
11	Biblioteca digital		
12	Biblioteca eletrônica		
13	Dspace	11	http://pt.dbpedia.org/page/DSpace
14	Empirico	12	http://pt.dbpedia.org/page/Empiricism
15	Fundamentação epistemológica	13	http://pt.dbpedia.org/page/Epistemology
16	Análise facetada	14	http://pt.dbpedia.org/page/Faceted_classification
17	Instituição bancária	15	http://pt.dbpedia.org/page/Financial_institution
18	Questionário	16	http://pt.dbpedia.org/page/Form_(document)
19	Instituições de saúde	17	http://pt.dbpedia.org/page/Health_system
20	AIDS	18	http://pt.dbpedia.org/page/HIV
21	Hipermidia	19	http://pt.dbpedia.org/page/Hypermedia
22	Hipertexto	20	http://pt.dbpedia.org/page/Hypertext
23	Uso e impacto das novas tecnologias de comunicação e informação	21	http://pt.dbpedia.org/page/Information_and_communications_technology
24	Competência informacional	22	http://pt.dbpedia.org/page/Information_literacy
25	Sistema de recuperação da Informação	23	http://pt.dbpedia.org/page/Information_retrieval
26	Recuperação da Informação		
27	Tecnologias da informação	24	http://pt.dbpedia.org/page/Information_technology
28	Infometria	25	http://pt.dbpedia.org/page/Informetrics
29	Interdisciplinaridade	26	http://pt.dbpedia.org/page/Interdisciplinarity
30	Entrevista	27	http://pt.dbpedia.org/page/Interview
31	Knorr-Cetina	28	http://pt.dbpedia.org/page/Karin_Knorr_Cetina
32	Organização do conhecimento	29	http://pt.dbpedia.org/page/Knowledge_organization
33	Prontuário médico	30	http://pt.dbpedia.org/page/Medical_record
34	modelo metadados (novo)	31	http://pt.dbpedia.org/page/Metadata
35	Metodologia da pesquisa	32	http://pt.dbpedia.org/page/Methodology

36	Microfilme	33	http://pt.dbpedia.org/page/Microform
37	Sintagmas nominais	34	http://pt.dbpedia.org/page/Noun_phrase
38	Educação física	35	http://pt.dbpedia.org/page/Physical_education
39	Bordieu	36	http://pt.dbpedia.org/page/Pierre_Bourdieu
40	Divulgação científica	37	http://pt.dbpedia.org/page/Popular_science
41	Prefeituras	38	http://pt.dbpedia.org/page/Prefecture
42	Protocolo verbal	39	http://pt.dbpedia.org/page/Protocol_analysis
43	Base Qualis	40	http://pt.dbpedia.org/page/Qualis_(CAPES)
44	Informação científica e tecnológica	41	http://pt.dbpedia.org/page/Scientific_literature
45	Pesquisa científica	42	http://pt.dbpedia.org/page/Scientific_research
46	Cienciometria	43	http://pt.dbpedia.org/page/Scientometrics
47	Planilhas	44	http://pt.dbpedia.org/page/Spreadsheet
48	Indexação	45	http://pt.dbpedia.org/page/Subject_indexing
49	Análise de assunto		
50	Norma técnica	46	http://pt.dbpedia.org/page/Technical_standard
51	Telecomunicações	47	http://pt.dbpedia.org/page/Telecommunication
52	Kuhn	48	http://pt.dbpedia.org/page/Thomas_Kuhn
53	Universidades	49	http://pt.dbpedia.org/page/University
54	Estudo de usuário	50	http://pt.dbpedia.org/page/User_study
55	Modelagem conceitual	51	https://en.wikipedia.org/wiki/Conceptual_model

Fonte: Elaborado pelo Autor

As URIs recuperadas, que estão expostas na Tabela 7, corresponderam ao objeto da tripla.

4.1.3 Identificação de relacionamentos - Construção das triplas

Nas etapas anteriores, foram identificadas as URIs dos documentos (sujeito), das classes (predicado) e dos termos (objeto). Com esses dados foram criadas as triplas que representam os relacionamentos entre os três elementos.

Como exemplo, apresenta-se a tripla para uma das teses do recorte desta pesquisa, que foi formada da seguinte maneira:

Tabela 8: 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 pelo autor

No exemplo, é possível observar a identificação da URIs para Sujeito (documento), predicado (classe) e objeto (termo). Essa mesma representação da relação entre os três elementos, formada pelas triplas, foi realizada para cada um dos dez documentos do recorte dessa pesquisa.

Com o conjunto de triplas pronto, foi necessária a sua transcrição para um editor de textos que, no caso desta pesquisa, foi o Bloco de Notas do Windows. Para a descrição foi empregado o formato Ntriple, por ter sido considerado mais intuitivo de construir do que o formato RDF.

A tripla em formato Ntriple segue o seguinte padrão:

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

Como exemplo, cita-se a tese “Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais”, cujas triplas foram construídas em formato *Ntriple*:

```
        <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> .
```

Esse processo foi realizado para todos os dez documentos. Após transcrever todas as triplas, separadamente, foi utilizada uma ferramenta. disponível

gratuitamente, denominada EasyRdf – Converter[7], que converte as descrições *Ntriple* para formato RDF, conforme a seguir.

```
<?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>
```

Dessa forma, todas as triplas foram declaradas em RDF, representando as relações entre os documentos (sujeito) e suas propriedades (predicado e objeto) para os dez documentos.

4.2 Análise de resultados

Nesta subseção é apresentada a análise da amostra proposta nesta pesquisa, para avaliar a contribuição do Linked Data como recurso incorporado à Biblioteca Digital de Teses e Dissertações da Universidade Federal de Minas Gerais, visando agregar novos dados às informações disponibilizadas.

Documento 1: Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais.

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

A Estrutura original de classes e termos da Taxonomia[8] existentes no documento 1 está de acordo com tabela 9 abaixo:

Tabela 9: Estrutura original de classes e termos da Taxonomia – D1

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: Elaborada pelo autor.

Ao analisar este documento, percebe-se que sua estrutura requer somente algumas classes disponíveis nos elementos do Dublin Core, para quem acesse os conteúdos informacionais dessas classes à DBPedia. As seguintes classes são interligadas: (1) tema; (2) objeto; (3) escopo e (4) ambientação. Elas permitem apontar para os seguintes termos: (1) Indexação; (2) Sintagmas Nominais; (3) Sistemas de recuperação da Informação.

Essas relações são demonstradas na Tabela 10:

Tabela 10: Relação de Classes e termos do Documento 1

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

Fonte: Elaborada pelo autor.

A taxonomia original apresentava, na classe “Métodos”, o termo Análise Documentária, entretanto, esse termo não foi encontrado na base de dados DBPedia. Consequentemente, não foi possível construir a tripla que descrevesse esta relação.

[8] Veja tabela 11

Assim, se houvesse necessidade de realizar uma busca sobre a análise documentaria, como um tipo de método, conforme utilizada neste documento, não seria possível.

Nota-se que a utilização dos padrões da Web Semântica, RDF e URI, permitiu apontar para a DBPedia, através das triplas construídas com os termos: (1) Indexação; (2) Sintagmas Nominais; e (3) Sistemas de Recuperação da Informação contidas no documento para se obter mais informações.

Ao final dessa análise foi gerado o arquivo RDF abaixo:

```
<?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>
```

Documento 2: Análise de citações da produção científica de uma comunidade - a construção de uma ferramenta e sua aplicação em um acervo de teses e dissertações do PPGCI-UFMG

A estrutura original de classes e termos da Taxonomia existentes no documento 2 está de acordo com tabela abaixo:

Tabela 11: Estrutura original de classes e termos da Taxonomia – D2

Classes	Termos
Tema	Bibliometria
Tema	Cienciometria
Tema	Infometria
Objeto	Documentação Científica

Escopo	Estudos de Citação
Escopo	Ferramenta de análise de citações
Ambientação	Dspace
Tipo de Pesquisa	Experimental
Coleta de dados	Gráficos e Planilhas
Métodos	Análise de citações

Fonte: Elaborada pelo autor.

Diferentemente do documento anterior, nota-se que no Documento 2 houve uma ampliação na representação das classes propostas de acordo com a Taxonomia. Na estrutura deste documento foi possível preencher sete classes, sendo que a classe “Tema” aparece 3 vezes e a “Escopo” 2 vezes. As classes foram: (1) tema; (2) escopo; (3) Ambientação; (4) Métodos. Elas permitiram apontar para os seguintes termos: (1) Bibliometria; (2) Cienciometria; (3) Infometria; (4) Dspace; (5) gráficos e planilhas e (6) Análise de citações.

Estas relações são demonstradas na Tabela 12:

Tabela 12: Relação de Classes e termos do Documento 2

Classe	Termo
Tema	Bibliometria
Tema	Cienciometria
Tema	Infometria
Objeto	Documentação Científica
Escopo	Estudos de Citação
Escopo	Ferramenta de análise de citações
Ambientação	Dspace
Métodos	Análise de citações

Fonte: Elaborada pelo autor.

Neste Documento, não foi possível descrever as triplas para as classes (1) tipo de pesquisa; (2) Coleta de dados, pois os termos resultantes destas classes não tiveram campos e nem qualificador do Dublin Core, que possibilitasse a criação

desses apontamentos. Assim, nessa amostra, também não foi possível utilizar todas as classes da Taxonomia original para complementar os conteúdos informacionais passíveis de serem usados.

O arquivo RDF gerado a partir dessa análise é mostrado abaixo:

```
<?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/VALA-6KFNN9">
        <dc11:subject rdf:resource="http://pt.dbpedia.org/page/Bibliometrics"/>
   <dc11:subject rdf:resource="http://pt.dbpedia.org/page/Scientometrics"/>
        <dc11:subject rdf:resource="http://pt.dbpedia.org/page/Informetrics"/>
        <dc:coverage rdf:resource="http://pt.dbpedia.org/page/Scientometrics"/>
   <dc:coverage rdf:resource="http://pt.dbpedia.org/page/Physical_education"/>
   <dc:spatial rdf:resource="http://pt.dbpedia.org/page/DSpace"/>
   <dc:instructionalMethod
rdf:resource=”http://pt.dbpedia.org/page/Citation_analysis”/>
 </rdf:Description>
</rdf:RDF>
```

Documento 3: Educação Física no PPGMH/UFRGS: uma visão a partir da análise de citações e perfil dos pesquisadores

A estrutura original de classes e termos da Taxonomia[9] existentes no documento 3 está de acordo com a tabela abaixo:

Tabela 13: Estrutura original de classes e termos da Taxonomia – D3

Classe	Termo
Tema	Bibliometria
Tema	Cientometria
Tema	Infometria
Objeto	Documentação Científica

[9] Veja página 54

Escopo	Fundamentação epistemológica
Escopo	Estudos de citação
Escopo	Educação física
Ambientação	Universidades (privada e pública)
Métodos	Análises de citação
Métodos	Método indiciário
Fundamento teórico	Divulgação científica
Fundamento teórico	Kuhn
Fundamento teórico	Bordieu
Fundamento teórico	Knorr - Cetina
Fundamento histórico/conceitual	Bibliometria
Fundamento histórico/conceitual	Cienciometria
Fundamento histórico/conceitual	Infometria
Fundamento histórico/conceitual	Sociologia do conhecimento e da ciência
Causa e efeito	Estudos de autoria
Causa e efeito	Estudos de citação
Causa e efeito	Estudos sobre fontes de informação

Fonte: Elaborada pelo autor.

Após a analise desse documento, utilizando a Taxonomia, obteve-se como resultado 7 classes com ocorrências de 22 termos. Dessas 7 classes, somente as 4 classes “Tema, Escopo, Ambientação e Método” obtiveram termos correspondentes nos elementos dos Dublin Core.

Tabela 14: Relação de Classes e termos do Documento 3

Classe	Termo
Tema	Bibliometria
Tema	Cienciometria
Tema	Infometria
Escopo	Fundamentação Epistemológica

Escopo	Educação Física
Ambientação	Universidade

Fonte: Elaborada pelo autor.

Os termos da classe “Tema” foram todos localizados na DBpedia, por possuírem uma URI válida, contemplando toda a temática do documento. A classe Ambientação, que teve somente um termo ocorrente, também foi contemplada. Já as classes Escopo e Método, tiveram respectivamente 2 termos (Fundamentação epistemológica e Educação física) e 1 termo (Análise de citações) correspondente na DBpedia, possibilitando a criação da tripla com a URI compatível.

Ressalta-se que não foi possível encontrar na DBpedia, URIs para os termos correspondentes às classes, Objeto, Fundamento teórico, Fundamento histórico/conceitual e Causa e efeito, demonstrando a ineficiência da Base para cobrir toda a temática.

Após essa análise foi criado o arquivo RDF abaixo:

```
<?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/VALA-6T7RH6">
          <dc11:subject rdf:resource="http://pt.dbpedia.org/page/Bibliometrics"/>
   <dc11:subject rdf:resource="http://pt.dbpedia.org/page/Scientometrics"/>
          <dc11:subject rdf:resource="http://pt.dbpedia.org/page/Informetrics"/>
          <dc:spatial rdf:resource="http://pt.dbpedia.org/page/University"/>
          <dc:coverage rdf:resource=”http://pt.dbpedia.org/page/Physical_education”/>
          <dc:coverage rdf:resource=”http://pt.dbpedia.org/page/Epistemology”/>
          <dc:instructionalMethod
rdf:resource="http://pt.dbpedia.org/page/Citation_analysis"/>
  </rdf:Description>
 </rdf:RDF>
```

Documento 4: Metadados para descrição de documentos remanescentes de fundo eclesiástico: uma proposta de metadados e biblioteca digital para os assentos de batismo, casamento e óbito da matriz do Pilar de Ouro Preto.

A estrutura original de classes e termos da Taxonomia existentes no documento 4 é a seguinte:

Tabela 15: Estrutura original de classes e termos da Taxonomia – D4

Classe	***Termo***
Tema	Organização do conhecimento
Objeto	Microfilme
Escopo	Outros sistemas e tecnologias de comunicação e informação
Escopo	Modelo de metadados
Ambientação	Bibliotecas (virtual, digital e eletrônica)
Tipo de pesquisa	Experimental
Coleta de dados	Gráficos e planilhas
Métodos	Estatística, mensuração
Métodos	Análise documentária
Fundamento teórico	Metadados Padrões MARC
Fundamento teórico	Metadados Padrões Dublin Core
Fundamentação histórico/conceitual	Usos da informação e de unidades de informação
Fundamentação histórico/conceitual	Publicações oficiais
Fundamentação histórico/conceitual	Bibliotecas (virtual, digital e eletrônica) [10]
Fundamentação histórico/conceitual	Recuperação da informação
Resultados	Modelo metadados (novo)

Fonte: Elaborada pelo autor.

A representação deste Documento foi realizada através de 10 classes e 16 termos. As classes “Escopo” Métodos, Fundamento teórico e Fundamentação histórico/conceitual tiveram mais de um termo na sua representatividade, já as outras classes, apenas uma ocorrência. Ressalta-se que a classe Fundamentação histórico/conceitual teve a ocorrência de 4 termos, e as classes Fundamentação

[10] Apesar de a literatura bibliotecônoma diferenciar os termos Biblioteca Digital, Biblioteca Virtual e Biblioteca eletrônica, a DBpedia considera tais termos como correlatos, gerando apenas uma URI para os três termos.

teórica, Métodos e Escopo, 2 termos cada; nas outras classes a representatividade foi de somente um termo.

Tabela 16: Relação de Classes e termos do Documento 4

Classe	Termo
Tema	Organização do Conhecimento
Objeto	Microfilme
Escopo	Metadados
Ambientação	Bibliotecas (virtual, digital e eletrônica)

Fonte: Elaborada pelo autor.

Das classes atribuídas para o documento, somente quatro - Tema, Objeto, Escopo e Ambientação - tiveram elementos correspondentes no Dublin Core. Entre essas classes, a classe “Escopo” teve 2 termos, porém somente um teve a URI correspondente.

Até o momento, foi o documento que menos obteve apontamentos para a DBpedia, se comparar a quantidade de classes versus termos.

Após essa análise foi criado o arquivo RDF abaixo:

```
<?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/VALA-6T7R3Q">
         <dc11:subject rdf:resource="http://pt.dbpedia.org/page/Knowledge_organization"/>
         <ns0:PhysicalObject rdf:resource="http://pt.dbpedia.org/page/Microform"/>
         <dc:coverage rdf:resource="http://pt.dbpedia.org/page/Metadata"/>
   <dc:spatial rdf:resource="http://pt.dbpedia.org/page/Digital_library"/>
 </rdf:Description>

</rdf:RDF>
```

Documento 5: Organização e uso das bases de informação para o atendimento a clientes em call centers.

A estrutura original de classes e termos da Taxonomia existentes no documento 5 é a seguinte:

Tabela 17: Estrutura original de classes e termos da Taxonomia – D5

Classe	Termo
Tema	Estudos de usuário, demanda e uso da informação e de unidades de informação
Objeto	Scripts de atendimento
Escopo	Avaliação de bases de dados
Escopo	Atendimento em Call Centers
Escopo	Usos da informação e de unidades de informação
Ambientação	Instituição bancária
Ambientação	Instituições de saúde
Ambientação	Telecomunicações
Tipo de pesquisa	Empírica
Coleta de dados	Questionários
Coleta de dados	Entrevista semi-estruturada
Métodos	Avaliação de serviços e de unidades de informação
Fundamento teórico	Teorias e conceitos de informação
Fundamento histórico/conceitual	Atendimento em call centers
Fundamento histórico/conceitual	uso da informação e de unidades de informação
Resultados	Norma técnica (vigente e proposta)

Fonte: Elaborada pelo autor.

A representação deste Documento foi realizada através de 10 classes e 16 termos. As classes “Escopo”, Ambientação, Coleta de dados, e Fundamentação histórico/conceitual tiveram mais de um termo na sua representatividade, já as outras classes, apenas uma ocorrência. Nota-se que as classes Ambientação e Escopo tiveram a ocorrência de 3 termos cada, e as classes Ambientação, Coleta de

Dados e Fundamentação histórico/conceitual tiveram ocorrência de 2 termos cada; e nas outras classes a representatividade foi de somente um termo.

Tabela 18: Relação de Classes e termos do Documento 5

Classe	Termo
Tema	Estudos de usuário, demanda e uso da informação e de unidades de informação
Escopo	Atendimento em Call Centers
Ambientação	Instituição bancária
Ambientação	Instituições de saúde
Ambientação	Telecomunicações

Fonte: Elaborada pelo autor.

Apesar da alta ocorrência de termos, foi possível criar triplas somente para 5 termos, por causa, tanto da disponibilidade de elementos representativos do Dublin Core, quanto da existência das URIs na base de dados DBPedia.

Portanto, as classes representadas foram somente: (1) tema; (2) escopo; (3) Ambientação. Foram criadas as triplas apontando para a DBPedia para os termos: (1) Estudo de usuário, (2) Atendimento em call centers, (3) Instituição bancária, (4) Instituições de saúde, (5) Telecomunicações.

Nota-se que, apesar de o conteúdo do Documento possibilitar uma boa representatividade de acordo com a estrutura original de classes e termos da Taxonomia, a ampliação de seu conteúdo informacional utilizando a tecnologia de Linked Data não atingiu todo o potencial.

Após essa análise, o arquivo RDF gerado foi este abaixo:

```
<?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/VALA-6T7QQE">
        <dc11:subject rdf:resource="http://pt.dbpedia.org/page/User_study"/>
        <dc:coverage rdf:resource="http://pt.dbpedia.org/page/Call_centre"/>
        <dc:spatial rdf:resource="http://pt.dbpedia.org/page/Telecommunication"/>
```

```
  <dc:spatial rdf:resource="http://pt.dbpedia.org/page/Health_system"/>
  <dc:spatial rdf:resource="http://pt.dbpedia.org/page/Financial_institution"/>
 </rdf:Description>
</rdf:RDF>
```

Documento 6: Prontuário Eletrônico do Paciente: estudo de uso pela equipe de saúde do Centro de Saúde Vista Alegre

A estrutura de classes e termos originais no documento 6:

Tabela 19: Estrutura original de classes e termos da Taxonomia – D6

Classes	***Termos***
Tema	Tecnologias da informação
Objeto	Prontuário médico
Escopo	Caracterização e comportamento do usuário
Escopo	Uso e impactos das novas tecnologias de comunicação e informação
Escopo	Usos da informação e de unidades de informação
Escopo	Competência informacional
Ambientação	Prefeituras
Ambientação	Instituições de saúde
Coleta de dados	Entrevistas

Fonte: Elaborada pelo autor.

O conteúdo do documento permitiu sua representação através de somente 5 classes da Taxonomia. A classe Escopo ocorreu 4 vezes, Ambientação 2 vezes, e as classes Tema, Objeto e Coleta de dados uma vez. Com isso, foi possível, ao final, obter a representatividade de 9 termos: (1) Tecnologias da informação, (2) Prontuário medico, (3) Caracterização e comportamento do usuário, (4) uso e impactos das novas tecnologias de comunicação e informação (5) Usos da informação e de unidades de informação, (6) Competência Informacional, (7) Prefeituras, (8) Instituições de saúde e (9) Entrevistas.

Porém, dentre essas classes, somente a Classe Coleta de dados não possuiu correspondência a nenhum elemento do Dublin Core

Após análise sua nova estrutura é mostrada abaixo.

Tabela 20: Relação de Classes e termos do Documento 6

Classe	Termo
Tema	Tecnologias da informação
Objeto	Prontuário Médico
Escopo	Competência Informacional
Ambientação	Prefeituras

Fonte: Elaborada pelo autor.

Em relação à representatividade utilizando as triplas, dos 9 termos resultantes da análise das classes taxonômicas, somente 4 termos, descritos na tabela acima, encontraram correspondência com os elementos do Dublin Core, também foi possível criar as triplas para todos, com apontamento para as URIs da BDPédia.

```
<?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/VALA-6K5LVK">
          <dc11:subject rdf:resource="http://pt.dbpedia.org/page/Information_technology"/>
          <ns0:PhysicalObject rdf:resource="http://pt.dbpedia.org/page/Medical_record"/>
          <dc:coverage
rdf:resource="http://pt.dbpedia.org/page/Information_and_communications_technology"/>
          <dc:coverage rdf:resource="http://pt.dbpedia.org/page/Information_literacy"/>
    <dc:spatial rdf:resource="http://pt.dbpedia.org/page/Prefecture"/>
  </rdf:Description>
 </rdf:RDF>
```

Documento 7: Mapa Hipertextual (MHTX): um modelo para organização hipertextual de documentos

A estrutura original de classes e termos da Taxonomia existentes no documento 7 é a seguinte:

Tabela 21: Estrutura original de classes e termos da Taxonomia – D7

Classes	***Termos***
Tema	Organização do conhecimento
Objeto	Hipertexto e hipermídia
Escopo	Outros sistemas e tecnologias de comunicação e informação
Ambientação	Bibliotecas (virtual, digital e eletrônica)
Métodos	Análise documentária
Métodos	Análise facetada
Métodos	Modelagem conceitual
Fundamento teórico	Mapa conceitual
Fundamento teórico	Análise facetada
Resultado	Mapa hipertextual MHTX

Fonte: Elaborada pelo autor.

Após a análise deste documento, utilizando a Taxonomia, obteve-se como resultado 7 classes com ocorrências de 10 termos. Dessas 7 classes, somente as 4 classes “Tema, Objeto, Escopo, Ambientação e Método”, obtiveram termos correspondentes nos elementos do Dublin Core, possibilitando assim, a criação das triplas para 6 URIs na DBpédia.

Tabela 22: Relação de Classes e termos do Documento 7

Classe	Termo
Tema	Organização do conhecimento
Objeto Objeto	Hipertexto Hipermidia
Ambientação	Bibliotecas (virtual, digital e eletrônica)
Métodos	Análise facetada
Métodos	Modelagem conceitual

Fonte: Elaborada pelo autor.

A classe Método teve 3 termos para serem apontados na DBpedia, mas somente modelagem conceitual e análise facetada teve URIs correspondentes para agregar valor informacional. Já a classe Escopo não obteve nenhuma URI correspondente ao termo “Outros sistemas e tecnologias de comunicação e informação”

Ressalta-se que o termo “Análise facetada” foi representado em duas classes, tanto na classe Método, quanto na classe Fundamento teórico.

Após essa análise foi criado o arquivo RDF abaixo:

```
<?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/LHLS-6BUPG9">
          <dc11:subject rdf:resource="http://pt.dbpedia.org/page/Knowledge_organization"/>
          <ns0:PhysicalObject rdf:resource="http://pt.dbpedia.org/page/Hypertext"/>
          <ns0:PhysicalObject rdf:resource="http://pt.dbpedia.org/page/Hypermedia"/>
          <dc:coverage>Outros sistema e tecnologias de comunicação e
informação</dc:coverage>
          <dc:spatial rdf:resource="http://pt.dbpedia.org/page/Digital_library"/>
          <dc:instructionalMethod>Análise Documentária</dc:instructionalMethod>
          <dc:instructionalMethod
rdf:resource="http://pt.dbpedia.org/page/Faceted_classification"/>
    <dc:instructionalMethod rdf:resource="https://en.wikipedia.org/wiki/Conceptual_model"/>
  </rdf:Description>
</rdf:RDF>
```

Documento 8: A análise de assunto na literatura ficcional infantil: categorias para ler o que você tem

A estrutura original de classes e termos da Taxonomia existentes no documento 8 é a seguinte:

Tabela 23: Estrutura original de classes e termos da Taxonomia – D8

Classes	***Termos***
Tema	Análise de assunto
Objeto (qualificador)	Literatura infantil
Escopo	Análise documentária
Escopo	Classificação
Escopo	Indexação (automática e manual)
Ambientação	Bibliotecas (virtual, digital e eletrônica)

Métodos	Protocolo verbal
Fundamento teórico	Metodologia da pesquisa
Fundamento teórico	Análise de citações

Fonte: Elaborada pelo autor.

No Documento 8 foram representadas 6 classes resultando em 9 termos representativos de acordo com a estrutura taxonômica. Nota-se que a classe Escopo é contemplada com 3 termos, por sua vez, Fundamentação teórica possui 2 termos. As outras classes obtiveram uma única ocorrência:

Tabela 24: Relação de Classes e termos do Documento 8

Classe	Termo
Tema	Análise de assunto
Objeto	Literatura Infantil
Escopo	Classificação
Escopo	Análise de assunto
Metodo	Protocolo Verbal

Fonte: Elaborada pelo autor.

Assim, foi possível criar triplas com os 5 termos para as URIs da DBpedia. De todos esses termos, somente “Análise documentária” (Escopo) não obteve resultado que permitisse a interligação dos conteúdos informacionais na DBpedia.

Após essa análise foi criado o arquivo RDF abaixo:

```
<?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/VALA-6T7RN">
          <dc11:subject rdf:resource="http://pt.dbpedia.org/page/Subject_indexing"/>
          <ns0:PhysicalObject
rdf:resource="http://pt.dbpedia.org/page/Children's_literature"/>
          <dc:coverage>Análise Documentária</dc:coverage>
          <dc:coverage rdf:resource="http://pt.dbpedia.org/page/Classification"/>
    <dc:coverage rdf:resource="http://pt.dbpedia.org/page/Subject_indexing"/>
    <dc:instructionalMethod rdf:resource="http://pt.dbpedia.org/page/Protocol_analysis"/>
  </rdf:Description>
</rdf:RDF>
```

Documento 9: Padrões de disciplinaridade no campo de pesquisa sobre a AIDS - uma prospecção a partir de publicações periódicas e pesquisadores.

A estrutura original de classes e termos da Taxonomia existentes no documento 9 é a seguinte:

Tabela 25: Estrutura original de classes e termos da Taxonomia – D9

Classes	***Termos***
Tema	Estudos de usuário, demanda e uso da informação e de unidades de informação
Objeto	Informação científica e tecnológica
Objeto	AIDS
Escopo	Fundamentação epistemológica
Escopo	Pesquisa científica
Ambientação	Bases de dados
Métodos	Análise documentária
Métodos	Classificação
Fundamento teórico	Divulgação científica
Fundamento teórico	Estudos sobre fontes de informação
Fundamento histórico/conceitual	Campo científico

Fonte: Elaborada pelo autor.

Ao analisar, este Documento obteve 7 classes representativas com 11 termos distribuídos entre elas. Percebeu-se que somente algumas classes que estavam disponíveis nos elementos do Dublin Core, para que acessassem os conteúdos informacionais dessas classes à DBPedia. As classes “Objeto”, “Escopo” e “Fundamentação Teórica” ocorrem 2 vezes cada.

Tabela 26: Relação de Classes e termos do Documento 9

Classe	Termo
Tema	Estudos de usuário, demanda e uso da informação e de unidades de informação
Objeto	Informação científica e tecnológica
Objeto	AIDS
Escopo	Fundamentação Epistemológica

Escopo	Pesquisa Científica
Ambientação	Base de dados
Metodo	Classificação

Fonte: Elaborada pelo autor.

Das 7 classes originais do documento, 5 possuem correspondência com os elementos do Dublin Core - Tema, Objeto, Escopo, Ambientação e Método. Foi possível, então, criar triplas para interligar as informações de 7 termos pertencentes a essas classes, com a DBpedia. Pode-se dizer que neste documento houve uma boa representatividade completar ao conteúdo do documento em questão.

Apenas para o termo “Análise documentária”, contido na classe Método, não foi encontrada URI compatível.

Após essa análise foi criado o arquivo RDF abaixo:

```
<?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/EARM-7HBQJU">
          <dc11:subject rdf:resource="http://pt.dbpedia.org/page/User_study"/>
          <ns0:PhysicalObject rdf:resource="http://pt.dbpedia.org/page/Scientific_literature"/>
          <ns0:PhysicalObject rdf:resource="http://pt.dbpedia.org/page/HIV"/>
          <dc:coverage rdf:resource="http://pt.dbpedia.org/page/Epistemology"/>
          <dc:coverage rdf:resource="http://pt.dbpedia.org/page/Scientific_research"/>
          <dc:spatial rdf:resource="http://pt.dbpedia.org/page/Database"/>
          <dc:instructionalMethod>Análise Documentária</dc:instructionalMethod>
          <dc:instructionalMethod rdf:resource="http://pt.dbpedia.org/page/Classification"/>
  </rdf:Description>

</rdf:RDF>
```

Documento 10: As relações interdisciplinares refletidas na ciência da informação

Tabela 27: Estrutura original de classes e termos da Taxonomia – D10

Classe	Termo
Tema	Interdisciplinaridade
Objeto	Avaliação de Periódicos

Escopo	Estudos da Produção e da produtividade científica
Ambientação	Base Qualis
Fundamento Histórico Conceitual	Interdisciplinaridade

Fonte: Elaborada pelo autor.

O termo Interdisciplinaridade foi encontrado na base de dados DBPedia, entretanto, a classe Fundamento Histórico Conceitual não possui um elemento do *Dublin Core* correspondente, tornando impossível explicitar a relação entre eles.

Tabela 28: Relação de Classes e termos do Documento 10

Classe	Termo
Tema	Interdisciplinaridade

Fonte: Elaborada pelo autor.

Os termos: avaliação de periódicos, estudos da produção e da produtividade científica e base qualis não possuem correspondente na base de dados DBPedia, o que impossibilitou a criação das triplas para a interligação dos dados.

Portanto, somente a temática interdisciplinaridade pode ter um apontamento para o uso do *Dublin Core*, para a DBPedia.

```
<?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>
```

Nesta análise, percebeu-se que a aplicação das 10 classes da Taxonomia, em relação aos 15 elementos e qualificadores do Dublin Core e às URIs da DBPedia, nem sempre pode ser realizada, por não ter correspondência entre os resultados desses três procedimentos. Um dos principais problemas que pode ser

levantado é que a Base de Dados - DBPedia, não é uma base especifica da área temática em questão. Isso dificulta, porém, que todas as URIs sejam criadas.

Notou-se também que alguns documentos, em sua análise, preencheram quase todas as classes da Taxonomia, algumas classes tiveram mais de um termo representativo, mas nem sempre esses termos podem ser considerados expressivos, ou seja, pode ser que a pesquisa descrita naquele documento tratou de várias temáticas, que não eram essencialmente relevantes para o conteúdo da pesquisa.

Outra observação, em números, que pode confirmar essa suposição, se baseia no resultado total de termos dos documentos, avaliados a partir da Taxonomia. Foram obtido 119 termos em potencial, sendo que somente em 48 termos deste total, foi possível criar as triplas com apontamento para a DBPedia, ou seja, menos de 50 % do total.

Assim, pode-se dizer que houve problemas de representatividade no uso da Taxonomia, com os elementos do Dublin Core e a disponibilidade destes termos na DBPedia, para que o uso do Linked Data fosse utilizado como mecanismo para agregar informações à BDTD.

5 CONSIDERAÇÕES FINAIS

Esta pesquisa foi iniciada tendo como objetivo verificar a contribuição do *Linked Data* como recurso incorporado às Bibliotecas Digitais de Teses e Dissertações (BDTDs), para possibilitar a agregação de novos dados às informações já disponibilizadas nos seus repositórios digitais, visando atender às necessidades de informação dos usuários desse tipo de biblioteca digital. A literatura estudada demonstra que o primordial usuário das BDTDs é a própria comunidade científica, que, ao procurar por documentos do tipo tese e dissertação, esses usuários esperam encontrar informações detalhadas sobre conteúdos tais como os temas de interesse, métodos e técnicas de pesquisa e as bases teóricas que fundamentam as investigações.

A partir do estudo das características e das iniciativas referentes ao *Linked Data*, verificou-se que é um esforço comunitário para fornecer um conjunto de dados (informações) vinculados na web, com licenças abertas, que provê princípios para a publicação desses dados de forma estruturada, usando URI (identificadores de dados) e RDF (pequeno fragmento disponível na web para de referenciar uma URI). Dessa maneira, visa definir ligações de dados entre os dados (coisas) de uma determinada fonte de dados para os dados dentro de outras fontes de dados. Assim, seria possível ligar os dados de uma BDTD, que é uma fonte de dados, aos dados do DBpedia, por exemplo, que é outra fonte de dados.

Nesta perspectiva, foi possível verificar que o *Linked Data* pode ser utilizado para agregar informações às BDTDs, uma vez que segue os princípios da Web Semântica, que é mais que simplesmente depositar dados na web. Trata-se de fazer *links* entre dados, ou seja, vincular dados para que uma pessoa ou máquina possa explorar a Web de Dados e recuperar ou encontrar outros dados relacionados. Dessa maneira, cria-se uma teia de dados que não é simplesmente um hipertexto, onde os links são relações âncoras em documentos de hipertexto escritos em HTML e que constroem ligações entre dados (coisas) de forma arbitrária, mas são dados identificados por URIs que apontam para qualquer tipo de objeto ou conceito, que, em geral, está em formato de licença aberta, disponibilizado na *web*. Esse é o cerne da Web Semântica: integração de dados vinculados, em grande escala e em vários níveis de complexidade, acessibilidade e possibilidade de raciocínio sobre os dados disponibilizados na Web.

Nesta pesquisa foi utilizado o conjunto de dados vinculados do DBpedia, que é um projeto que torna o conteúdo da Wikipedia disponível em RDF. Nesse projeto não há apenas dados da Wikipédia, mas, também, incorpora dados advindos de outros conjuntos de dados na web como, por exemplo, para Geonames. Pelo fato de integrar dados de vários conjuntos de dados, o DBpedia pode fornecer ao usuário de uma BDTD uma gama maior de conteúdos para atender à sua necessidade de maiores informações sobre o teor contido nos documentos do tipo tese e dissertação.

É preciso destacar que as iniciativas do Linked Data ainda apresentam algumas limitações para a integração de dados, tal como foi verificado na literatura sobre os dados contidos na Wikipédia, que são utilizados no projeto do DBpedia, uma vez que há um uso desordenado de infoboxes, muitas vezes sem um rigor de padronização de representações. Por outro lado, as inconsistências encontradas podem gerar propostas de resoluções dos problemas na Wikipédia, podendo levar a um aprimoramento da qualidade dos dados disponibilizados. Dessa forma, de acordo com os feedbacks encontrados na literatura, por exemplo, faz-se necessário, a geração de hierarquias de classe mais consistentes, a eliminação de propriedades múltiplas com o mesmo significado e a construção de extratores de infobox mais inteligentes, que possam identificar os diferentes infoboxes para uma mesma classe. A literatura também demonstrou que o fato de a DBpedia possuir ligações com outras fontes de dados torna possível desenvolver uma extensão de MediaWiki, trazendo informação adicional aos artigos da Wikipédia, a exemplo dos dados Geonames e de outros dados tais como imagens do Flickr e dados estatísticos do Eurostat e do CIA Factbook.

De modo geral, para se atingir a plenitude do uso da tecnologia Linked Data para recuperação de dados em um sistema, é preciso a construção de novas Triple Stores para suprir a demanda de URI’s em uso pelos diversos sistemas. Apesar das aplicações nesta pesquisa terem sido realizados com URIs já disponíveis no DBpedia, seria melhor se houvesse um maior número de datasets e triple stores disponíveis.

Embora sejam percebidos os problemas já elencados acima, devido à riqueza e diversidade de conhecimento estruturado que pode ser disponibilizado, se as BDTDs fossem capazes de promover a gestão do conhecimento da Web de Dados, por meio do Linked Data, tornando-se, assim, centros de excelência e ampliando

suas fronteiras, os usuários poderiam se beneficiar com essa troca e compartilhamento de conhecimento.

No campo da Ciência da Informação, a contribuição desta pesquisa se evidencia porque trouxe o tema do Linked Data para discussão, pois considera-se que seus princípios são importantes na integração de dados na web. Ao disponibilizar dados vinculados na web, a partir de triplas em RDF, aplicativos de diferentes sistemas e repositórios digitais podem explorar e reutilizar o conhecimento contido nas fontes de dados que utilizam os princípios do Linked Data, e, assim, são baseados na integração de dados.

Como estudos futuros aponta para necessidade de estudos sobre a aplicação do linked data em outras bases de dados bibliográficos que utilizam outros padrões que possam permitir uma maior representatividade, utilizando bases de dados mais especificas da área, e também, estudar essa mesma metodologia em outra tipologia de documentos.

6 REFERÊNCIAS

ADOBE Systems. **Document Management**: portable document format part 1 PDF 1.7. 2008. Disponível em: http://goo.gl/HLltkL0000000000

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.

ALVES, R. C. V. **Web Semântica**:uma análise focada no uso de metadados. 180 f. 2005. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual de Paulista, Marília, 2005) Disponível em: <http://repositorio.unesp.br/handle/11449/93690> Acesso em:07/11/2015

AUER, S. et al. OntoWiki: **A tool for social, semantic collaboration**. In: NOY, N. F. et al. 16TH INTERNATIONAL WORLD WIDE WEB CONFERENCE (WWW2007), 16., Banff, Canada, May 8, 2007. Proceedings… of the Workshop on Social and Collaborative Construction of Structured Knowledge (CKC 2007). Banff, Canada, CEUR Workshop Proceedings, 2007. v. 273.

AUER, Sören et al. Dbpedia: A nucleus for a web of open data. In: **The semantic web**. Springer Berlin Heidelberg, 2007. p. 722-735.

AZEVEDO, PATRÍCIA CAROLINA NEVES DE. Uma proposta para visualização de linked data sobre enchentes na Bacia do Rio Doce. **Projetos e Dissertações em Sistemas de Informação e Gestão do Conhecimento**, v. 3, n. 1, 2014.

BAKER, Thomas. Dublin Core in Multiple Languages: Esperanto, Interlingua, or Pidgin. In: **Proceedings of the International Symposium on Research, Development and Practice in Digital Libraries**. 1997. p. 8-15.

BARRIENTOS, EDER ÁVILA. **Linked Open Data en la Biblioteca Digital Semántica Académica**. Disponível em: http://migre.me/rA1DN Acesso em: 11 de Setembro de 2015

BERMES, E. Convergence and Interoperability: a Linked Data perspective. In: **WORLD LIBRARY AND INFORMATION CONGRESS:** 77th IFLA GENERAL CONFERENCE AND ASSEMBLY, 77., 13 a 18 de agosto, San Juan. Anais... San Juan: IFLA, 2011. Disponível em: <http://www.ifla.org/past-wlic/2011/149-bermes-en.pdf>. Acesso em: 12 fev. 2017.

Berners Lee, T., Hendler, j., Lassila, O. The semantic web. **Scientific american**, v. 284, n. 5, p. 28-37, 2001.

BERNERS-LEE, T. **Sir Tim Berners-Lee talks with Talis about the Semantic Web**. [Online]. 2008. Disponível em: <talis-podcasts.s3.amazonaws.com/twt20080207_TimBL.html>. Acesso em: 9 abr. 2017.

BERNERS-LEE, Tim. **Linked Data**. 2006. Disponível em: <https://goo.gl/jK6GwE> Acesso em 22 dez. 2016.

BERNERS-LEE, Tim. **Uniform Resource Identifiers (URI)**: Generic Syntax. 1998. Disponível em: < http://www.ietf.org/rfc/rfc2396.txt > Acesso em 25 jan. 2017.

BIZER, C. et al. **How to publish linked data on the web**. 2007.

BIZER, C. The emerging web of linked data. **IEEE Intelligent Systems**, v. 24, n. 5, p. 87–92, 2009.

BIZER, C.; CYGANIAK, R.; GAUß, T. **The RDF Book Mashup**: From Web APIs to a Web of Data. Proceedings of the 3rd Workshop on Scripting for the Semantic Web (SFSW2007). 2007. Disponível em: <http://richard.cyganiak.de/2008/papers/bookmashup-sfsw2007.pdf>

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked Data: The Story so far. **International Journal on Semantic Web and Information Systems**, v. 5, n. 3, p. 1-22, 2009. Disponível em: <http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf>

BLATTMANN, Ursula; WEBER, Claudiane. Dspace como repositório digital na Organização. **Revista ACB Biblioteconomia** v. 13, n. 2, p. 467-485, 2008. Disponível em: < https://revista.acbsc.org.br/racb/article/view/593>. Acesso em: 05/04/2017.

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

CARDOSO, Olinda Nogueira Paes. Recuperação de Informação. **INFOCOMP Journal of Computer Science**, v. 2, n. 1, p. 33-38, 2004. Disponível em: <https://goo.gl/CYHoYI> Acesso em: 11/02/2017.

CASTELLS, Manuel. **A Galáxia Internet**: reflexões sobre a Internet, negócios e a sociedade. Zahar, 2003.

CASTRO, F. F.; Santos, P. L. V. A. C. Os metadados como instrumentos tecnológicos na padronização e potencialização dos recursos informacionais no âmbito das bibliotecas digitais na era da web semântica. Informação e Sociedade: Estudos, João Pessoa, v.17, n.2, p.13-19, maio/ago. 2007. Disponível em: < http://www.ies.ufpb.br/ojs2/index.php/ies/article/view/840> . Acesso em: 30/01/2017.

CATARINO, Maria Elisabete; CERVANTES, Brígida Maria Nogueira; DE ALMEIDA, Ilza Andrade. A representação temática no contexto da web semântica. **Informação & Sociedade**, v. 25, n. 3, 2015.

CATARINO, M. E.; SOUZA, T. B. A representação descritiva no contexto da web semantica. **Transinformação**, Campinas, v. 24, n. 2, p. 77-90, ago. 2012. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-37862012000200001&lng=pt&nrm=iso&gt>. Acesso em: 13 abr. 2017.

CHOI, Y. ; RASMUSSEN, E. What is need to educate future digital librarians. **D-lib magazine**, v. 12, n.9, p. 1-8, set. 2006. Disponível em: <http://www.dlib.org/dlib/september06/choi/09choi.html> Acesso em 01 ago. 2016.

CUNHA, Danusa R.B.; LÓSCIO, Bernadette F.; SOUZA, Damires. **Linked Data**: Da Web de Documentos para a Web de Dados. III Escola Regional de Computação-Ceará, Maranhão e Piauí-ERCEMAPI, 2011.

CUNHA, Murilo Bastos da. Biblioteca digital: bibliografia internacional anotada. **Ci. Inf. [online**]. 1997, vol.26, n.2 ISSN 1518-8353. Disponível em <http://dx.doi.org/10.1590/S0100-19651997000200013>. Acesso em 11 de Out. 2015.

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

CUNHA, Murilo Bastos da. Desafios na construção de uma biblioteca digital. **Ci. Inf,** v. 28, n. 3, p. 257-268, 1999.

DIAS, Eduardo Wense. Contexto digital e tratamento da informação. **DataGramaZero-Revista de Ciência da Informação**, v. 2, n. 5, 2001.

DUBLIN CORE METADATA INITIATIVE. **Dublin Core metadata element set, version 1.1**: reference description. [S.l.], 2004. Disponível em: <http://dublincore.org/documents/dces/>. Acesso em: 27 jul. 2015.

FELIPE, Eduardo Ribeiro, **A importância dos metadados em bibliotecas digitais**: da organização à recuperação da informação. 2012. 110 f. Dissertação (Mestrado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2012.

FICHE, R. H. et al. A utilização dos repositórios digitais encontrados na organizações. **RACIn**, João Pessoa, v. 5, n. 1, p. 61-80, jan.- jun. 2017. Disponível em http://racin.arquivologiauepb.com.br/edicoes/v5_n1/racin_v5_n1_artigo04.pdf. Acesso em 16/12/2017.

FUJITA, Mariângela Spotti Lopes; LACRUZ, Maríadel Carmem Augustín; DÍAZ, Raquel Gómez. A situação atual da indexação nas tarefas bibliotecárias**. Perspect. Ciênc. Inf**, v. 17, n. 1, p. 94-109, 2012.

FURRIE, B. **Understanding MARC bibliographic**: machine-readable cataloging. 7th. ed.

GIL, Antônio Carlos. **Métodos e Técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2008.

GLOBO.COM. **Astronautas recebem acesso à internet na Estação Espacial**. G1 Tecnologia, Rio de Janeiro, 22 jan. 2010. Disponível em: <https://goo.gl/jEDmW2> Acesso em 21 jan. 2017.

HEATH, Tom; BIZER, Christian. **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.

HU, B.; SVENSSON, G**.** A Case Study of Linked Enterprise Data. In: PATEL-SCHNEIDER, P.;PAN, Y., et al (Ed.). **The Semantic Web – ISWC 2010**: Springer Berlin Heidelberg, v.6497, 2010. cap. 9, p.129-144. (Lecture Notes in Computer Science). ISBN 978-3-642-17748-4.

Instituto Brasileiro de Informação em Ciência e Tecnologia, IBICT. **Sobre a BDTD**. Disponível em <http://goo.gl/mMmI4W> Acesso em 23 de Novembro de 2015

KONTOKOSTAS, Dimitris et al. Test-driven evaluation of linked data quality. In: **23RD INTERNATIONAL CONFERENCE ON WORLD WIDE WEB**. Proceedings… [S.l.], ACM, 2014. p. 747-758.

KRÖTZSCH, M., VRANDECI'C, D.; VÖLKEL, M. Semantic MediaWiki. In:**THE SEMANTIC WEB, Lecture Notes in Computer Science, Springer, Heidelberg, DE. Proceedings**… [S.l.], ISWC, 2006. v. 4273, p. 935-942.

LANCASTER, F. W. **Indexação e Resumos:** teoria e prática. 2 ed. Brasília, DF: Briquet de Lemos, 2004.

LANCASTER, F. W. **Toward paperless information systems**. New York: Academic Press, 1978.

LARA, Marilda Lopes Ginez de. Documentary languages and knowledge organization systems in the context of the semantic web. **Transinformação**, Campinas , v. 25, n. 2, p. 145-150, Ago. 2013 . Disponível em <http://goo.gl/Dwy0d9>. Acesso em 10 de abril de 2016.

LASLIE, M. **The People's Encyclopedia**. Science, v. 301, p. 1299, Sep. 2003.

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.

LAUFER, C. **Guia de Web Semântica**. São Paulo: Secretaria do Governo de SP; Brasília: Embaixada Britânica, 2015.

LEVACOV, Marília. Bibliotecas Digitais: (r)evolução?. **Ciência da Informação**, Brasília, v. 26, n. 2, Maio 1997. Disponível em: <http://ref.scielo.org/6qd48z>. Acesso em: 27 jul. 2015

MACULAN, **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.

MARCONDES, C. H. Metadados: descrição e recuperação na Web In: MARCONDES, C. H. et al. (Org.). **Bibliotecas digitais**: saberes e práticas. Salvador, BA : EDUFBA; Brasília; IBICT, 2005. p. 77-143

MARCONDES, Carlos Henrique; Mendonça, Marília A.; Carvalho, Suzana M. Serviços via Web em bibliotecas universitárias brasileiras. **Perspectivas em Ciência da Informação**, Belo Horizonte v. 11, n. 2, p. 174-186. 2006.

MARCONDES, Carlos Henrique; SAYÃO, Luis Fernando. Documentos digitais e novas formas de cooperação entre sistemas de informação em C&T. **Ciência da Informação**, Brasília, v. 31, n. 3, p. 42-54, 2002.

MARCONDES, Carlos Henrique; SAYÃO, Luís Fernando. Integração e interoperabilidade no acesso a recursos informacionais eletrônicos em C&T: a proposta da Biblioteca Digital Brasileira. **Ciência da Informação**, v. 30, n. 3, p. 24-33, 2001.

MENDONÇA, Marília A.; CARVALHO, Suzana M.; MARCONDES, Carlos Henrique. **Serviços via Web em bibliotecas universitárias brasileiras**. 2006.

MOLON, J. T. **Armazenando registros de colaboração utilizando triple store**. 2013. 79f. TCC (Bacharelado em Ciência de Computação). Centro de Computação e Tecnologia da Informação, Universidade de Caxias do Sul. Caxias do Sul, dezembro de 2013.

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.

OLIVEIRA, Marlene de. **Ciência da informação e biblioteconomia**: novos conteúdos e espaços de atuação. Editora UFMG, 2005.

Washington, D. C.: **Library of Congress**; Follet Software, 2003. Disponível em: <http://www.loc.gov/marc/umb/>. Acesso em: 27 jul. 2015.

PIZZOL, Leandro Dal. **Uso da web de dados como fonte de informação no processo de inteligência competitiva setorial**. 2014. Tese de Doutorado. Universidade Federal de Santa Catarina.

RIECHERT, T. et al. Knowledge engineering for historians on the example of the catalogus professorum lipsiensis. In: PATEL-SCHNEIDER, P. F. et al. (Ed**). 9TH INTERNATIONAL SEMANTIC WEB CONFERENCE, 9., Lecture Notes in Computer Science, Springer, Shanghai/China**. Proceedings… of the ISWC2010, Shanghai/China, 2010. v. 6497 of, p. 225–240.

ROSETTO, Marcia; NOGUEIRA, Adriana Hypólito. **Aplicação de elementos metadados Dublin Core para descrição de dados bibliográficos on-line da biblioteca digital de teses da USP**. Seminário Nacional de Bibliotecas Universitárias, v. 12, 2002.

RUSSO, Mariza. **Fundamentos em Biblioteconomia e Ciência da informação**. Editora E-Papers Serviços Editoriais, 2010.

SANTAREM SEGUNDO; SEMÂNTICA, JE Web. introdução a recuperação de dados usando Sparql. **ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO**, v. 15, p. 3863-3882, 2014.

SAYÃO, L. F.; MARCONDES, C. H. O desafio da interoperabilidade e as novas perspectivas para as bibliotecas digitais**. Transinformação**, v. 20, n. 2, p. 133–148, 2008.

SEQUEDA, Juan. **Introduction to: Tiplestores**. 2003. Disponível em: <http://www.dataversity.net/introduction-to-triplestores/> Acesso em 27 de nov. 2016.

SILVA, Renata Eleuterio da et al. **As tecnologias da Web Semântica no domínio bibliográfico**. 2013. Disponível em: <http://base.repositorio.unesp.br/handle/11449/93653> . Acesso em: 27 jul. 2015.

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

SOUZA, T. B. De; CATARINO, M. E.; SANTOS, P. C. Dos. Metadados: catalogando dados na internet. **Transinformação**. v. 9, n. 2, p. 93-105, Maio/agosto, 1997. Disponível em: <http://periodicos.puc-campinas.edu.br/seer/index.php/transinfo/article/download/1586/1558> Acesso em 29 set. 2015.

Universidade Federal de Minas Gerais, UFMG. **Biblioteca Digital de Teses e Dissertações da UFMG.** Disponível em: <https://goo.gl/RVAkic> Acesso em: 23 nov. 2015.

VANDER SANDE, Miel et al. **Everything is connected: Using Linked Data for multimedia Narration os Connections between Concepts**. In: 11th Internation Semantic Web Conference (ISWC-2012). 2012 Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.416.6162&rep=rep1&type=pdf

VERGUEIRO, Waldomiro; DE CARVALHO, Telma. Definição de indicadores de qualidade: a visão dos administradores e clientes de bibliotecas universitárias. **Perspectivas em Ciência da Informação**, v. 6, n. 1, 2001.

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