Artigo Original

Encontros Bibli

ENRIQUECIMENTO SEMÂNTICO PARA A DISPONIBILIZAÇÃO DE DADOS ABERTOS: TEORIA E PRÁTICA

Semantic enrichment for open data availability: theory and practice

Emanuelle TORINO
Doutoranda em Ciência da Informação. Programa de Pós-Graduação em Ciência da Informação, Universidade Estadual Paulista - Unesp, Marília, Brasil
Universidade Tecnológica Federal do Paraná (UTFPR), Curitiba, Brasil
emanuelle@utfpr.edu.br
https://orcid.org/0000-0002-3791-9884

Gustavo Lunardelli TREVISAN
Doutorando em Ciência da Informação. Programa de Pós-Graduação em Ciência da Informação, Universidade Estadual Paulista - Unesp, Marília, Brasil
Universidade Estadual Paulista - Unesp, Programa de Pós-Graduação em Ciência da Informação, Marília, Brasil
g.trevisan@unesp.br
https://orcid.org/0000-0002-4175-7910

Caio Saraiva CONEGLIAN
Doutorando em Ciência da Informação. Programa de Pós-Graduação em Ciência da Informação, Universidade Estadual Paulista - Unesp, Marília, Brasil
Docente do Centro Universitário Eurípedes de Marília (UNIVEM). Marília, Brasil.
caio.coneglian@gmail.com
https://orcid.org/0000-0002-6126-9113

Leonardo Castro BOTEGA
Doutor em Ciência da Computação
Docente do Centro Universitário Eurípedes de Marília (UNIVEM). Marília, Brasil.
Docente do Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista – Unesp, Marília, Brasil.
leonardo.botega@unesp.br
https://orcid.org/0000-0003-1495-5935

José Eduardo Santarem SEGUNDO
Doutor em Ciência da Informação
Docente. Universidade de São Paulo (USP). Departamento de Educação, Informação e Comunicação, Ribeirão Preto, Brasil.
Docente do Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista – Unesp, Marília, Brasil.
santarem@usp.br
https://orcid.org/0000-0003-3360-7872

Silvana Aparecida Borsetti Gregorio VIDOTTI
Doutora em Ciência da Informação
Docente. Universidade Estadual Paulista - Unesp, Departamento de Ciência da Informação. Programa de Pós-Graduação em Ciência da Informação, Marília, Brasil
silvana.vidotti@unesp.br
https://orcid.org/0000-0002-4216-0374

A lista completa com informações dos autores está no final do artigo

RESUMO

Objetivo: Apresentar o processo de formalização necessário à disponibilização de dados abertos no contexto do *linked open data*, utilizando-se, neste caso, dados que estão no contexto da pesquisa científica.
Método: Para a elaboração do estudo, foram selecionados três *datasets*, dois disponíveis no portal Dados Abertos CAPES e o terceiro, consiste dos dados do *Open Researcher and Contributor* ID (ORCID). Como procedimentos metodológicos foram utilizados a pesquisa bibliográfica para embasamento teórico-conceitual do estudo e a pesquisa descritiva para explorar e fornecer informações quanto ao processo de enriquecimento de dados abertos.
Resultado: Apresenta como resultado a descrição do processo necessário à disponibilização de um conjunto de dados enriquecido, a partir da modelagem e estruturação, pronto para a conversão ferramental e recuperação.
Conclusões: A partir desse estudo, é possível visualizar como ocorre o processo de enriquecimento semântico de dados no contexto do *linked open data*, abarcando a seleção, análise, processamento e preparação visando a disponibilização, o uso e o reuso.
PALAVRAS-CHAVE: Dados Abertos. Ligação de Dados. Enriquecimento semântico.

ABSTRACT

Objective: Present the formalization process necessary to make open data available in the context of the linked open data, using, in this case, data that are in the context of scientific research.
Methods: For the elaboration of the study, tree datasets were selected, two available on the Dados Abertos CAPES portal and the third consists of data from the Open Researcher and Contributor ID (ORCID). As methodological procedures were used the bibliographic research for theoretical-conceptual basis of the study and the descriptive research to explore and provide information about the process of enrichment of open data.

Results: It presents as a result the description of the process required to make available a rich data set, from the modeling and structuring, ready for tool conversion and recovery.
Conclusions: From this study, it is possible to visualize how the process of semantic data enrichment occurs in the context of linked open data, encompassing the selection, analysis, processing and preparation for availability, use and reuse.
KEYWORDS: Open Data. Linked Data. Semantic enrichment.

1 INTRODUÇÃO

Na atualidade, a quantidade de dados gerados, seu contínuo e ininterrupto crescimento e a facilidade de disponibilização desses na web, contribuem para um crescimento quantitativo exponencial e, geralmente, desordenado. Concomitantemente, soluções que se apropriam de conceitos recentes, como os da web semântica, termo cunhado por Tim Bernes-Lee, surgem com o propósito de estruturar as páginas da web, de forma que os dados tenham significados passíveis de interpretação por humanos e máquinas (BERNERS-LEE; HENDLER; LASSILA, 2001).

A tendência internacional de disponibilização de dados, visando à transparência, sobretudo governamental, culminou no chamado movimento de dados abertos (MURRAY-RUST, 2008; SAYÃO; SALES, 2014) que, posteriormente atingiu outras esferas, como a dos dados de pesquisa e de produção científica.

É inegável que nesse cenário, apresentam-se também novos olhares acerca das formas de produção, disponibilização, disseminação e uso de dados abertos, ao passo que o tratamento e a recuperação visando ao reuso ainda demandam estudos, na perspectiva da Ciência da Informação.

Assim, faz-se necessário não apenas a observação dos conceitos aplicados à produção de dados abertos, limitando-se à sua teorização, mas também da sua *práxis* no que tange ao tratamento adequado, que possibilite recuperação para reuso. Tal discussão amplia em interesse e complexidade quando a ela é acrescido o contexto de *linked open data*.

Neste contexto, a web semântica e o *linked data* fornecem solução a essa problemática, uma vez que estão vinculados a uma série de tecnologias e padrões que possibilitam a publicação de dados, de forma que sejam expressivos computacionalmente para recuperação e interoperabilidade, favorecendo seu reuso.

Portanto a utilização de tecnologias da web semântica e princípios do *linked data* podem colaborar na inserção desses dados no âmbito dos dados abertos e com relações entre si. Além disso, ao seguir as principais diretrizes das Boas Práticas de Publicação de Dados Abertos, preconizadas pelo *World Wide Web Consortium* (W3C) (LÓSCIO; BURLE; CALEGARI, 2017), os dados são disponibilizados com um nível de semântica

formal elevado, permitindo que tais informações sejam recuperadas futuramente com mais eficiência, possibilitando a realização de inferências por aplicações computacionais.

Vale destacar que o processo de disponibilização de dados exige a utilização de diversas técnicas e tecnologias que visam a torná-los mais expressivos. O uso das Boas Práticas de Publicação de Dados Abertos pode auxiliar nesta tarefa, além da definição de ontologias que estão relacionadas ao contexto dos dados e a sua ligação com outras bases relacionadas, permitindo que as informações estejam mais bem contextualizadas.

No entanto, o processo de publicação de dados exige reflexões sobre como emana a relação desses dados, desde o momento de sua geração, publicação até o uso e reuso, de forma a refletir o contexto em que está inserido. Assim, visando contribuir com a discussão, este estudo objetiva apresentar o processo de formalização necessário à disponibilização de dados abertos no contexto do *linked open data*, utilizando, neste caso, dados que estão no contexto da pesquisa científica.

Para tanto, foi realizada a pesquisa bibliográfica para embasamento teórico-conceitual do estudo e a pesquisa descritiva e aplicada para explorar e fornecer informações quanto ao processo de enriquecimento e conexão de dados abertos.

As etapas necessárias ao enriquecimento dos dados são explicitadas de forma a apoiar o leitor na práxis de modelagem, utilizando-as para o enriquecimento semântico, ontologias e vocabulário controlado, observando-se ainda, as recomendações de boas práticas para a produção e disponibilização de dados abertos no contexto do *linked open data* (LÓSCIO; BURLE; CALEGARI, 2017).

O resultado deste tratamento traduz-se no processo necessário ao enriquecimento e ao relacionamento de conjuntos de dados, adequadamente preparados para a etapa de disponibilização, visando a recuperação, uso e reuso por humanos e aplicações computacionais.

2 DADOS ABERTOS E *LINKED OPEN DATA*

O acesso à informação é uma das temáticas amplamente discutidas em diferentes cenários, nos âmbitos nacional e internacional. Tal fato está intimamente ligado à disponibilização de dados na web, o que ocorre, na maioria das vezes, de forma desordenada e desestruturada, tornando o ambiente digital um espaço de difícil tratamento e recuperação.

No Brasil, a partir da Lei de Acesso à Informação (LAI), as instituições foram obrigadas a disponibilizar informações e a publicar anualmente dados administrativos em sites da web. Nos termos do art. 4º da LAI, considera-se “I - informação: dados, processados ou não, que podem ser utilizados para produção e transmissão de conhecimento, contidos em qualquer meio, suporte ou formato” (BRASIL, 2011).

É pertinente esclarecer que a simples disponibilização de dados na web não os torna dados abertos. A abertura de dados está atrelada à adoção de preceitos básicos necessários à sua preparação para a disponibilização, visando, além da recuperação, ao uso e ao reuso. Neste contexto, dados abertos são aqueles disponibilizados sem barreiras de acesso e em formatos abertos e legíveis por humanos e aplicações computacionais, com a utilização de vocabulário de metadados padronizado internacionalmente e licença de direitos autorais pouco restritiva que estabelecem as formas autorizadas de uso.

Dados abertos são aqueles cujas utilização, reutilização e redistribuição podem ser realizadas mediante atribuição de créditos e adoção de licenças (OPEN KNOWLEDGE INTERNATIONAL, 2018). É destacado ainda que os dados devem ser: disponibilizados integralmente, preferencialmente baixados pela internet, de forma que possam ser modificados; passíveis de uso, reuso, redistribuição e conexão com outros dados; universal, sem distinção pessoal, computacional ou legal.

Open Knowledge International (2018) esclarece ainda que a importância da disponibilização de dados abertos seguindo as características supracitadas está na interoperabilidade.

Partindo dessa premissa, podemos expandir a relevância ao tratarmos dados abertos na perspectiva da web de dados que “[...] sugere a ligação entre dados publicados em bases de dados disponíveis livremente na Internet.” (SANTAREM SEGUNDO, 2014, p. 3865). Para tanto, apoia-se na web semântica para a estruturação dos conjuntos de dados, utilizando vocabulário de metadados padronizado, ontologias e inferências, o que culmina no conceito de *linked data*, proposto por Tim Berners-Lee com o objetivo de ligar conjuntos de dados estruturados disponíveis na web.

Holland e Verborgh (2014) esclarecem que o termo *linked data* refere-se ao conjunto de boas práticas para a estruturação e disponibilização de dados na web. Tal afirmação é corroborada no levantamento bibliográfico apresentado por Arakaki (2016), no qual 34 autores apresentam definições e a maioria define *linked data* como “melhores práticas para estruturação de dados na Web”. O trabalho mais citado por esses autores é intitulado "*Linked data: the story so far*", de autoria de Bizer, Heath e Berners-Lee (2009).

Quanto ao conceito e definição do termo, em Arakaki (2016) foi observada uma diferença primordial entre os termos *linked data* e *linked open data*. Para o autor, *linked data* caracteriza-se como melhores práticas para estruturar e ligar dados, de forma a facilitar a busca por agentes humanos e computacionais e direcioná-los a diferentes bases a partir desses dados ligados. Enquanto linked open data refere-se aos dados abertos no contexto do *linked data*.

Acerca do *linked data* e do *linked open data*, Berners-Lee (2006, tradução nossa) definiu quatro princípios para a estruturação de dados:

a) use URIs como nomes para as coisas;

b) use HTTP URIs para que as pessoas possam procurar esses nomes;

c) quando alguém procurar um URI, forneça informações úteis, usando os padrões (RDF, SPARQL);

d) inclua *links* para outros URIs, para que outros possam descobrir mais coisas.

Adicionalmente, as boas práticas de publicação de dados (LÓSCIO; BURLE; CALEGARI, 2017) são diretrizes que auxiliam na elaboração de subsídios para os que pretendem disponibilizar dados na web, com um padrão de qualidade que atenda aos requisitos iniciais de Bernes-Lee. Esse documento guia a publicação de dados, contemplando os principais elementos que devem ser considerados para que os dados estejam adequados aos princípios do *linked data*.

Assim, as boas práticas estão estruturadas em um conjunto de 35 recomendações, agrupadas em 13 categorias - metadados, licença, proveniência, qualidade, versionamento, identificação, formatos, vocabulários, acesso, preservação, feedback, enriquecimento e republicação - que norteiam a organização de dados de forma coerente para a obtenção de melhor resultado, cujos benefícios estão atrelados a: descoberta, acesso, confiabilidade, compreensão, processabilidade, reuso, interoperabilidade e conexão.

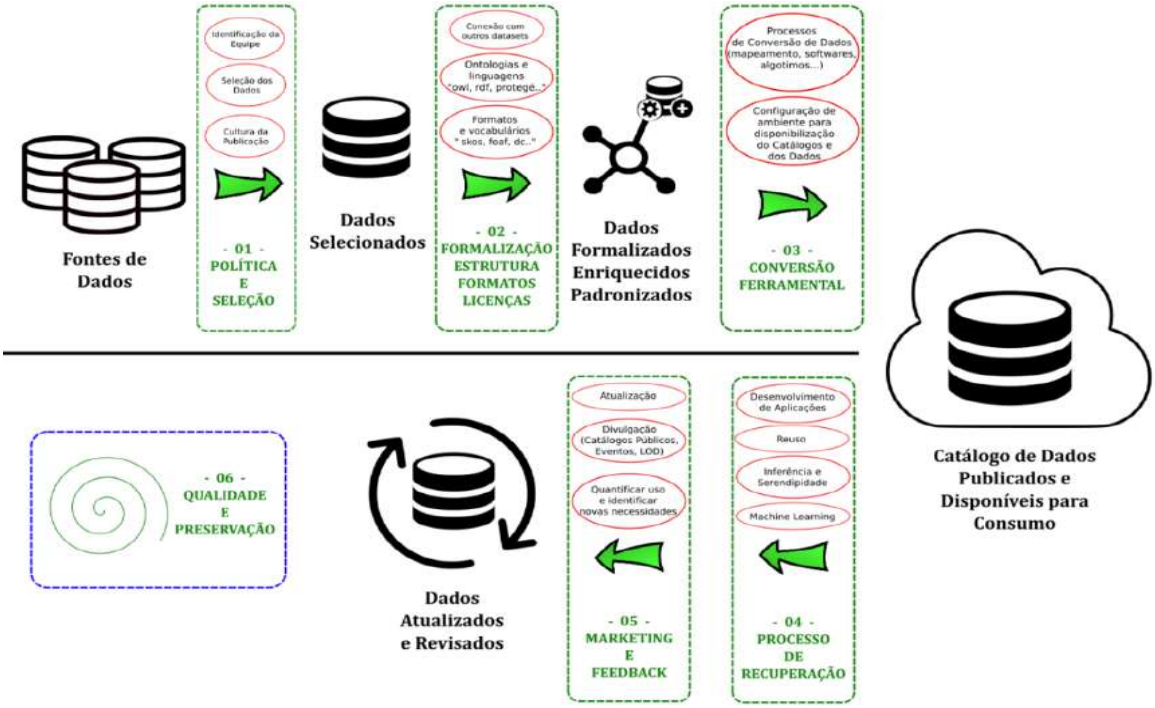
A seguir apresenta-se como este trabalho considera as boas práticas de publicação, adotando-as de forma a atender ao processo de formalização para estruturação e enriquecimento dos conjuntos de dados.

3 PUBLICAÇÃO DE DADOS SEGUINDO OS PRINCÍPIOS DO *LINKED DATA*

A publicação de dados contempla diversas etapas que devem ser consideradas, para que a disponibilização e a recuperação futura sejam adequadas. Neste estudo, considera-se que a publicação extrapola a simples conversão de dados para formato estruturado, utilizando *Resource Description Framework* (RDF), e contempla aspectos como as políticas, a formalização e o próprio aprimoramento da recuperação.

Assim, conforme Santarem Segundo (2018), o processo de publicação de dados no contexto do *linked data* está dividido nas seguintes etapas: política de seleção dos dados, formalização (estrutura, formatos e licenças), conversão ferramental, processo de recuperação, marketing e feedback (Figura 1).

Figura 1 – Processo de publicação de dados no contexto do *linked data*



Fonte: Santarem Segundo (2018).

Tal estrutura, permeada pelos conceitos de qualidade e preservação, estabelece que os dados constantemente estão sendo requisitados, seja na etapa de seleção ou na de recuperação, de forma que o planejamento da sua disponibilização precisa ser adequadamente estruturado, o que requer atenção na etapa de seleção, formalização, padronização e enriquecimento (parte das etapas do processo de publicação), fase preliminar à disponibilização efetiva dos dados em um ambiente digital, sobre o qual serão realizadas consultas e recuperação para uso e reuso.

Para fins deste estudo, cujo objetivo é apresentar o processo de formalização necessário à disponibilização de dados abertos no contexto do *linked open data*, cumpre

ressaltar que focamos na primeira etapa do processo de publicação de dados, que consiste na seleção, formalização, padronização e enriquecimento de dados. Para tanto, foram observadas as boas práticas do W3C referentes a esta etapa do processo.

Os dados devem ser disponibilizados utilizando um padrão de metadados compreensível por humanos e aplicações computacionais, adequado ao seu domínio e em quantidade suficiente para representar seu conteúdo e contexto, condições estas relevantes para a descoberta, uso, reuso, compreensão e processabilidade dos dados. Além disso, devem ser acompanhados de uma licença, disponível também em seus metadados, de forma a apresentar a humanos e aplicações computacionais as condições de uso. Para que possam ser trabalhados no contexto do *linked open data*, os dados devem estar disponíveis em um formato não-proprietário e legível por máquina.

Desta forma, selecionados os conjuntos de dados, é possível proceder ao enriquecimento semântico, que se constitui em um conjunto de processos utilizados para aprimorá-los ou melhorá-los, possibilitando conectá-los ou gerar dados novos.

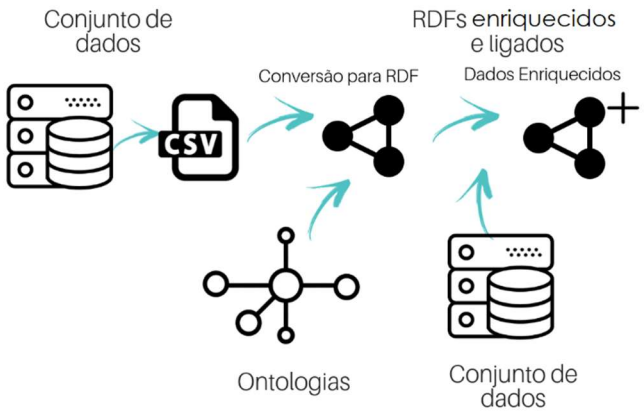
3.1 Enriquecimento Semântico de Dados

Disponibilizar dados abertos, sobretudo no contexto do *linked open data,* requer planejamento e enriquecimento, considerando os preceitos da web semântica, visando à compreensão por humanos e por agentes computacionais. Para tanto, à medida que a web semântica se desenvolve, são incorporados conceitos, técnicas e tecnologias que objetivam o tratamento semântico dos dados disponibilizados na web. Santarem Segundo, Coneglian e Lucas (2017) destacam algumas dessas tecnologias: *Resource Description Framework* (RDF), *eXtensible Markup Language* (XML), SPARQL *Protocol and RDF Query Language* (SPARQL), *Ontology Web Language* (OWL), dentre outras.

Os autores afirmam que o ferramental supracitado é amplamente conhecido, apesar de pouco utilizado em ambientes digitais, permanecendo como desafio o desenvolvimento da web de dados, que objetiva interligação dos dados.

A Figura 2 apresenta o enriquecimento semântico de dados iniciado pela conversão de um arquivo *Comma-separated values* (CSV) em RDF, ao qual é aplicada uma ontologia para posterior ligação a outro conjunto de dados, gerando assim um RDF enriquecido e pronto para a disponibilização em um ambiente digital de dados.

Figura 2 – Enriquecimento Semântico de Dados



Fonte: Autoria própria (2019).

Na Figura 2 destaca-se a aplicação de ontologias visando tornar os dados disponíveis em RDF enriquecidos pela formalização de propriedades e de axiomas que tornam o nível de semântica formal mais elevado. Complementarmente, o processo apresentado contempla a ligação dos dados com outras bases de dados, que é um dos princípios do *linked data*, em que os dados disponibilizados devem ter ligações com outros conjuntos, como, por exemplo, realizar uma ligação com o Open Researcher and Contributor ID (ORCID).

Destaca-se ainda que esse processo permite aos usuários a navegação pelos dados, aumentando a capacidade de inferências e de gerar mais valor a eles. Esse processo enriquece os dados, uma vez que possibilita a inserção de outros vocabulários e ontologias que fornecem informações contextuais daqueles conjuntos e relações. Um exemplo seria possibilitar os nomes de citação que um autor pode possuir, por meio de relações como *owl:sameAs*, permitindo, assim, que o ORCID forneça as diferentes citações que um determinado autor pode ter em suas publicações.

4 RESULTADOS E DISCUSSÕES

Visando apresentar a aplicação prática do processo de seleção, formalização, padronização, enriquecimento e ligação de dados, foram selecionados conjuntos de dados da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). Os dados estão disponíveis no portal Dados Abertos CAPES, na temática Avaliação da Pós-Graduação *Stricto Sensu*.

O primeiro conjunto de dados selecionado - Autor da Produção Intelectual de Programas de Pós-Graduação *Stricto Sensu* no Brasil 2013 a 2016, tipo produção

bibliográfica e subtipo artigo em periódico[1] - disponibiliza dados sobre os autores com produção no período, tais como: identificação do tipo e subtipo de produção, tipo de vínculo com o programa de pós-graduação, tipo de atuação profissional, código de identificação do produto no ano base e no programa na base de dados da CAPES, identificador do autor (docente, discente e participante externo), nome do autor, forma de citação do nome do autor, dentre outras.

Enquanto o segundo - Produção Intelectual de Programas de Pós-Graduação *Stricto Sensu* no Brasil 2013 a 2016, tipo bibliográfica e Subtipo artigo em periódico[2] - disponibiliza dados sobre a produção no período, tais como: código de identificação do produto no ano base e no programa na base de dados da CAPES, título do produto, código de identificação da linha de pesquisa do programa a qual está vinculada a produção, projeto de pesquisa do programa ao qual está vinculada a produção, dentre outras.

A partir destes conjuntos de dados, foi iniciado o processo de enriquecimento, utilizando como base para a ligação o campo código de identificação do produto no ano base e no programa na base de dados da CAPES, codificado em ambos os conjuntos como ID_ADD_PRODUCAO_INTELECTUAL. O Quadro 1 apresenta um fragmento do CSV enriquecido a partir de dados disponíveis nos dois conjuntos de dados extraídos do portal Dados Abertos CAPES.

Quadro 1 – Fragmento dos Conjuntos de Dados em CSV

Subtipo	Código Programa	Nome programa	Sigla instituição	Identificação produto	Título	Autor
Artigo em periódico	28001010042P7	FILOSOFIA	UFBA	14715238	A DESCOBERTA DO EFEITO COMPTON: DE UMA ABORDAGEM SEMI-CLÁSSICA À QUÂNTICA	OLIVAL FREIRE JUNIOR; INDIANARA LIMA SILVA
Artigo em periódico	31075010001P2	MATEMÁTICA EM REDE NACIONAL	UTFPR	18479425	SIMULAÇÃO DE UM MODELO MATEMÁTICO QUE ANALISA AS DEFORMAÇÕES DE DOIS METAIS DEVIDO A PROBLEMAS DE CONTATO	LUZ DELICIA CASTILLO VILLALOBOS

[1] Disponível em: https://dadosabertos.capes.gov.br/dataset/7f7d09b0-f2fb-40d4-af4b-dccee9121763/resource/58536483-db4b-4040-a68a-f1ac042d6d11/download/br-capes-colsucup-prod-autor-2013a2016-2017-03-01-bibliografica-artpe.csv. Acesso em: 13 set. 2019.

[2] Disponível em: https://dadosabertos.capes.gov.br/dataset/04c63a4e-e9c9-4473-b6a1-244a935a7cae/resource/61218c67-b6ba-47cf-8234-2e190b846daf/download/br-capes-colsucup-producao-2013a2016-2017-11-01-bibliografica-artpe.csv. Acesso em: 13 set. 2019.

Subtipo	Código Programa	Nome programa	Sigla instituição	Identificação produto	Título	Autor
Artigo em periódico	33004030082P3	REABILITAÇÃO ORAL	UNESP/ARAR	12130538	EFFECT OF THERMAL AND MECHANICAL CYCLES ON THE HARDNESS AND ROUGHNESS OF ARTIFICIAL TEETH	JANAINA HABIB JORGE; HELOISA DE PAULA LEMOS TENAN; PAULA VOLPATO SANITA DANTAS; ANA CLAUDIA PAVARINA; EWERTON GARCIA DE OLIVEIRA MIMA
Artigo em periódico	33002010002P2	FÍSICA	USP	11796572	HEAT PRODUCTION OF NONINTERACTING FERMIONS SUBJECTED TO ELECTRIC FIELDS	WALTER ALBERTO DE SIQUEIRA PEDRA; C. HERTLING; J.-B. BRU
Artigo em periódico	12001015035P2	CIÊNCIA E ENGENHARIA DE MATERIAIS	UFAM	11831570	AMAZON RIVER DISSOLVED LOAD: TEMPORAL DYNAMICS AND ANNUAL BUDGET FROM THE ANDES TO THE OCEAN	TEREZA CRISTINA SOUZA DE OLIVEIRA
Artigo em periódico	24009016013P3	ZOOTECNIA	UFCG	16181505	MANEJO DA CAATINGA PARA PRODUÇÃO DE CAPRINOS E OVINOS	ADERBAL MARCOS DE AZEVEDO SILVA

Fonte: Adaptado de Dados Abertos CAPES (2019).

A adequação às 13 dimensões e 35 boas práticas para a publicação de dados na web não é objetivo do presente estudo, de forma que não há pretensão de esgotar esta discussão e sim analisar o conjunto de dados selecionados para proceder à sua formalização. Contudo o atendimento a duas das boas práticas consistiu em fator decisivo para a seleção: a disponibilização dos dados com uma licença Creative Commons Atribuição, que por seu caráter pouco restritivo permite reuso dos dados e sua ligação com outros dados; e a disponibilização em formato CSV, cuja leitura e processamento são facilmente realizados por humanos e aplicações computacionais.

Destaca-se ainda que, embora os dois conjuntos de dados tenham o mesmo período de referência, no momento da ligação dos dados foi possível verificar que não há coincidência exata nos dados disponibilizados. Desta forma, embora haja um código de identificação que permita ligar os dois conjuntos de dados, algumas produções não dispunham dos dados correspondentes aos autores.

Após a seleção dos dados, iniciou-se o processo de conversão e de enriquecimento, em que os dados disponíveis em CSV foram convertidos para RDF. Para tanto, utilizou-se a ferramenta Sparqlify[3], um reescritor SPARQL-SQL, que permite definir as visões RDF em bancos de dados relacionais e consultá-las com SPARQL.

Em outras palavras, o Sparqlify utiliza a linguagem SPARQL para converter os conjuntos de dados CSV em RDF, permitindo realizar as ligações necessárias para estruturar as informações. O Sparqlify permite, ainda, a ligação e a estruturação dos

[3] Disponível em: http://aksw.org/Projects/Sparqlify.html. Acesso em: 1 maio 2019.

dados a partir de esquemas e de ontologias. Neste estudo as ligações foram realizadas utilizando a ontologia VIVO, elaborada para tratar os relacionamentos entre as diversas áreas do domínio acadêmico.

Em suma, o processo de conversão utilizando a ferramenta Sparqlify contempla questões de modelagem de dados, seguindo os princípios da web semântica, estruturando as informações de acordo com o esquema da ontologia.

Nesse processo, refletiu-se sobre o esquema em que os dados estariam estruturados, a partir da ontologia VIVO, dos próprios dados disponíveis em CSV e de propriedades do OWL que poderiam ser aplicadas a partir das relações existentes entre os dados.

As ligações entre os dados utilizam algumas das propriedades do OWL, ao mesmo tempo em que permitem a ligação das informações que foram obtidas no CSV. Tais ligações expressam como o nível de semântica formal dos dados foi aumentado, o que possibilita, no momento da recuperação, a realização de inferências e a geração de mais valor na utilização ou na construção de aplicações que usam os dados para fornecer informações para os usuários. Vale destacar que para montar a estrutura proposta utilizou-se a Ontologia VIVO, como base para os relacionamentos propostos, como *bfo:Entity*, para fazer referência a entidades, e *bfo:Occurrent*, fazendo referência a localização e tempo que uma publicação foi realizada.

Esses elementos são utilizados por diversas aplicações que estão envoltas no contexto de dados acadêmicos, como a própria aplicação VIVO, que relaciona dados de publicação e autoria, focada em uma perspectiva mais institucional. Cabe ressaltar que a adoção de ontologias amplamente utilizadas pela comunidade favorece a disseminação e a interoperabilidade dos dados publicados, uma vez que permite que outras aplicações criem ligações e reutilizem os conjuntos de dados.

Para montar a estrutura utilizou-se a ferramenta Sparqlify, para que essa estrutura fosse adaptada para uma linguagem semelhante ao SPARQL, chamada de *Sparqlification Mapping Language* (SML)[4] a fim de realizar a conversão e a estruturação. A sintaxe do SML está disponível no documento construído pelo Sparqlify (2018). A Figura 3 apresenta um fragmento do código utilizado para realizar a conversão dos dados para RDF.

[4] Disponível em: http://sml.aksw.org/. Acesso em: 13 set. 2019.

Figura 3 – Fragmento do Código SML

```
Create View dados As
  Construct {
    ?s capes:ID_ADD_PRODUCAO_INTELECTUAL ?o .
    ?o dc:title ?titulo .
  }
  with
   ?s = uri(?x)
   ?o = uri(?x)
   ?titulo = plainLiteral(?name)
  Constrain
    ?x prefix "http://dados.artigo.capes/"
  From
    datacapes
```

Fonte: Autoria própria (2019).

No código apresentado na Figura 3, é possível verificar a semelhança com o SPARQL e como as relações foram construídas de modo a permitir a realização das ligações dos dados que são parte do conjunto disponibilizado pela CAPES.

A Figura 3 ilustra um trecho do processo da estruturação dos dados, seguindo os princípios da tripla RDF. Esse fragmento demonstra como se liga um objeto que se refere a uma produção intelectual ao seu título, o que pode, por sua vez, ser expandido para quaisquer outras informações vinculadas. Na sequência, são definidos os prefixos, ou seja, quais são os *namespaces* aos quais esses dados pertencem, no caso dos objetos, além de apontar que o título é um valor literal, como um texto. Na última parte do trecho, utiliza-se o termo “From”, que aponta a origem dos dados, no caso um arquivo contendo as informações dos dados da CAPES.

Adicionalmente, com o código em SML, realizou-se a conversão dos dados de CSV para RDF/XML, no qual todas as relações e a estrutura representada na Figura 2 podem ser visualizadas. Assim, obtiveram-se os dados em RDF com o nível de semântica formal adequado, atendendo aos princípios das boas práticas para a publicação de dados.

O RDF/XML demonstra que todo o processo realizado com a estruturação conceitual e transposta para o Sparqlify gerou dados no formato RDF/XML, um formato que pode ser inserido em bases de dados e ser disponibilizado para o acesso futuro. Sob esse conjunto, já podem ser aplicadas consultas em SPARQL, promovendo a recuperação da informação por meio das tecnologias da web semântica.

Por fim, o último passo para o enriquecimento dos dados seguindo os princípios do *linked data*, da web semântica e das boas práticas está na interligação dos dados com outras bases de dados que estão no contexto do *linked open data*.

A principal base de dados utilizada para realizar a interligação com o conjunto de dados selecionado e convertido para o RDF é o ORCID[5]. Utilizado por pesquisadores e acadêmicos, o ORCID fornece um identificador persistente individual que objetiva reduzir as ambiguidades de nomes, além de fazer ligações entre pesquisadores, suas atividades e produções.

Nos processos aplicados a partir do RDF obtido com a conversão do Sparqlify, inseriram-se os identificadores persistentes ORCID iD dos pesquisadores. Esse processo permite que os dados sejam ligados com outras bases, o que possibilita que os autores e pesquisadores sejam tratados de maneira única, usando uma base global. A inserção desses identificadores aprimora a recuperação dos dados, pois os usuários poderão encontrar as informações utilizando um registro adotado internacionalmente. Complementarmente, a inserção do ORCID permite que os dados estejam interligados a outra base, favorecendo a descoberta futura e o relacionamento com outras bases que também utilizam o identificador.

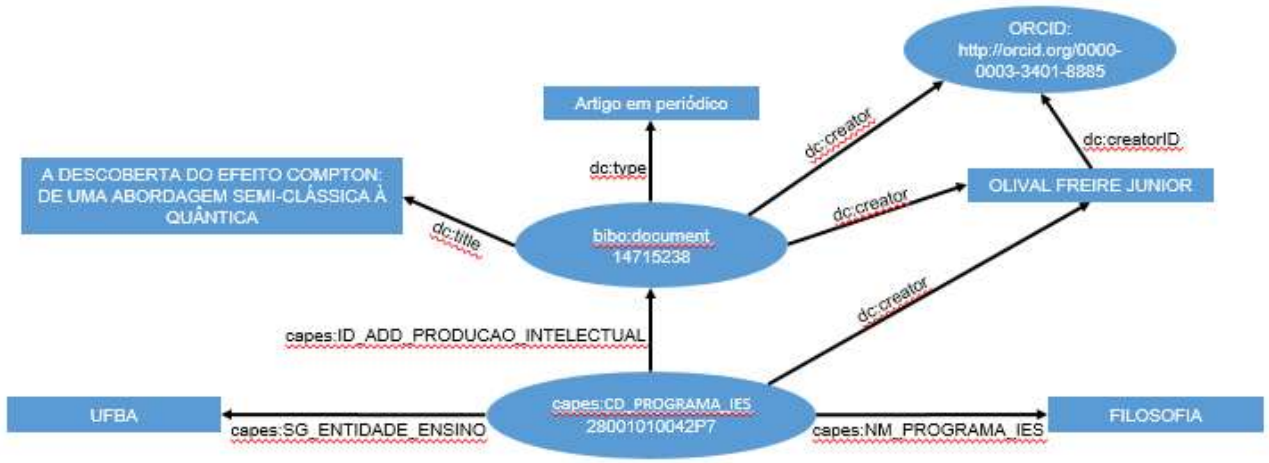
Vale observar que outras bases de dados podem ser utilizadas para interligar os recursos do portal Dados Abertos CAPES e fornecer relações importantes para contextualizar e relacionar os dados com as diversas bases do *linked open data*. Destacam-se o *Virtual International Authority File* (VIAF)[6], de registro de autoridade, e o DBpedia[7], de domínio geral.

A Figura 4 é um fragmento da estrutura conceitual concebida a partir dos dados obtidos das bases de dados da CAPES e apresenta, conceitualmente, um exemplo de dados estruturados, utilizando vocabulários diversos, inclusive de registro de autoridade.

Na Figura 4, identifica-se a presença de objetos diversos, como um programa de pós-graduação, que tem como produção um artigo, produzido por um pesquisador, que por sua vez está vinculado ao programa citado. Essa estrutura demonstra como, a partir dos dados obtidos no portal Dados Abertos CAPES, em que há uma série de relações de publicações, programas e autores, as informações foram relacionadas e enriquecidas com outros dados, neste contexto, a ligação com o ORCID.

[5] Disponível em: https://orcid.org/. Acesso em: 13 set. 2019.
[6] Disponível em: https://viaf.org/. Acesso em: 13 set. 2019.
[7] Disponível em: https://wiki.dbpedia.org/. Acesso em: 13 set. 2019.

Figura 4 – Representação Esquemática da Coleta e Modelagem de Dados



Fonte: Autoria própria (2018).

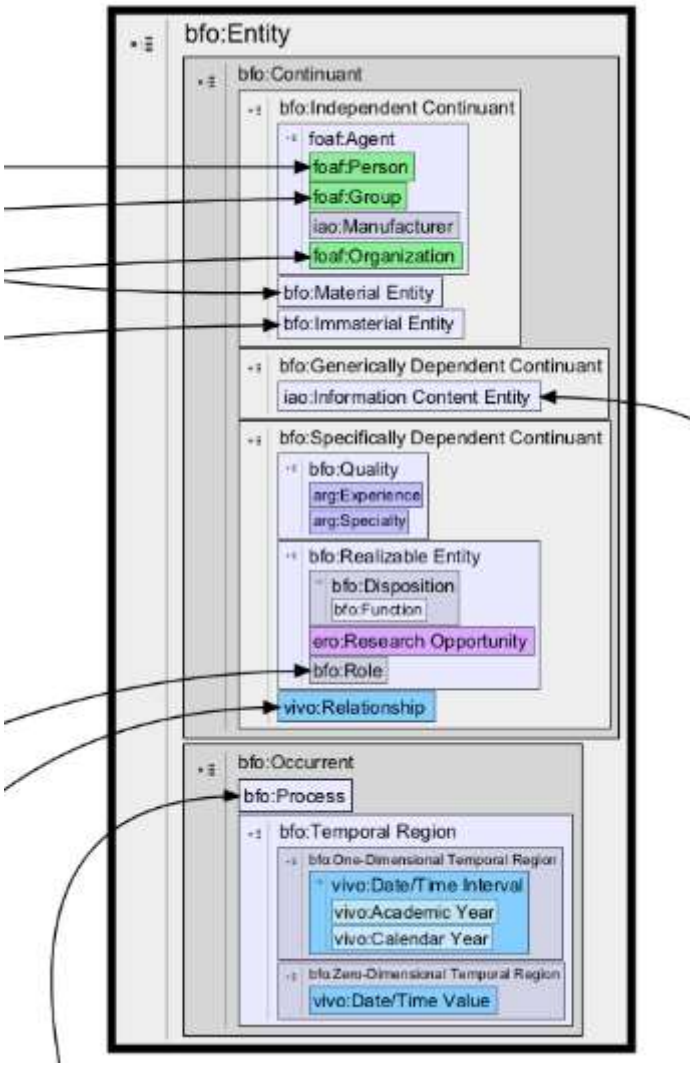
Destaca-se, ainda, o uso de vocabulários padronizados, como o Dublin Core, ao realizar as relações utilizando as propriedades *dc:title*, *dc:creator* e *dc:type*. Além deste foram utilizados *namespaces* específicos dos conjuntos de dados extraídos do portal Dados Abertos CAPES[8], para a inserção dos dados obtidos nos conjuntos de dados em CSV, o que é aderente aos princípios do *linked data* e da web semântica.

Um dos pontos relatados anteriormente, que pode ser visualizado na representação da figura 4, está na aplicação da ontologia VIVO. Nesta figura, utiliza-se o bibo:document, que é uma entidade da ontologia VIVO que traz características da *The Bibliographic Ontology*, utilizada pela ontologia VIVO para realizar a estruturação e os relacionamentos entre as diversas propriedades que os itens bibliográficos possuem. Desta forma, é possível ter a interoperabilidade com outros documentos que utilizam essa estrutura, tanto da VIVO quanto da *The Bibliographic Ontology.*

Assim, a Figura 5 demonstra como uma parte da ontologia da VIVO está vinculada a estas ligações com o *bibo:document* e o *bfo:entity*, utilizadas pela representação realizada neste trabalho. Destaca-se que o *bibo:document* é uma subclasse de *bfo:entity*, herdando assim as características desta segunda classe.

[8] Disponível em: https://metadados.capes.gov.br/index.php/home. Acesso em: 18 dez. 2019.

Figura 5 – Parte da representação da ontologia VIVO



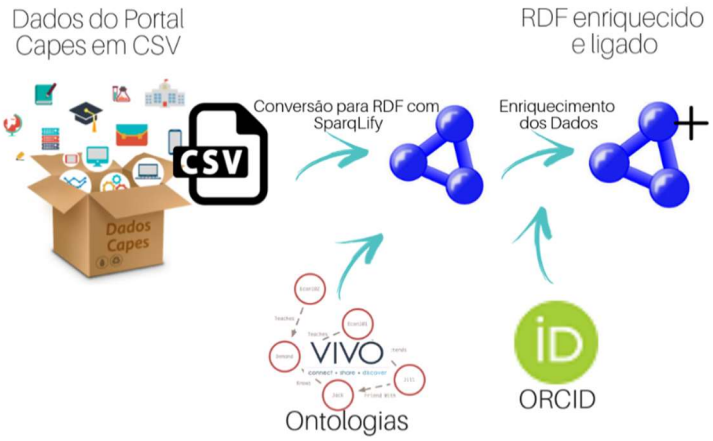
Fonte: Duraspace (2014).

No que tange à ligação realizada com a base ORCID, é possível acessar e buscar informações de ORCID iD a partir dos nomes de autores utilizando uma *Application Programming Interface* (API) disponibilizada pela própria base. Desta forma, foi possível realizar a leitura e a escrita dos códigos ORCID iD nos registros gerados, partindo dos nomes obtidos dos dados do portal Dados Abertos CAPES.

Ressalta-se que o uso da base ORCID permite o processo de reconciliação por meio de seu identificador único, favorecendo futura identificação de equivalência com outros dados, incluindo outras publicações não listadas na base.

Finalmente, a Figura 6 demonstra o processo aplicado nesse estudo, partindo da seleção do conjunto de dados, em CSV, obtido do portal Dados Abertos CAPES, até a disponibilização dos dados em RDF, enriquecidos e ligados.

Figura 6 – Representação Esquemática da Coleta e Modelagem de Dados



Fonte: Autoria própria (2018).

A Figura 6 origina-se da Figura 2, na qual os conceitos para o enriquecimento são apresentados para representar as fontes dos dados (portal Dados Abertos CAPES e ORCID), os formatos (CSV e RDF), as tecnologias (Ontologias) e a ferramenta (Sparqlify) utilizadas para realizar a conversão dos dados. Ilustra, ainda, como os diversos elementos e bases de dados podem ser utilizados para realizar o processo de enriquecimento semântico.

5 CONSIDERAÇÕES FINAIS

A discussão realizada neste estudo esclarece que a simples disponibilização de conjuntos de dados na web não assegura que possam ser acessados, utilizados e reutilizados, seja por humanos ou agentes computacionais. Neste contexto, as boas práticas para a publicação de dados na web, atreladas às ferramentas da web semântica, possibilitam a formalização dos conjuntos de dados visando ampliar seu alcance de uso e ligação com outros *datasets*.

Nesse estudo, descreveu-se a etapa de formalização de dados abertos, que compreende a seleção de dados, a formalização (estrutura, formatos e licenças) e o enriquecimento semântico, seguindo as boas práticas para disponibilização de dados na web e visando favorecer as etapas posteriores, de conversão ferramental, recuperação, marketing e feedback para a publicação de dados no contexto do *linked open data*.

As camadas abordadas nesse artigo foram: dados em sua forma sintática, conversão e análise, e modelagem de dados para estruturação ontológica. A partir deste estudo é possível visualizar como ocorre o processo de enriquecimento semântico de

dados no contexto do *linked open data*, abarcando a seleção, análise, processamento e preparação visando a disponibilização, o uso e o reuso.

Vale ressaltar que o presente estudo não pretendeu esgotar as possibilidades de enriquecimento semântico dos dados obtidos a partir do portal Dados Abertos CAPES, de forma que há a possibilidade de fazê-lo utilizando-se de outras bases de dados como, por exemplo, o *Digital Object Identifier* (DOI).

Destaca-se que o processo apresentado foi aplicado a conjuntos de dados abertos disponíveis na web no portal Dados Abertos CAPES em atendimento à Lei de Acesso à Informação. Desta forma, seria recomendável que a disponibilização de dados das instituições fosse planejada e efetivada no contexto do *linked open data* e, em atendimento às boas práticas preconizadas pelo W3C, para favorecer a recuperação, o uso e o reuso, bem como o enriquecimento semântico.

Espera-se que este estudo contribua com pesquisadores e profissionais no desenvolvimento de novas abordagens, bem como no processo prático de formalização de dados para a disponibilização na web no contexto do *linked open data*.

REFERÊNCIAS

ARAKAKI, F. **Linked data**: ligação de dados bibliográficos. 2016. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2016. Disponível em: https://repositorio.unesp.br/handle/11449/147979. Acesso em: 1 maio 2018.

BERNERS-LEE, T. **Linked data**. 2006. Disponível em: http://www.w3.org/DesignIssues/LinkedData.html. Acesso em: 4 jun. 2018.

BERNERS-LEE, T.; CONNOLLY, D. **Notation3 (n3)**: a readable RDF syntax. 2011. Disponível em: https://www.w3.org/TeamSubmission/n3/. Acesso em: 4 jun. 2018.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. **Scientific American**, v. 284, n. 5, p. 34-43, 2001. Disponível em: http://www.jstor.org/stable/26059207. Acesso em: 6 jun. 2018.

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. Acesso em: 4 jun. 2018.

BRASIL. Lei nº 12.527, de 18 de novembro de 2011. **Diário Oficial da União**, Brasília, DF, 18 nov. 2011. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm. Acesso em: 1 maio 2018.

DADOS abertos CAPES. Disponível em: https://dadosabertos.capes.gov.br/dataset/producao-intelectual-de-programas-de-pos-graduacao-2013-a-2016/resource/61218c67-b6ba-47cf-8234-2e190b846daf. Acesso em: 7 maio 2019.

DURASPACE. **VIVO-ISF Ontology v1.6 overviews**: classes. Beaverton: DuraSpace, 2014. Disponível em: https://wiki.duraspace.org/display/VTDA/VIVO-ISF+Ontology+v1.6+Overview%3A+Classes. Acesso em: 3 dez. 2019.

HAUSENBLAS, M. Exploiting linked data to build web applications. **IEEE Internet Computing**, v. 13, p. 68-73, jul./aug. 2009. Disponível em: http://doi.ieeecomputersociety.org/10.1109/MIC.2009.79. Acesso em: 4 jun. 2018.

HOOLAND, S.; VERBORGH, R. **Linked data for libraries, archives and museums**: how to clean, link and publish your metadata. Chicago: Neal-Schuman, 2014.

LÓSCIO, B. F.; BURLE, C.; CALEGARI, N. (Ed.). **Data on the web best practices**. 2017. Disponível em: https://www.w3.org/TR/dwbp/. Acesso em: 26 abr. 2018.

MURRAY-RUST, P. Open data in science. **Serials Review**, v. 34, n. 1, p. 52-64, 2008. Disponível em: https://openresearch-repository.anu.edu.au/bitstream/1885/46791/7/npre20081526-1%5B1%5D.pdf. Acesso em: 1 jun. 2018.

OPEN KNOWLEDGE INTERNATIONAL. **The open data handbook**. Disponível em: http://opendatahandbook.org/. Acesso em: 10 maio 2018.

SANTAREM SEGUNDO, J. E. **Web semântica**: fluxo para publicação de dados abertos e ligados. Informação em Pauta, Fortaleza, v. 3, n. esp., nov. 2018. Disponível em: http://www.periodicos.ufc.br/informacaoempauta/article/view/39721. Acesso em: 26 abr. 2018.

SANTAREM SEGUNDO, J. E. Web semântica: introdução a recuperação de dados usando SPARQL. **Encontro Nacional de Pesquisa em Ciência da Informação**, v. 15, 2014. Disponível em: http://www.brapci.inf.br/index.php/article/view/0000015784/2c3210137bae1fd29171ee96d39d888b/. Acesso em: 26 abr. 2018.

SANTAREM SEGUNDO, J. E.; CONEGLIAN, C. S.; LUCAS, E. R. de O. Conceitos e tecnologias da web semântica no contexto da colaboração acadêmico-científica: um estudo da plataforma Vivo. **Transinformação**, Campinas, v. 29, n. 3, p. 297-309, dez. 2017. Disponível em: http://www.scielo.br/pdf/tinf/v29n3/0103-3786-tinf-29-03-00297.pdf. Acesso em: 3 jun. 2018.

SAYÃO, L. F.; SALES, L. F. Dados abertos de pesquisa: ampliando o conceito de acesso livre. **Revista Eletrônica de Comunicação, Informação & Inovação em Saúde**, v. 8, n. 2, p. 76-92, jun. 2014. Disponível em: https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/611/1252. Acesso em: 1 jun. 2018.

SPARQLIFY. **Sparqlification Mapping Language (SML).** Disponível em: http://sparqlify.org/wiki/SML. Acesso em: 1 jun. 2018.

NOTAS

AGRADECIMENTOS
Não se aplica.

CONTRIBUIÇÃO DE AUTORIA
Os papéis descrevem a contribuição específica de cada colaborador para a produção acadêmica inserir os dados dos autores conforme exemplo, excluindo o que não for aplicável. Iniciais dos primeiros nomes acrescidas com o último Sobrenome, conforme exemplo.

Concepção e elaboração do manuscrito: E. Torino, G. L. Trevisan, C. S. Coneglian
Coleta de dados: E. Torino
Análise de dados: E. Torino, C. S. Coneglian
Discussão dos resultados: E. Torino, G. L. Trevisan, C. S. Coneglian. L. C. Botega, J. E. Santarem Segundo. S. A. B. G. Vidotti
Revisão e aprovação: L. C. Botega, J. E. Santarem Segundo. S. A. B. G. Vidotti

CONJUNTO DE DADOS DE PESQUISA

1) Os conjuntos de dados que dão suporte aos resultados deste estudo estão disponíveis publicamente no portal Dados Abertos CAPES:

Disponível em: https://dadosabertos.capes.gov.br/dataset/7f7d09b0-f2fb-40d4-af4b-dccee9121763/resource/58536483-db4b-4040-a68a-f1ac042d6d11/download/br-capes-colsucup-prod-autor-2013a2016-2017-03-01-bibliografica-artpe.csv. Acesso em: 13 set. 2019.
Disponível em: https://dadosabertos.capes.gov.br/dataset/04c63a4e-e9c9-4473-b6a1-244a935a7cae/resource/61218c67-b6ba-47cf-8234-2e190b846daf/download/br-capes-colsucup-producao-2013a2016-2017-11-01-bibliografica-artpe.csv. Acesso em: 13 set. 2019.

FINANCIAMENTO
CNPq - Bolsa em Produtividade em Pesquisa - CNPq PQ.

CONSENTIMENTO DE USO DE IMAGEM
Não se aplica.

APROVAÇÃO DE COMITÊ DE ÉTICA EM PESQUISA
Não se aplica.

CONFLITO DE INTERESSES
Não se aplica.

PUBLISHER
Universidade Federal de Santa Catarina. Programa de Pós-graduação em Ciência da Informação. Publicação no Portal de Periódicos UFSC. As ideias expressadas neste artigo são de responsabilidade de seus autores, não representando, necessariamente, a opinião dos editores ou da universidade.

EDITORES
Enrique Muriel-Torrado, Edgar Bisset Alvarez, Camila Barros.

HISTÓRICO
Recebido em: 14-10-2019 – Aprovado em: 12-01-2020 – Publicado em: 17-04-2020