

Análise do processo de recuperação da informação em bases de dados publicadas como dados abertos ligados utilizando a abordagem RDB2LOD

Clayton Martins Pereira

Doutorando; Universidade Estadual Paulista, Marília, SP, Brasil;
clayton.martins@unesp.br

Edberto Ferneda

Doutor; Universidade Estadual Paulista, Marília, SP, Brasil;
edberto.ferneda@unesp.br

José Eduardo Santarem Segundo

Doutor; Universidade de São Paulo, Ribeirão Preto, SP, Brasil;
santarem@usp.br

Resumo: Os dados abertos ligados têm se tornado um padrão para publicação e enriquecimento de dados, o que promove a transição de uma Web orientada a documentos para uma Web de dados e, por conseguinte, para a Web Semântica. Por outro lado, bases de dados relacionais compõem o núcleo da maioria dos sistemas de informação atualmente em operação. Assim, a publicação da imensa quantidade de dados mantidos em bases de dados relacionais, seguindo as boas práticas e recomendações do *Linked Data*, pode contribuir significativamente para a consolidação das ferramentas e tecnologias da Web Semântica. É nesse contexto que surgiu a abordagem RDB2LOD para publicação de dados abertos ligados obtidos a partir de bases de dados relacionais. Porém, depois de os dados serem efetivamente publicados, o passo seguinte é recuperá-los de forma eficiente para o seu devido consumo. Este trabalho, de natureza qualitativa e do tipo exploratório, tem como objetivo fazer uma análise do processo de recuperação da informação na abordagem RDB2LOD, a fim de averiguar se a utilização desta abordagem pode facilitar a formulação de consultas SPARQL e, conseqüentemente, melhorar a visualização e exploração dos dados recuperados. Para isso foi realizada uma pesquisa bibliográfica e documental, além de um experimento no qual a abordagem teve seu processo de recuperação da informação avaliado em dois casos distintos. Ficou demonstrado que, ao levar em consideração os aspectos semânticos dos termos empregados nas expressões de consulta, por meio da aplicação de ontologias, é possível tornar mais eficiente e precisa a recuperação de dados.

Palavras-chave: Recuperação da informação. Bases de dados relacionais. Dados ligados. Dados abertos ligados. Web Semântica.

1 Introdução

A Web Semântica é considerada uma extensão da Web atual, na qual os dados recebem um significado bem definido, facilitando o trabalho de computadores e pessoas (BERNERS-LEE; HENDLER; LASSILA, 2001). Na Web Semântica os significados das palavras são levados em consideração na formulação das consultas, o que possibilita a procura de informações considerando aspectos semânticos dos dados, de forma a obter resultados capazes de satisfazer as necessidades de informação do usuário (PABÓN; GONZÁLEZ, 2014). Para isso são usadas as ontologias, caracterizadas como “[...] um dos principais elementos da Web Semântica na construção de informações relacionadas que apresentam significado.” (SANTAREM SEGUNDO; CONEGLIAN, 2016, p. 220).

Assim, essa capacidade de organizar e usar as informações a partir de análise semântica, baseada em ontologias e vocabulários, pode permitir uma representação mais adequada e eficiente do conteúdo dos documentos. Com essa intenção, foram criadas linguagens e ferramentas que descrevem e representam com maior eficiência as informações acumuladas na Internet, como a *eXtensible Markup Language* (XML) e o *Resource Description Framework* (RDF) (CUBA RODRÍGUEZ; OLIVEIRA BATISTA, 2018). A partir do desenvolvimento dessas ferramentas, torna-se evidente o potencial da Web Semântica, também conhecida como Web inteligente, que é capaz de gerar avanços em diversos campos científicos, como a busca e recuperação da informação (CONEGLIAN *et al.*, 2017).

A evolução da Web e da Web Semântica propiciou o surgimento de iniciativas de abertura de dados por diversos governos e instituições, viabilizada pelo uso da tecnologia de dados ligados, ou *Linked Data*. Nesta tecnologia, um conjunto de dados ligados (*dataset*) é formado por um conjunto de triplas, representadas no formato RDF. Neste formato, uma tripla é composta por: (1) um sujeito/recurso (*subject*); (2) um objeto/valor (*object*); e (3) um relacionamento entre sujeito e objeto (*predicate*). Dessa forma, uma tripla pode ser lida como uma sentença composta por sujeito, predicado e objeto.

A promoção de dados ligados em um contexto de dados abertos (*open data*), isto é, dados que estão publicamente disponíveis na Web, ocorre por meio das recomendações e boas práticas do projeto *Linked Open Data* (LOD), ou dados abertos ligados, um refinamento do *Linked Data* (CRISTOVÃO; FERNANDES, 2018).

Com o uso crescente de *Linked Open Data*, torna-se cada vez mais importante para os provedores de dados não apenas publicar seus dados como LOD, mas também modelá-los de uma maneira fácil de processar, ou seja, tornar os dados mais legíveis por humanos e processáveis por máquina (SCHAIBLE; GOTTRON; SCHERP, 2014).

Por outro lado, bases de dados relacionais compõem o núcleo da maioria dos sistemas de informação atualmente em operação devido à sua maturidade e eficiência na forma de armazenagem e consulta dos dados, além de apresentarem alta confiabilidade e escalabilidade (LING; ZHOU, 2010).

Nesse sentido, a publicação da imensa quantidade de dados mantidos em bases de dados relacionais ao redor do mundo, seguindo as boas práticas e recomendações do *Linked Data*, pode contribuir significativamente para a consolidação das ferramentas e tecnologias da Web Semântica, promovendo a interoperabilidade entre sistemas e aplicações, bem como estimular o reuso de dados tanto por humanos quanto por máquinas, de forma a automatizar o conteúdo da Web que conhecemos hoje.

Foi nesse contexto que Pereira (2012) propôs uma abordagem, chamada de *Relational DataBase to Linked Open Data* (RDB2LOD), para a publicação de dados abertos ligados obtidos a partir de bases de dados relacionais por meio da integração entre as diversas ferramentas de software aplicadas neste processo. Utilizando essa abordagem, foi implementado um aplicativo que possibilita a personalização, de forma semiautomática, do arquivo de mapeamento entre a base de dados relacional e o modelo de dados RDF: o usuário pode incorporar a esse mapeamento uma ontologia de domínio com o intuito de atribuir significados bem definidos (aproximados à linguagem natural) aos sujeitos, predicados e objetos das triplas RDF a serem geradas. Além disso, a abordagem

oferece uma interface para consulta, visualização e exploração dos conjuntos de dados ligados.

Apesar de terem decorrido alguns anos desde a concepção da abordagem RDB2LOD, é possível considerar que ela ainda continue relevante para aplicação nos processos de publicação de dados na Web Semântica, uma vez que alguns trabalhos recentes neste campo (PATEL; JAIN, 2019; LANTI; XIAO; CALVANESI, 2019; ULUTAŞ KARAKOL *et al.*, 2018; DEVI; MEHROTRA; BAAZAOUI-ZGHAL, 2018) ainda utilizem como base as linguagens, ferramentas e *frameworks* (R2RML, D2RQ, Jena) integrados nesta abordagem.

Assim, o presente trabalho tem como objetivo fazer uma análise do processo de recuperação da informação na abordagem RDB2LOD, a fim de averiguar se a utilização desta abordagem pode facilitar a formulação de consultas *Sparql Protocol and RDF Query Language* (SPARQL) e, conseqüentemente, melhorar a visualização e exploração dos dados recuperados.

A metodologia empregada neste trabalho, de natureza qualitativa e do tipo exploratório, consistiu em uma pesquisa bibliográfica sobre os temas *publicação de dados e recuperação da informação na Web Semântica*, a fim de apresentar uma breve revisão sobre estes assuntos. Além disso, foi realizada uma pesquisa documental sobre a abordagem RDB2LOD, possibilitando seu detalhamento.

Em seguida, a capacidade de recuperação da informação da abordagem RDB2LOD foi avaliada por meio de um experimento, no qual a abordagem foi aplicada em duas bases de dados relacionais: uma contendo dados acadêmicos de um programa de pós-graduação, mantida por uma instituição de ensino superior, e outra contendo dados referentes ao acompanhamento (mortes, internações hospitalares e atendimentos ambulatoriais) de casos de câncer, mantida pelo Sistema Único de Saúde (SUS).

A escolha de tais bases de dados para a realização do experimento aqui apresentado se justifica pela disponibilidade de acesso a essas bases no âmbito do projeto de pesquisa em que este trabalho teve origem. Além disso, a diferença entre tamanho e domínio de aplicação dessas bases permitirá

demonstrar, como se verá, que a abordagem pode ser aplicada independentemente desses fatores.

Convém ainda esclarecer que os dados constantes na base de dados de acompanhamento de casos de câncer não permitem a identificação nominal de pacientes, tratando-se de dados gerais para fins de acompanhamento estatístico e de acompanhamento do impacto financeiro-orçamentário do tratamento desses pacientes na rede pública de saúde. Com relação à base de dados de teses defendidas em um programa de pós-graduação, trata-se de dados públicos provenientes do sistema de coleta de dados de pós-graduação da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

As seções seguintes deste trabalho estão assim organizadas: a seção 2 apresenta uma breve revisão sobre os processos de publicação de dados e de recuperação de informação na Web Semântica; a seção 3 mostra um detalhamento da abordagem RDB2LOD; a seção 4 faz a análise do processo de recuperação da informação desta abordagem, a partir de um experimento, e, por fim, a seção 5 apresenta as considerações finais e as sugestões de trabalhos futuros.

2 Publicação de dados e recuperação da informação na Web Semântica

Nos últimos anos, o paradigma do *Linked Open Data* tem se tornado um padrão para a publicação e o enriquecimento de dados, o que promove a transição de uma Web orientada a documentos para uma Web de dados e, por conseguinte, para a Web Semântica. O LOD permite a abertura de dados em formatos legíveis por máquina, deixando-os prontos para consumo e para reutilização e enriquecimento por meio de conexões com outros conjuntos de dados, possibilitando a criação de novos conhecimentos (SILVELLO *et al.*, 2017).

Isso tem se refletido num crescente interesse pela publicação de dados e, ao mesmo tempo, pelo desenvolvimento de aplicações que possam consumir esses dados publicados na Web Semântica. Além disso, contribuiu para o surgimento de um quarto paradigma da ciência, o *e-Science*, no qual a pesquisa

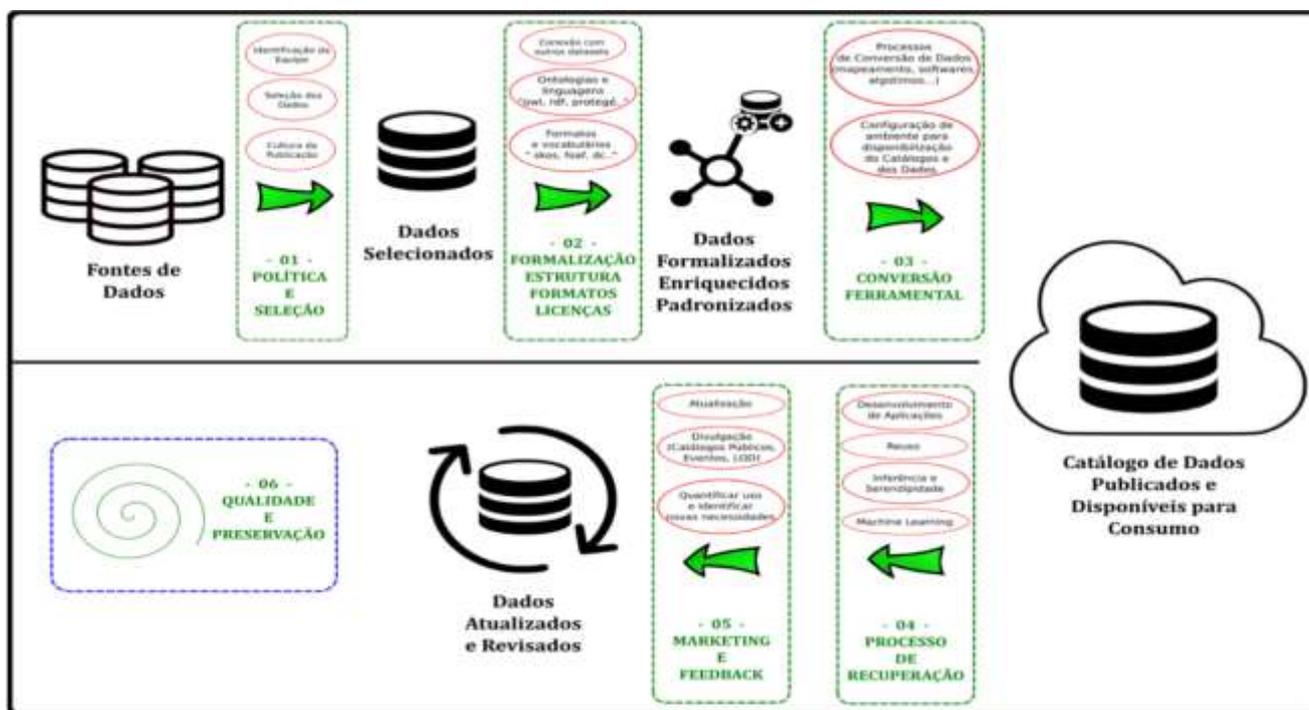
e desenvolvimento são baseados predominantemente em dados (SANTAREM SEGUNDO, 2018).

Entretanto, de acordo com Santarem Segundo (2018), “O processo de publicação de dados, que em diversas situações parece uma tarefa trivial, tem se tornado o grande problema das equipes ou pessoas que se propõe [sic] a realizá-lo [...]” (SANTAREM SEGUNDO, 2018, p. 120). Esse problema é o que motivou esse autor a apresentar uma proposta de um

[...] fluxo organizacional, segmentado em fases, que descreva as atividades que devem ser desenvolvidas no processo de publicação de dados em formato aberto e semântico, seguindo as melhores práticas de Dados Ligados. (SANTAREM SEGUNDO, 2018, p. 120).

A Figura 1 ilustra este modelo de fluxo organizacional para a publicação de dados ligados.

Figura 1 – Fluxo organizacional para publicação de dados ligados



Fonte: Santarem Segundo (2018).

Mesmo depois de terem sido publicados, os dados devem ser constantemente revistos e atualizados, de forma que estejam sempre disponíveis

para atender às necessidades de informação dos usuários, as quais exibem um comportamento dinâmico, ou seja, se modificam ao longo do tempo.

Depois de os dados serem efetivamente publicados, o próximo passo é buscá-los e recuperá-los de forma eficiente e precisa para seu devido consumo, transformação e eventual nova disponibilização para reuso.

Para tornar isso possível, a recuperação da informação abrange os aspectos intelectuais da descrição da informação e sua especificação para a busca, bem como quaisquer sistemas, técnicas ou máquinas que são utilizadas para realizar a operação (MOOERS, 1951¹ *apud* FERNEDA, 2019). Neste processo, as expressões de busca são o meio que o usuário emprega para comunicar a sua necessidade informacional para o sistema. Essas expressões podem ser especificadas em linguagem natural ou por meio de uma linguagem artificial, dependendo dos recursos oferecidos pelo sistema (FERNEDA, 2019).

Um sistema de recuperação da informação tem como meta encontrar a informação exigida para satisfazer uma necessidade de informação do usuário (FRANTZ; SHAPIRO; VOISKUNSKII, 1997² *apud* GONZALEZ; LIMA, 2003). Para tanto, esse tipo de sistema deve ser capaz de realizar, além das buscas, o armazenamento e manutenção de informação (KOWALSKI, 1997³ *apud* GONZALEZ; LIMA, 2003). Em outras palavras, deve representar, organizar e dar acesso a itens de informação (BAEZA-YATES, 1999⁴ *apud* GONZALEZ; LIMA, 2003).

Para tentar satisfazer as necessidades informacionais do usuário mais adequadamente, as pesquisas em recuperação da informação têm desviado seu foco da sintaxe para a semântica. Enquanto a sintaxe corresponde à análise de como as palavras se agrupam (estrutura gramatical) para formar frases, a semântica procura analisar os possíveis significados de uma frase, incluindo a desambiguação de palavras no seu contexto. Esse desvio de foco nas pesquisas surgiu, de acordo com Liddy (1998), a partir da utilização do processamento de linguagem natural em nível semântico nos sistemas de recuperação da informação, o qual busca implantar a desambiguação de palavras com múltiplos sentidos e incorporar, às consultas formuladas, todos os sinônimos dos termos utilizados nessas consultas, a fim de expandi-las.

Embora a tecnologia dos mecanismos de busca venha se aprimorando nas últimas décadas, as técnicas de descrição de conteúdo e processamento de consultas, que são predominantemente empregadas no campo da Recuperação da Informação, ainda se baseiam principalmente em palavras-chave e, portanto, fornecem recursos limitados para capturar e explorar as conceituações envolvidas nas necessidades de informação do usuário e nos significados do conteúdo recuperado (FERNÁNDEZ *et al.*, 2011).

Para tentar superar as limitações dos modelos de consulta baseados em palavras-chave, o desenvolvimento das tecnologias de busca semântica, esta entendida como a busca por significados, em vez de cadeias literais, tem sido o foco de diversas pesquisas nas áreas de recuperação da informação e de Web Semântica. Essas pesquisas empregam as ontologias como elementos-chave para a representação do conhecimento, de forma a constituir o núcleo destas tecnologias. Portanto, a ênfase destas pesquisas está no desenvolvimento de mecanismos capazes de capturar os termos de busca do usuário e convertê-los em uma representação formal de consulta por meio de uma linguagem semântica, como a SPARQL (FERNÁNDEZ *et al.*, 2011).

A linguagem SPARQL, considerada a linguagem *Structured Query Language* (SQL) da Web Semântica, permite a consulta de triplas RDF obtidas a partir de *datasets* ou de *triple stores* (depósitos de triplas) e consiste em um modelo em que a expressão de busca é formulada no formato de triplas, mas com a possibilidade de definir variáveis em quaisquer dos termos desta tripla, seja no sujeito, no predicado ou no objeto. Como exemplo, a Figura 2a mostra um conjunto de triplas RDF que contêm dados (identificador, título, autor e editora) de livros da literatura brasileira. Sobre essas triplas é efetuada a consulta SPARQL mostrada na Figura 2b, que busca pelos títulos (*name*) e autores (*creator*) dos livros cuja editora seja a *Companhia da Letras*. O resultado dessa consulta é mostrado na Figura 2c.

Figura 2a – Exemplo de conjunto de triplas RDF referentes a livros da literatura brasileira

```
base <http://example.org/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .

<#gabriela-cravo-canela>
  dc:identifier "9788535912388" ;
  foaf:name "Gabriela, Cravo e Canela" ;
  dc:creator "Jorge Amado" ;
  dc:publisher <#companhia-das-letras> .

<#vidas-secas>
  dc:identifier "9788501067340" ;
  foaf:name "Vidas Secas" ;
  dc:creator "Graciliano Ramos" ;
  dc:publisher <#grupo-record> .

<#antologia-poetica>
  dc:identifier "9788535921199" ;
  foaf:name "Antologia Poética" ;
  dc:creator "Carlos Drummond de Andrade" ;
  dc:publisher <#companhia-das-letras> .

<#companhia-das-letras>
  foaf:name "Companhia das Letras".

<#grupo-record>
  foaf:name "Grupo Editorial Record" .
```

Fonte: Adaptado de Laufer (2015).

Figura 2b – Consulta SPARQL formulada sobre o conjunto de triplas RDF da Figura 2a

```
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX ex: <http://example.org/>
SELECT ?name ?creator
WHERE
{
  ?book dc:publisher ex:companhia-das-letras .
  ?book foaf:name ?name .
  ?book dc:creator ?creator .
}
```

Fonte: Adaptado de Laufer (2015).

Figura 2c – Resultado da consulta SPARQL formulada na Figura 2b

| NAME | creator |
|----------------------------|------------------------------|
| "Gabriela, Cravo e Canela" | "Jorge Amado" |
| "Antologia Poética" | "Carlos Drummond de Andrade" |

Fonte: Adaptado de Laufer (2015).

Conforme visto, a linguagem SPARQL abre caminho para que uma expressão de consulta possa ser redigida de uma forma muito aproximada à linguagem natural. Os resultados obtidos nas consultas podem ainda assumir a

forma de dados ligados, os quais podem ser explorados e levar a outros resultados não obtidos inicialmente.

3 Abordagem RDB2LOD

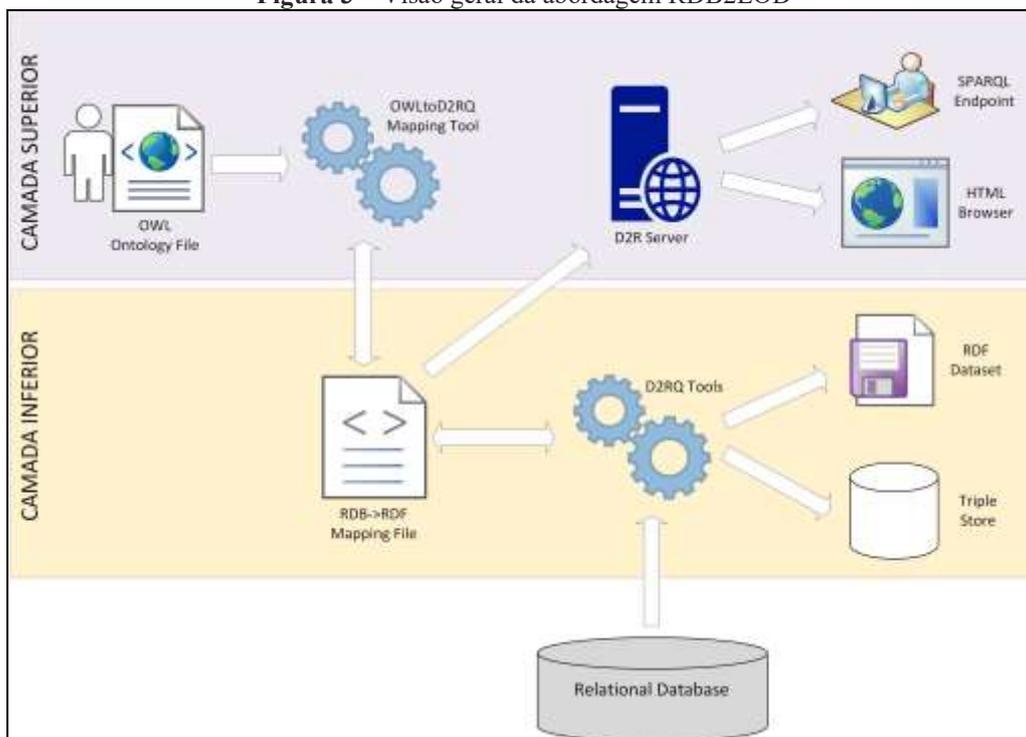
As ferramentas atualmente disponíveis para efetuar a conversão de dados mantidos em bases de dados relacionais para dados ligados, na forma de triplas RDF, têm diversas limitações que exigem do usuário um elevado nível de conhecimento técnico e de interação manual.

Para suprir essas limitações, a abordagem RDB2LOD oferece um aplicativo que integra alguma dessas ferramentas, por meio de uma interface gráfica pela qual é possível configurá-las e acioná-las, além de uma ferramenta que possibilita a personalização do arquivo de mapeamento pela incorporação de uma ontologia do domínio fornecida pelo usuário. Dessa forma, ficam bastante reduzidas as necessidades de interação manual e de conhecimento técnico por parte do usuário. A Figura 3 apresenta uma visão geral da abordagem, que é composta de duas camadas.

A camada inferior da abordagem é composta pelas ferramentas da plataforma D2RQ: *D2RQ Engine*, *D2RQ Generate Mapping*, *D2RQ Dump-RDF* e *D2R-Query*. A partir destas ferramentas, é possível efetuar o acesso à base de dados relacional e gerar o mapeamento de seu esquema para o modelo RDF (*RDB→RDF Mapping File*), o que possibilita a geração das triplas para um formar um *dataset* RDF ou o armazenamento delas em uma base de dados do tipo *triple store*.

Por sua vez, a camada superior conta com os seguintes componentes: a ferramenta *D2R-Server*, que permite publicar de forma dinâmica o conteúdo da base de dados relacional, ou seja, efetuar sob demanda a conversão dos dados da base mapeada para o formato de triplas RDF, sem a necessidade de replicar seu conteúdo em um *triple store*; e a ferramenta *OWLtoD2RQ-Mapping*, que tem por finalidade personalizar, de forma interativa e semiautomática, o arquivo de mapeamento para o formato RDF (*RDB→RDF Mapping File*) a partir de uma ontologia (*OWL Ontology File*) fornecida pelo usuário.

Figura 3 – Visão geral da abordagem RDB2LOD

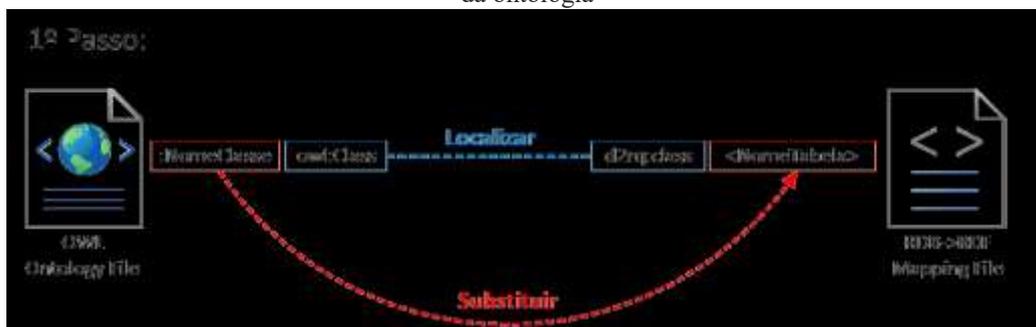


Fonte: Adaptado de Pereira (2012).

O método para incorporar os elementos de uma ontologia ao arquivo de mapeamento entre a base de dados relacional e o modelo RDF (*RDB→RDF Mapping File*), implementado pela ferramenta *OWLtoD2RQ-Mapping*, consiste inicialmente na substituição, no corpo do arquivo de mapeamento, dos nomes das tabelas da base de dados relacional pelos nomes das classes da ontologia que foram associados pelo usuário, por meio da interface gráfica da ferramenta.

Assim, para cada rótulo de classe a ser substituído, deve ser localizada no corpo do arquivo de mapeamento uma cláusula *d2rq:class* acompanhada do nome de uma tabela da base de dados mapeada. Uma vez localizada a cláusula, o nome da tabela é substituído pelo nome da classe da ontologia, de acordo com a associação feita pelo usuário. A Figura 4 ilustra este procedimento.

Figura 4 – Localização e substituição do nome da tabela da base de dados pelo nome da classe da ontologia



Fonte: Adaptado de Pereira (2012).

No passo seguinte, são localizados e substituídos, para cada classe no arquivo de mapeamento, os rótulos das respectivas propriedades, ou seja, os nomes das colunas de cada tabela da base de dados mapeada.

Assim, para cada rótulo de propriedade a ser substituído, deve ser localizada no corpo do arquivo de mapeamento uma cláusula *d2rq:property* acompanhada do nome de uma das colunas de uma tabela da base de dados mapeada. Uma vez localizada a cláusula, o nome da coluna é substituído pelo nome da propriedade da ontologia, também de acordo com a associação feita pelo usuário por meio da interface gráfica da ferramenta. A Figura 5 permite visualizar este procedimento.

Figura 5 – Localização e substituição do nome da coluna da base de dados pelo nome da propriedade da ontologia



Fonte: Adaptado de Pereira (2012).

Ao final destes procedimentos, o arquivo de mapeamento entre a base de dados relacional e o modelo RDF, já com os elementos da ontologia incorporados, pode ser utilizado para gerar os dados ligados na forma de triplas RDF. Assim, o valor (ou objeto) de cada tripla RDF a ser gerada corresponderá a um campo de uma coluna (predicado) referente a uma tabela (sujeito) da base

de dados mapeada, ou ainda ao campo chave de outra tabela relacionada (expressa por meio de uma URI, *Unified Resource Identifier*), quando essa coluna (predicado) for uma chave estrangeira da tabela mapeada.

Como resultado, a aplicação desse método possibilitará a atribuição de significados bem definidos (aproximados à linguagem natural) aos sujeitos, predicados e objetos das triplas RDF a serem geradas, a partir da personalização do mapeamento da base de dados relacional pela incorporação de uma ontologia fornecida pelo usuário.

Para prover as interfaces gráficas necessárias para a automatização e integração das ferramentas que compõem a abordagem, foi disponibilizado um aplicativo, em linguagem Java, chamado RDB2LOD. Este aplicativo possibilita a coleta de parâmetros, por meio de campos e listas ou botões de seleção, com a finalidade de automatizar a elaboração e execução das linhas de comando para acionamento e operação de cada uma das ferramentas aplicadas na abordagem.

A seguir será efetuada uma análise do processo de recuperação da informação nesta abordagem, a partir da realização de um experimento.

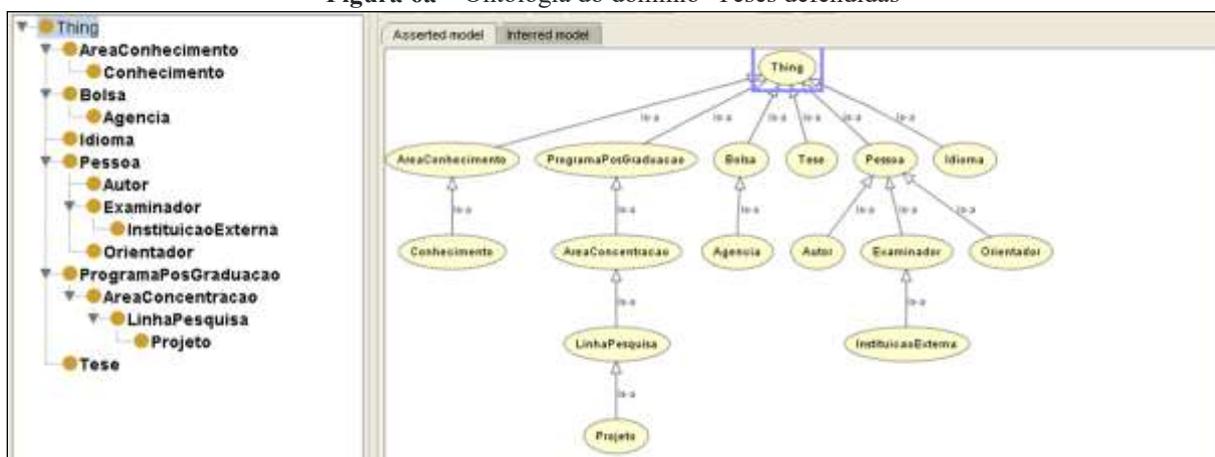
4 Análise do processo de recuperação da informação na abordagem RDB2LOD

Para analisar o processo de recuperação da informação na abordagem RDB2LOD, foi realizado um experimento em que essa abordagem foi aplicada em duas bases de dados relacionais: uma contendo dados acadêmicos de um programa de pós-graduação, mantida por uma instituição de ensino superior, e outra contendo dados referentes ao acompanhamento (mortes, internações hospitalares e atendimentos ambulatoriais) de casos de câncer, mantida pelo SUS.

Conforme justificou-se na introdução deste trabalho, por serem bases de dados de tamanhos e domínios de aplicações diferentes, será possível mostrar também neste experimento que a abordagem RDB2LOD pode ser aplicada independentemente destes fatores.

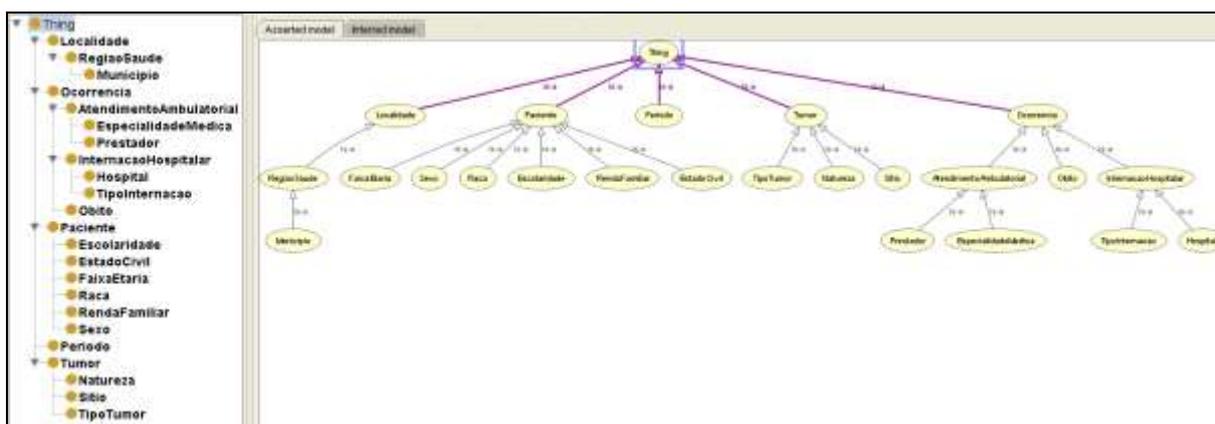
Inicialmente, para aplicar a abordagem RDB2LOD às bases de dados escolhidas, a fim de gerar as triplas RDF, na forma de dados ligados, foram criadas as ontologias dos respectivos domínios, em formato OWL (*Ontology Web Language*), utilizando o software para construção de ontologias *Protégé*. Representações destas ontologias são apresentadas na Figura 6a (para o domínio ‘teses defendidas’) e na Figura 6b (para o domínio ‘acompanhamento de casos de câncer’).

Figura 6a – Ontologia do domínio ‘Teses defendidas’



Fonte: Pereira (2012).

Figura 6b – Ontologia do domínio ‘Acompanhamento de casos de câncer’

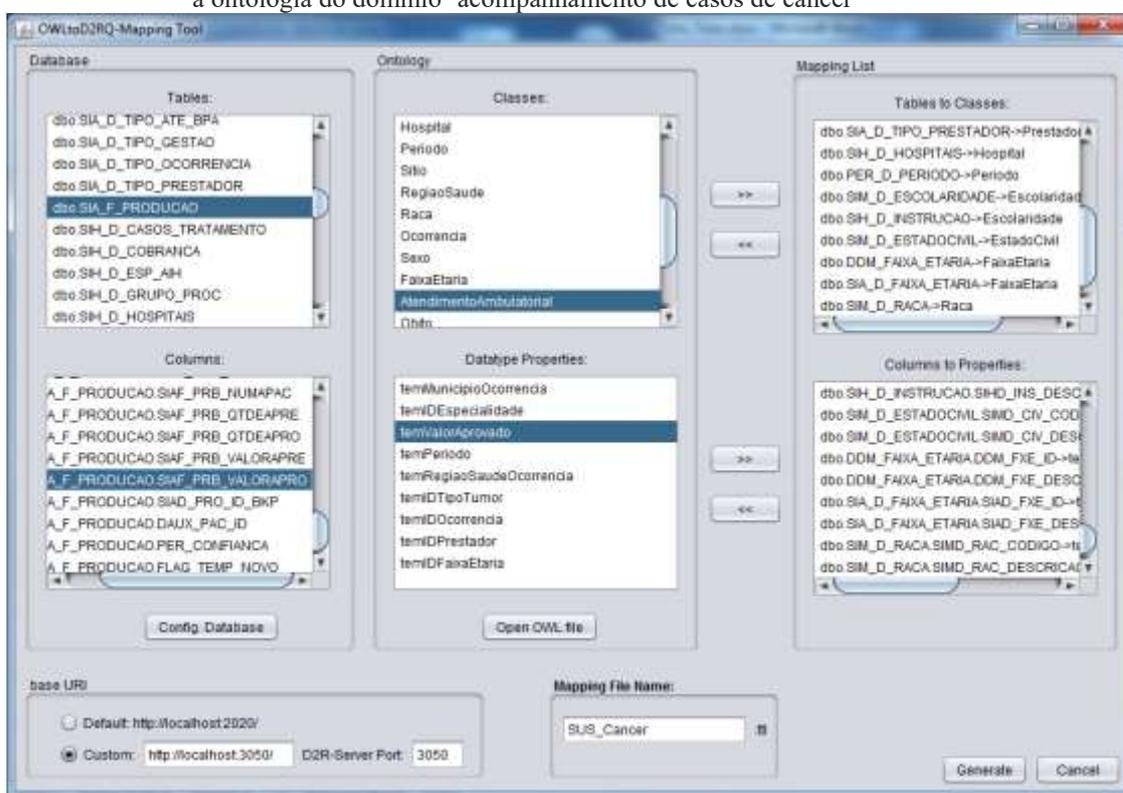


Fonte: Pereira (2012).

A partir das ontologias criadas e dos parâmetros de conexão aos respectivos servidores de gerenciamento, torna-se possível efetuar os mapeamentos das referidas bases de dados por meio da ferramenta *OWLtoD2RQ-Mapping*, disponibilizada pela abordagem RDB2LOD.

Para construir os mapeamentos nesta ferramenta, as tabelas e as colunas das bases de dados, assim como as classes e as propriedades das ontologias, são exibidas em uma janela do aplicativo, na forma de caixa de seleção, permitindo ao usuário efetuar a associação entre as tabelas e classes e entre as colunas e propriedades, tal como exhibe a Figura 7 (para o domínio ‘acompanhamento de casos de câncer’).

Figura 7 – Tela da ferramenta *OWLtoD2RQ-Mapping* para mapeamento entre a base de dados e a ontologia do domínio ‘acompanhamento de casos de câncer’



Fonte: Pereira (2012).

Uma vez construídos os mapeamentos entre as bases de dados e as respectivas ontologias do domínio, os dados estarão prontos para formarem as triplas RDF e serem publicados de forma aberta e ligada na Web. A partir daí é possível utilizar as ferramentas da abordagem RDB2LOD para o processo de recuperação da informação, com consultas (*queries* SPARQL) e com a respectiva visualização dos dados ligados (triplas RDF) recuperados.

Para prosseguir com a análise do processo de recuperação da informação na abordagem RDB2LOD, formulou-se uma consulta em linguagem SPARQL

para cada base de dados utilizada neste experimento. Para a base de dados de teses defendidas, a consulta formulada teve por finalidade obter a relação dos docentes que orientaram teses escritas no idioma inglês. Para a base de dados de acompanhamento de casos de câncer, a consulta formulada teve por finalidade obter a relação dos municípios do Estado de São Paulo, com população entre dez mil e cem mil habitantes, nos quais se registraram ocorrências de internação hospitalar de pacientes acometidos de câncer.

As Figuras 8a e 8b apresentam as consultas formuladas, utilizando o nome padrão de cada tabela e coluna da base de dados mapeada, ou seja, sem a aplicação de ontologia. Nestas figuras é importante observar os predicados de cada sentença das consultas formuladas e tentar compreender seus significados.

É possível observar nas Figuras 8a e 8b que as consultas formuladas a partir dos nomes padrão das tabelas e colunas das bases de dados são semelhantes à formulação de consultas SQL tradicionais. Neste caso é preciso que o usuário saiba antecipadamente tais nomes.

Figura 8a – Consulta SPARQL formulada utilizando o nome padrão das tabelas e colunas da base de dados de teses defendidas

```
SPARQL:
PREFIX db: <http://localhost:2020/assessor/>
PREFIX edf: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX vocab: <http://localhost:2020/assessor/vocab/>

SELECT DISTINCT ?NomeOrientador ?IDOrientador ?Titulo ?Tese ?NomeAutor ?IDAutor WHERE {
  (?Orientador vocab:col_r_teses_orientadores_IDIESE ?Tese .
  ?Tese vocab:col_teses_IDIDIOMA ?Idioma .
  ?Idioma vocab:g_idiomas DESCRICAO "Inglês" .)
  OPTIONAL (?Orientador vocab:col_r_teses_orientadores_IDPESSOALORIENTADOR ?IDOrientador .
  ?IDOrientador vocab:col_pessoal_NOME ?NomeOrientador .)
  OPTIONAL (?Tese vocab:col_teses_TITULOTESE ?Titulo .)
  OPTIONAL (?Tese vocab:col_teses_IDPESSOAL ?IDAutor .
  ?IDAutor vocab:col_pessoal_NOME ?NomeAutor .)
}
```

Fonte: Pereira (2012).

Figura 8b – Consulta SPARQL formulada utilizando o nome padrão das tabelas e colunas da base de dados de acompanhamento de casos de câncer

```

SPARQL:
PREFIX db: <http://localhost:2020/resource/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX map: <http://localhost:2020/resource/#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX vocab: <http://localhost:2020/resource/vocab/>

SELECT DISTINCT ?Municipio ?NomeMunicipio ?Populacao
WHERE {
  {?Internacao vocab:dbo_SIH_F_AIH_UNT_MUN_IDUNI ?Municipio .
  ?Municipio vocab:dbo_UNT_MUNICIPIOS_UNT_MUN_NOME ?NomeMunicipio .
  ?Municipio vocab:dbo_UNT_MUNICIPIOS_UNT_MUN_CODUF ?IDUF .
  ?IDUF vocab:dbo_UNT_UF_UNT_UF_SIGLA "SP" .
  ?Municipio vocab:dbo_UNT_MUNICIPIOS_UNT_MUN_POPTOTAL ?Populacao .
  FILTER (?Populacao > 10000 && ?Populacao < 100000) }
}
ORDER BY ?Populacao
Results: Browse Go! Reset
  
```

Fonte: Pereira (2012).

A seguir, as Figuras 9a e 9b apresentam as mesmas consultas formuladas anteriormente, porém agora com o mapeamento entre as tabelas e colunas das bases de dados e as classes e propriedades das respectivas ontologias de domínio. Mais uma vez é importante observar os predicados de cada sentença e identificar se a aplicação de ontologias, feita por meio da abordagem, proporcionou melhor compreensão de seus significados.

Figura 9a – Consulta SPARQL formulada com o mapeamento entre as classes e propriedades da ontologia e as tabelas e colunas da base de dados de teses defendidas

```

SPARQL:
PREFIX : <file:///C:/RDB2LOD/ontologies/TesesColeta_rdf.owl#>
PREFIX db: <http://localhost:3040/resource/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX d2r: <http://sites.wiwiw.de/suhl/bizer/d2r-server/config.rdf#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX map: <http://localhost:3040/resource/#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX vocab: <http://localhost:3040/resource/vocab/>

SELECT DISTINCT ?NomeOrientador ?IDOrientador ?Titulo ?Tese ?NomeAutor ?IDAutor WHERE {
  {?Orientador :temTeseOrientada ?Tese .
  ?Tese :temIDIIdioma ?Idioma .
  ?Idioma :temNomeIdioma "Inglês" .}
  OPTIONAL {?Orientador :temIDOrientador ?IDOrientador .
  ?IDOrientador :temNome ?NomeOrientador .}
  OPTIONAL {?Tese :temTitulo ?Titulo .}
  OPTIONAL {?Tese :temIDAutor ?IDAutor .
  ?IDAutor :temNome ?NomeAutor .}
}
Results: Browse Go! Reset
  
```

Fonte: Pereira (2012).

Figura 9b – Consulta SPARQL formulada com o mapeamento entre as classes e propriedades da ontologia e as tabelas e colunas da base de dados de acompanhamento de casos de câncer

```

SPARQL:
PREFIX : <file:///C:/RDB2LOD/ontologies/SUSCancer_rdf.owl#>
PREFIX db: <http://localhost:3050/resource/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX d2r: <http://sites.wiwiw.fu-berlin.de/suhl/bizer/d2r-server/config.rdf#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX map: <http://localhost:3050/resource/#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX vocab: <http://localhost:3050/resource/vocab/>

SELECT DISTINCT ?Municipio ?NomeMunicipio ?Populacao
WHERE {
  {?Internacao :temMunicipioInternacao ?Municipio .
  ?Municipio :temNomeMunicipio ?NomeMunicipio .
  ?Municipio :temIDUFMunicipio ?IDUF .
  ?IDUF :temSiglaUF "SP" .
  ?Municipio :temPopulacaoMunicipio ?Populacao .
  FILTER (?Populacao > 10000 && ?Populacao < 100000) }
}
ORDER BY ?Populacao
Results: Browse Go! Reset
  
```

Fonte: Pereira (2012).

Nestas figuras é possível notar que a formulação das mesmas consultas SPARQL, quando utilizados os termos das respectivas ontologias do domínio, se assemelha à escrita em linguagem natural, como pode ser observado nos predicados de cada sentença, o que possibilitou uma compreensão mais clara dos significados de cada predicado. Com isso, é possível notar também que houve uma redução no tamanho das sentenças formuladas em ambas as consultas.

No próximo passo, será analisado o impacto do uso da abordagem RDB2LOD nos resultados das consultas SPARQL formuladas neste experimento. As Figuras 10a e 10b exibem os resultados obtidos da execução das consultas ilustradas nas Figuras 8a e 8b, respectivamente.

Figura 10a – Resultados da consulta SPARQL ilustrada na Figura 8a

| SPARQL results: | | | | | |
|------------------------|-----------------------------------|---|---------------------------------|--|-------------------------------------|
| NomeOrientador | IDOrientador | Titulo | Tese | NomeAutor | IDAutor |
| "Elder Moreira Hemery" | db:col_pessoal/38 | "INDIRECT ADAPTIVE PREDICTIVE CONTROL APPLIED TO AN INDUSTRIAL TANK LEVEL PLANT" | db:col_teses/21 | "Ivan Garcia Martinez" | db:col_pessoal/4023 |
| "Takashi Yoneyama" | db:col_pessoal/47 | "FAULT-TOLERANT STATE ESTIMATION OF LINEAR GAUSSIAN SYSTEMS SUBJECT TO ADDITIVE FAULTS" | db:col_teses/22 | "Davi Antonio dos Santos" | db:col_pessoal/2762 |
| "Celso Massaki Hirata" | db:col_pessoal/1 | "A TIMESTAMP-BASED TWO PHASE COMMIT PROTOCOL FOR WEB SERVICES" | db:col_teses/34 | "Luiz Alexandre Hiane da Silva Maciel" | db:col_pessoal/4208 |
| "Takashi Yoneyama" | db:col_pessoal/47 | "FAILURE PROGNOSIS METHODS AND OFFLINE PERFORMANCE EVALUATION" | db:col_teses/39 | "Bruno Paes Leão" | db:col_pessoal/3486 |

Fonte: Pereira (2012).

Figura 10b – Resultados da consulta SPARQL ilustrada na Figura 8b

SPARQL results:

| Município | NomeMunicípio | Populacao |
|------------------------------|---------------------------|-----------|
| db:dbo/UNT_MUNICIPIOS/353600 | "Parapuã " | 10907 |
| db:dbo/UNT_MUNICIPIOS/351070 | "Cardoso " | 11178 |
| db:dbo/UNT_MUNICIPIOS/354860 | "São Bento do Sapucaí " | 11395 |
| db:dbo/UNT_MUNICIPIOS/354290 | "Ribeirão Bonito " | 11819 |
| db:dbo/UNT_MUNICIPIOS/351390 | "Divinolândia " | 12142 |
| db:dbo/UNT_MUNICIPIOS/353630 | "Patrocínio Paulista " | 12481 |
| db:dbo/UNT_MUNICIPIOS/355080 | "São Sebastião da Grama " | 12858 |
| db:dbo/UNT_MUNICIPIOS/352280 | "Itaporanga " | 14314 |
| db:dbo/UNT_MUNICIPIOS/355270 | "Tabatinga " | 14367 |
| db:dbo/UNT_MUNICIPIOS/351100 | "Castilho " | 15159 |
| db:dbo/UNT_MUNICIPIOS/351540 | "Fartura " | 15436 |
| db:dbo/UNT_MUNICIPIOS/353700 | "Pedregulho " | 15788 |
| db:dbo/UNT_MUNICIPIOS/352600 | "Junqueirópolis " | 16564 |
| db:dbo/UNT_MUNICIPIOS/352800 | "Macatuba " | 17183 |
| db:dbo/UNT_MUNICIPIOS/352740 | "Lucélia " | 18625 |
| db:dbo/UNT_MUNICIPIOS/352460 | "Jacupiranga " | 18676 |
| db:dbo/UNT_MUNICIPIOS/354000 | "Pompéia " | 18754 |

Fonte: Pereira (2012).

Nos resultados apresentados acima, é possível observar que a utilização dos nomes padrão das colunas das bases de dados não dificultou a compreensão dos significados das URIs geradas, porém é importante ressaltar que os nomes de colunas em esquemas de bases de dados nem sempre são triviais como neste caso.

A seguir, nas Figuras 11a e 11b, são exibidos os resultados obtidos da execução das consultas ilustradas nas Figuras 9a e 9b, que utilizam as propriedades das respectivas ontologias do domínio. Nestes resultados é possível observar que as URIs geradas permitem uma compreensão mais objetiva de seus significados.

Figura 11a – Resultados da consulta SPARQL ilustrada na Figura 9a

SPARQL results:

| NomeOrientador | IDOrientador | Titulo | Tese | NomeAutor | IDAutor |
|------------------------|--------------|---|------------|--|----------------|
| "Elder Moreira Hemery" | db:Pessoa/38 | "INDIRECT ADAPTIVE PREDICTIVE CONTROL APPLIED TO AN INDUSTRIAL TANK LEVEL PLANT" | db:Tese/21 | "Ivan Garcia Martinez" | db:Pessoa/4023 |
| "Takashi Yoneyama" | db:Pessoa/47 | "FAULT-TOLERANT STATE ESTIMATION OF LINEAR GAUSSIAN SYSTEMS SUBJECT TO ADDITIVE FAULTS" | db:Tese/22 | "Davi Antonio dos Santos" | db:Pessoa/2762 |
| "Celso Massaki Hirata" | db:Pessoa/1 | "A TIMESTAMP-BASED TWO PHASE COMMIT PROTOCOL FOR WEB SERVICES" | db:Tese/34 | "Luiz Alexandre Hiane da Silva Maciel" | db:Pessoa/4208 |
| "Takashi Yoneyama" | db:Pessoa/47 | "FAILURE PROGNOSIS METHODS AND OFFLINE PERFORMANCE EVALUATION" | db:Tese/39 | "Bruno Paes Leão" | db:Pessoa/3486 |

Fonte: Pereira (2012).

Figura 11b – Resultados da consulta SPARQL ilustrada na Figura 9b

| SPARQL results: | | |
|---------------------|----------------------------|-----------|
| Município | NomeMunicípio | Populacao |
| db:Município/353600 | "Parapuã " | 10907 |
| db:Município/351070 | "Cardoso " | 11178 |
| db:Município/354860 | "São Bento do Sapucaí " | 11395 |
| db:Município/354290 | "Ribeirão Bonito " | 11819 |
| db:Município/351390 | "Divinolândia " | 12142 |
| db:Município/353630 | "Patrocínio Paulista " | 12481 |
| db:Município/355080 | "São Sebastião da Gramma " | 12858 |
| db:Município/352280 | "Itaporanga " | 14314 |
| db:Município/355270 | "Tabatinga " | 14367 |
| db:Município/351100 | "Castilho " | 15159 |
| db:Município/351540 | "Fartura " | 15436 |
| db:Município/353700 | "Pedregulho " | 15788 |
| db:Município/352600 | "Junqueirópolis " | 16564 |
| db:Município/352800 | "Macatuba " | 17183 |
| db:Município/352740 | "Lucélia " | 18625 |
| db:Município/352460 | "Jacupiranga " | 18676 |
| db:Município/354000 | "Pompéia " | 18754 |

Fonte: Pereira (2012).

Na próxima etapa deste experimento, a partir dos resultados obtidos em cada consulta, será acionada uma das URIs geradas, o que levará à execução automática de uma nova consulta SPARQL, possibilitando a obtenção de novos dados não previstos na consulta inicial.

Assim, nos resultados exibidos na Figura 10a, será acionada a URI <db:col_pessoal/1> como exemplo, que apresentará como resultado, na Figura 12a, mais informações relacionadas à pessoa (docente orientador de tese) a que a URI se refere, tais como as URIs para outras teses de que esse docente foi orientador e/ou de cuja banca examinadora esse docente foi membro.

Da mesma forma, nos resultados exibidos na Figura 10b, será acionada a URI <db:dbo/UNT_MUNICIPIOS/355080> como exemplo, que apresentará como resultado, na Figura 12b, mais informações relacionadas ao município a que a URI se refere, tais como população, latitude, longitude e as URIs para a região de saúde e para os casos de morte, internação e/ou atendimento ambulatorial decorrente de câncer.

É possível perceber que nos novos resultados exibidos, que tiveram origem em consultas formuladas com a utilização dos nomes padrão das colunas das bases de dados (sem aplicação de ontologia), o usuário pode encontrar maior dificuldade para entender os significados das URIs geradas.

Figura 12a – Resultados da nova consulta executada automaticamente a partir de uma URI (nome padrão) acionada nos resultados apresentados na Figura 10a

| property | hasValue | isValueOf |
|---|------------------------|---------------------------------------|
| db:vocab/col_pessoal_NOME | "Celso Massaki Hirata" | - |
| db:vocab/col_pessoal_SEQUENCIAL | 1 | - |
| rdf:type | db:vocab/col_pessoal | - |
| rdfs:label | "col_pessoal #1" | - |
| db:vocab/col_r_teses_banca_examinadora_IDPESSOAEXAMINADOR | - | db:col_r_teses_banca_examinadora/1/31 |
| db:vocab/col_r_teses_banca_examinadora_IDPESSOAEXAMINADOR | - | db:col_r_teses_banca_examinadora/1/34 |
| db:vocab/col_r_teses_banca_examinadora_IDPESSOAEXAMINADOR | - | db:col_r_teses_banca_examinadora/1/50 |
| db:vocab/col_r_teses_orientadores_IDPESSOAORIENTADOR | - | db:col_r_teses_orientadores/1/34 |

Fonte: Pereira (2012).

Figura 12b – Resultados da nova consulta executada automaticamente a partir de uma URI (nome padrão) acionada nos resultados apresentados na Figura 10b

| property | hasValue | isValueOf |
|---|-------------------------------|---------------------------------|
| db:vocab/dbo_UNT_MUNICIPIOS_UNT_MUN_CODUF | db:dbo/UNT_UF/35 | - |
| db:vocab/dbo_UNT_MUNICIPIOS_UNT_MUN_ID | 355080 | - |
| db:vocab/dbo_UNT_MUNICIPIOS_UNT_MUN_LATITUDE | -22 | - |
| db:vocab/dbo_UNT_MUNICIPIOS_UNT_MUN_LONGITUDE | -47 | - |
| db:vocab/dbo_UNT_MUNICIPIOS_UNT_MUN_NOME | "São Sebastião da Gramma" | - |
| db:vocab/dbo_UNT_MUNICIPIOS_UNT_MUN_POPTOTAL | 12858 | - |
| db:vocab/dbo_UNT_MUNICIPIOS_UNT_MUN_S_IDHM | 0.78 | - |
| db:vocab/dbo_UNT_MUNICIPIOS_UNT_RGS_ID | db:dbo/UNT_REGIOES_SAUDE/3520 | - |
| rdf:type | db:vocab/dbo_UNT_MUNICIPIOS | - |
| rdfs:label | "UNT_MUNICIPIOS #355080" | - |
| db:vocab/dbo_SIA_F_PRODUCAO_UNT_MUNATE_ID | - | db:dbo/SIA_F_PRODUCAO/124030766 |
| db:vocab/dbo_SIH_F_AIH_UNT_MUN_IDUNI | - | db:dbo/SIH_F_AIH/18596316 |
| db:vocab/dbo_SIM_F_OBITOS_UNT_MUNRES_ID | - | db:dbo/SIM_F_OBITOS/1655040 |
| db:vocab/dbo_SIM_F_OBITOS_UNT_MUNRES_ID | - | db:dbo/SIM_F_OBITOS/191237 |

Fonte: Pereira (2012).

Por outro lado, são apresentados nas Figuras 13a e 13b os mesmos resultados, porém agora originados das consultas formuladas utilizando o mapeamento de ontologias, como visto nas Figuras 11a (<db:Pessoal/1>) e 11b (<db:Municipio/355080>). É possível notar como a exibição dos resultados em ambas as consultas ficou mais simples e como o entendimento do significado das URIs geradas, que se assemelham à escrita em linguagem natural, ficou mais fácil.

Figura 13a – Resultados da nova consulta executada automaticamente a partir de uma URI (nome personalizado) acionada nos resultados apresentados na Figura 11a

| SPARQL results: | | |
|------------------|------------------------|--------------------|
| property | hasValue | isValueOf |
| :temIDPessoal | 1 | - |
| :temNome | "Celso Massaki Hirata" | - |
| rdf:type | :Pessoa | - |
| rdfs:label | "Pessoa #1" | - |
| :temIDExaminador | - | db:Examinador/1/31 |
| :temIDExaminador | - | db:Examinador/1/34 |
| :temIDExaminador | - | db:Examinador/1/50 |
| :temIDOrientador | - | db:Orientador/1/34 |

Fonte: Pereira (2012).

Figura 13b – Resultados da nova consulta executada automaticamente a partir de uma URI (nome personalizado) acionada nos resultados apresentados na Figura 11b

| property | hasValue | isValueOf |
|--------------------------|--------------------------|--------------------------------------|
| :temIDHMunicipio | 0.78 | - |
| :temIDMunicipio | 355080 | - |
| :temIDRegiaoMunicipio | db:RegiaoSaude/3520 | - |
| :temIDUFMunicipio | db:Localidade/35 | - |
| :temLatitudeMunicipio | -22 | - |
| :temLongitudeMunicipio | -47 | - |
| :temNomeMunicipio | "São Sebastião da Grama" | - |
| :temPopulacaoMunicipio | 12858 | - |
| rdf:type | :Municipio | - |
| rdfs:label | "Municipio #355080" | - |
| :temMunicipioAtendimento | - | db:AtendimentoAmbulatorial/124030766 |
| :temMunicipioInternacao | - | db:InternacaoHospitalar/18596316 |
| :temMunicipioObito | - | db:Obito/1655040 |
| :temMunicipioObito | - | db:Obito/191237 |

Fonte: Pereira (2012).

Neste experimento foi possível observar que houve um **ganho de informação** quando foram adicionadas as classes e propriedades de uma ontologia ao processo de mapeamento para geração de conjuntos de dados ligados a partir dos dados mantidos em uma base de dados relacional.

Nos conjuntos de dados ligados produzidos a partir desta abordagem, os sujeitos, predicados e objetos das triplas geradas são de fácil compreensão e permitem que expressões de busca na linguagem SPARQL sejam formuladas levando-se em conta a semântica, ou os significados, de cada termo que as compõe. Isso certamente trará como resultado uma informação mais precisa e de maior valor para o usuário, que ainda poderá navegar pelos resultados, sem a formulação de novas consultas, com o intuito de obter novas informações ou

relacionamentos que não estavam previstos na consulta inicial, de forma a atender mais efetivamente sua necessidade de informação. Esta navegação pelos dados, ou ainda, os dados que levam a outros dados, é um dos benefícios proporcionados pelos dados ligados (*Linked Data*), que possibilitarão a consolidação da Web de Dados.

5 Considerações Finais

Neste trabalho foi realizada uma análise do processo de recuperação da informação na abordagem RDB2LOD, na qual foi possível averiguar que a utilização desta abordagem pôde facilitar a formulação de consultas SPARQL e, conseqüentemente, melhorar a visualização e exploração dos dados recuperados nessas consultas.

Para avaliar a recuperação da informação na abordagem RDB2LOD, foi realizado um experimento que consistiu na aplicação dessa abordagem a duas bases de dados distintas (de tamanhos e domínios diferentes), a fim de gerar triplas RDF que puderam ser recuperadas a partir de consultas formuladas em linguagem SPARQL em duas situações: utilizando os nomes padrão das tabelas e colunas das bases de dados, tal como numa consulta SQL tradicional, e aplicando o mapeamento das ontologias dos respectivos domínios, por meio de ferramenta disponibilizada pela abordagem, o que proporcionou significados bem definidos aos sujeitos, predicados e objetos das triplas geradas.

A partir do experimento, ficou demonstrado que, quando aplicado o mapeamento de ontologias ao processo de geração de triplas RDF, a formulação de consultas SPARQL, no processo de recuperação, apresentou escrita muito próxima à linguagem natural, o que contribui para que o usuário encontre menor dificuldade em expressar sua necessidade de informação, levando-o a uma experiência satisfatória no uso de um sistema de recuperação da informação baseado neste mecanismo de busca. Por conseguinte, os resultados retornados por essas consultas também apresentaram sujeitos, predicados e objetos, na forma de URIs, com rótulos escritos da mesma forma (linguagem próxima à natural), o que facilita a compreensão de seus significados e possibilita a

navegação pelos dados ligados gerados de forma intuitiva e sem a necessidade de formulação de novas consultas.

Desta forma, quando se levam em consideração os aspectos semânticos (significados) dos termos empregados nas expressões de consulta, o que é proporcionado pela aplicação de ontologias, fica possível tornar mais eficiente e precisa a recuperação de dados para consumo, transformação e eventual disponibilização para reuso, o que vai ao encontro da literatura consultada.

Espera-se assim que este trabalho contribua para o avanço das áreas de recuperação da informação e de Web Semântica, bem como estimule o aumento da publicação de dados ligados e o consequente reuso dos dados mantidos em bases de dados relacionais, o que pode impulsionar a consolidação da Web de dados e ampliar a interoperabilidade e a integração de aplicações e sistemas na Web.

Como trabalhos futuros são sugeridos: a adaptação da abordagem RDB2LOD, com a respectiva análise do processo de recuperação da informação a partir do mapeamento simultâneo de múltiplas bases de dados a uma ontologia (domínio com dados distribuídos em mais de uma base de dados), e o acréscimo, ao mapeamento efetuado pela abordagem, da associação entre as tabelas da base de dados e as *object properties* da ontologia, além do mapeamento de regras, restrições e cardinalidade para as classes, o que certamente aumentará a precisão do mecanismo de busca.

Referências

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. **Scientific American**, New York, v. 284, p. 28-37, 2001.

CONEGLIAN, C. S. *et al.* O papel estratégico da Web Semântica no contexto do big data. In: WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA, 1., 2017, Florianópolis. **Anais [...]**. Florianópolis: UFSC, 2017. p. 1-6.

CRISTOVÃO, H. M.; FERNANDES, J. H. C. Recuperação de informação em dados ligados: um modelo baseado em mapas conceituais e análise de redes complexas. **Transinformação**, Campinas, v. 30, n. 2, p. 193-207, 2018.

CUBA RODRÍGUEZ, Y.; OLIVERA BATISTA, D. Los metadatos, la búsqueda y recuperación de información desde las Ciencias de la Información. **e-Ciencias de la Información**, San José, v. 8, n. 2, p. 3-13, 2018.

DEVI, R.; MEHROTRA, D.; BAAZAOUI-ZGHAL, H. Pubworld - A R2RML mapping driven approach to transform relational database data into shareable format. *In*: INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS INTERNATIONAL ADVANCE COMPUTING CONFERENCE, 8., 2018, Greater Noida. **Proceedings** [...]. Piscataway: IEEE, 2018. p. 221-227.

FERNÁNDEZ, M. *et al.* Semantically enhanced Information Retrieval: an ontology-based approach. **Journal of Web Semantics**, Amsterdam, v. 9, n. 4, p. 434-452, 2011.

FERNEDA, E. **Material da disciplina Recuperação de Informação: técnicas e tecnologias**. Marília: Unesp, 2019. 1 diapositivo. Acesso em: 6 set. 2019.

GONZALEZ, M.; LIMA, V. L. S. Recuperação de Informação e Processamento da Linguagem Natural. *In*: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 23., 2003, Campinas. **Anais** [...]. Porto Alegre: SBC, 2003. p. 347-395.

LANTI, D.; XIAO, G.; CALVANESE, D. VIG: Data scaling for OBDA benchmarks. **Semantic Web**, Amsterdam, v. 10, n. 2, p. 413-433, 2019.

LAUFER, C. **Guia de Web semântica**. São Paulo: Projeto SPUK, 2015.

LIDDY, E. D. Enhanced Text retrieval using natural language processing. **Bulletin of the American Society for Information Science and Technology**, New Jersey, v. 24, n. 4, p. 14-16, 1998.

LING, H.; ZHOU, S. Translating relational databases into RDF. *In*: INTERNATIONAL CONFERENCE ON ENVIRONMENTAL SCIENCE AND INFORMATION APPLICATION TECHNOLOGY, 2., 2010, Wuhan. **Proceedings** [...]. Piscataway: IEEE, 2010. p. 464-467.

PABÓN, O. S.; GONZÁLEZ, M. E. del S. M. Propuesta para extender semánticamente el proceso de recuperación de información. **Escuela de ingeniería de Antioquia**, Envigado, v. 11, n. 22, p. 51-65, 2014.

PATEL, A.; JAIN, S. Present and future of semantic web technologies: a research statement. **International Journal of Computers and Applications**, Abingdon, p. 1-10, 2019.

PEREIRA, C. M. **Uma abordagem para a publicação de dados abertos ligados obtidos a partir de bases de dados relacionais**. 2012. Dissertação (Mestrado em Informática) - Curso de Pós-Graduação em Engenharia Eletrônica

e Computação, Instituto Tecnológico de Aeronáutica, São José dos Campos, 2012.

SANTAREM SEGUNDO, J. E.; CONEGLIAN, C. S. Web semântica e Ontologias: um estudo sobre construção de axiomas e uso de inferências. **Informação & Informação**, Londrina, v. 21, n. 2, p. 217-244, 2016.

SANTAREM SEGUNDO, J. E. Web semântica: fluxo para publicação de dados abertos e ligados. **Informação em Pauta**, Fortaleza, v. 3, n. esp., p. 117-140, 2018.

SCHAIBLE, J.; GOTTRON, T.; SCHERP, A. Survey on common strategies of vocabulary reuse in Linked Open Data modeling. *In*: PRESUTTI, V.; D'AMATO, C.; GANDON, F.; D'AQUIN, M.; STAAB, S.; TORDAI, A. (ed.). **ESWC 2014: the Semantic Web: trends and challenges**. Berlin: Springer, 2014. p. 457-472 (Lecture Notes in Computer Science, v. 8465).

SILVELLO, G. *et al.* Semantic representation and enrichment of information retrieval experimental data. **International Journal of Digital Libraries**, Berlin, v. 18, p. 145-172, 2017.

ULUTAŞ KARAKOL, D. *et al.* Semantic linking spatial RDF data to the web data sources. **International Archives of Photogrammetry and Remote Sensing Spatial Information Science**, Delft, v. XLII-4, p. 639-645, 2018.

Analysis of the information retrieval process in databases published as linked open data using the RDB2LOD approach

Abstract: Linked Open Data has become a standard for data publishing and data enrichment, and it supports the transition from a document-driven Web to an interconnected Web of data and thus to the Semantic Web. On the other hand, relational databases make up the core of most information systems currently in operation due to their maturity and efficiency in the form of storing and querying data. Thus, publishing the vast amount of data maintained in relational databases around the world in line with the good practices and recommendations of Linked Data can contribute significantly to the widespread adoption of Semantic Web tools and technologies. It is in this context that appeared the RDB2LOD approach for publishing Linked Open Data obtained from relational databases. However, once data is effectively published, the next step is efficiently and accurately searching and retrieving it for suitable use. This qualitative and exploratory work aims to analyze the information retrieval process in the RDB2LOD approach, in order to find out if the use of this approach can help to formulate SPARQL queries and, consequently, to improve the visualization and exploration of the retrieved data. For this, a bibliographic

and documentary study was carried out, along with an experiment where the RDB2LOD approach's information retrieval process was evaluated in two different cases. It was demonstrated that the consideration of the semantic aspects of terms in query expressions and the application of ontologies might improve data retrieval efficiency and accuracy.

Keywords: Information retrieval. Relational databases. Linked Data. Linked Open Data. Semantic Web.

Recebido: 04/12/2019

Aceito: 13/04/2020

Como citar:

PEREIRA, Clayton Martins; FERNEDA, Edberto; SANTAREM SEGUNDO, José Eduardo. Análise do processo de recuperação da informação em bases de dados publicadas como dados abertos ligados utilizando a abordagem RDB2LOD. **Em Questão**, Porto Alegre, v. 26, n. 3, p. 94-120, set./dez. 2020. DOI: <https://doi.org/10.19132/1808-5245263.94-120>



¹ MOOERS, C. Zatocoding applied to mechanical organization of knowledge. **American Documentation**, [s. l.], v. 2, n. 1, p. 20-32, 1951. *Apud* Ferneda (2019).

² FRANTZ, V.; SHAPIRO, J.; VOISKUNSKII, V. **Automated information retrieval: theory and methods**. San Diego: Academic Press, 1997. *Apud* Gonzalez e Lima (2003).

³ KOWALSKI, G. **Information retrieval systems: theory and implementation**. Berlin: Kluwer Academic Publishers, 1997. *Apud* Gonzalez e Lima (2003).

⁴ BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. New York: ACM Press, 1999. *Apud* Gonzalez e Lima (2003).