relato de pesquisa

MODELO DE DADOS ABERTOS CONECTADOS PARA INFORMAÇÃO LEGISLATIVA

Mariana Baptista Brandt[*]
Silvana Aparecida Borsetti Gregorio Vidotti[**]
José Eduardo Santarem Segundo[***]

RESUMO A presente pesquisa objetiva propor um modelo de dados abertos conectados (*linked open data - LOD*), para um conjunto de dados abertos legislativos da Câmara dos Deputados. Para tanto, procede-se à revisão de literatura sobre os conceitos de dados abertos, dados abertos governamentais, dados conectados (*linked data*), e dados abertos conectados (*linked open data*), seguido de pesquisa aplicada, com a modelagem de dados legislativos no modelo LOD. Para esta pesquisa foi selecionado o conjunto de dados "Deputados", que contém informações como partido político, unidade federativa, e-*mail*, legislatura, entre outras, sobre os parlamentares. Desse modo, observa-se que a estruturação do conjunto de dados em RDF (*Resource Description Framework*) é possível com reuso de vocabulários e padrões já estabelecidos na Web Semântica como Dublin Core, Friend of a Friend (FOAF), RDF e RDF Schema, além de vocabulários de áreas correlatas, como a Ontologia da Câmara dos Deputados italiana e a da Assembleia Nacional Francesa. Conforme recomendação do padrão *Linked Data*, os recursos foram relacionados também a outros conjuntos de LOD para enriquecimento semântico, como as bases Geonames e DBpedia. O estudo que permite concluir que a disponibilização dos dados governamentais, em especial, dados legislativos, pode ser feita seguindo as recomendações da W3C (*World Wide Web Consortium*) e, assim, integrar os dados legislativos à Web de Dados e ampliar as possibilidades de reuso e aplicações dos dados em ações de transparência e fiscalização, aproximando os cidadãos do Congresso e de seus representantes.

Palavras-chave: Dados abertos conectados. *Linked Data*. Web Semântica. Dados abertos governamentais. Dados legislativos.

* Mestre em Ciência da Informação pela Universidade de Brasília, Brasil. Doutoranda no Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Julio de Mesquita Filho, Brasil.
E-mail: marianabrandt@gmail.com.

** Doutora em Educação pela Universidade Estadual Paulista Júlio de Mesquita Filho, Brasil. Docente permanente do Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Julio de Mesquita Filho, Brasil.
E-mail: vidotti@marilia.unesp.br.

*** Doutor em Ciência da Informação pela Universidade Estadual Paulista Júlio de Mesquita Filho, Brasil. Docente permanente do Programa de Pós-Graduação em Ciência da Informação da Universidade Estadual Paulista Julio de Mesquita Filho, Brasil.
E-mail: santarem@usp.br.

1 INTRODUÇÃO

A *web* vem evoluindo de um modelo baseado em documentos para um modelo baseado em dados. A produção de dados tem crescido de forma exponencial nos últimos anos e esses dados podem ser produzidos por humanos, por processamento de dados que geram novos dados ou por agentes computacionais e sensores, que produzem dados a cada segundo. Essa realidade vem gerando enormes conjuntos de dados e o chamado fenômeno do *Big Data*. Passados mais de 30 anos da existência da *web*, as instituições entenderam e absorveram a necessidade da criação de padrões, regras, formatos e orientações para a publicação desses dados na *web* da melhor forma para serem compreendidos e reutilizados.

Os dados do setor público configuram conjuntos de informações muito importantes para a sociedade e, por força de lei (Lei

12.527/2011 - Lei de Acesso à Informação), devem estar disponíveis em formato aberto, isto é, de acordo com princípios que permitam que eles sejam manipulados, reutilizados e trabalhados de maneira livre, adequando-se ao conceito de dados abertos governamentais. Além disso, é essencial que a publicação desses dados esteja também alinhada com as melhores práticas e diretrizes das comunidades de prática da *web*. Assim, os dados governamentais configuram instrumento de transparência das ações dos governos no mundo todo e aplicações bastante úteis já foram desenvolvidas para a sociedade a partir desses conjuntos de dados.

Nesse contexto, destacam-se os dados legislativos como de grande importância para o cidadão brasileiro, principalmente no que diz respeito aos parlamentares e aos projetos de lei em tramitação no Congresso. Os deputados, por terem sido eleitos e representarem o cidadão, e os projetos de lei por serem propostas de reguladores das atividades que impactam na vida de todos.

A disponibilidade desses dados é de grande importância para que essas informações sejam acessadas e compreendidas pelo cidadão por meio de implementações utilizando esses conjuntos de dados.

Berners-Lee (2009) afirma que os propósitos da disponibilização de dados governamentais online são melhor atendidos quando se usa técnicas de *Linked Data*. O objetivo deste trabalho é propor um modelo de *linked open data* para dados abertos da Câmara dos Deputados, utilizando os vocabulários conhecidos da Web Semântica, *Dublin Core - DC* e *Friend of a Friend – FOAF*, os vocabulários temáticos *Vocabulario de la estrutura de organismos públicos* e *Vocabulario de Resultados Electorales*, além das ontologias da Câmara dos Deputados da Itália, da Assembleia Nacional Francesa e da DBpedia. Para isso, foi selecionado, entre os conjuntos de dados abertos da instituição, um conjunto para projeto piloto de modelagem em RDF – *Resource Data Framework*, criando um conjunto de dados abertos conectados para os dados legislativos. Além disso, a proposta contempla enriquecimento semânticos dos conjuntos de dados abertos modelados por meio de ligações com outros conjuntos de dados e vocabulários pertinentes.

O trabalho justifica-se pela importância dos dados legislativos e pela necessidade de estruturação desses dados em formato aberto, legível por máquina e com ligações semânticas que possam enriquecer seus conteúdos, fornecendo contexto ao usuário. Conforme afirma Santarem Segundo (2015, p. 221): "Entende-se que cultura de publicação de dados abertos no Brasil é muito tímida e quando acontece é carente de melhor infraestrutura, tanto do ponto de vista do armazenamento quanto da real possibilidade de acesso pelos usuários finais, a sociedade civil." Acredita-se, portanto, que a observância das melhores práticas do *linked data* na publicação dos dados legislativos trará qualidade a esse processo e aos dados publicados, o que deverá resultar em dados e informações mais acessíveis, atualizadas, confiáveis e contextualizadas ao público, permitindo melhor aplicação dos dados nos diferentes propósitos.

2 DADOS ABERTOS, GOVERNAMENTAIS, CONECTADOS

A emergência da *web* de dados trouxe consigo novos conceitos e características a esses dados. O conceito de dados abertos está relacionado com a disponibilização de conjuntos de dados em formato aberto, o que significa dados brutos, acessíveis por máquina e livres de restrições de uso. Um grande foco das iniciativas relacionadas a dados abertos são os dados governamentais. As ações de abertura de dados de governo dão origem ao conceito de dados abertos governamentais. Tais dados ganharam ênfase nas iniciativas de dados abertos devido ao valor que possuem para a sociedade, conforme afirma Santarem Segundo (2015, p. 223): "Há também no mundo uma tendência de publicação de dados governamentais, com o objetivo de criar a cultura de participação do cidadão na gestão do Estado, construindo um modelo conhecido como transparência.". Além da transparência, o *World Wide Web Consortium* – W3C, consórcio mundial para o desenvolvimento de padrões para a *Web*, identifica como áreas em que os dados abertos governamentais estão gerando valor: participação popular, empoderamento dos cidadãos, melhores ou novos produtos e serviços privados, inovação, melhoria na eficiência e efetividade de serviços governamentais, entre outros. Segundo Heath e Bizer (2011, não paginado, tradução nossa):

> Tornar esses dados acessíveis possibilita que organizações e membros da sociedade trabalhem com os dados, analisem-nos para descobrir novas ideias e criem ferramentas que ajudem a comunicar essas descobertas para outros, ajudando assim os cidadãos a tomar decisões informadas e a cobrar responsabilidade dos servidores públicos.

No Brasil, o que impulsionou as iniciativas de dados abertos governamentais foi a Lei de Acesso à Informação, que tornou obrigatória a publicação de dados abertos pelos órgãos públicos brasileiros e, mais recentemente, o Decreto 8.777/2016, que instituiu a Política de Dados Abertos do Poder Executivo Federal e tornou obrigatória a publicação de um Plano de Dados Abertos para os órgãos do poder executivo federal brasileiro.

A *web* de dados que surgiu e vem crescendo encontra-se no contexto da Web Semântica, proposta de atribuição de significado semântico aos recursos disponíveis na *web*. Imaginada por Tim Berners-Lee, a Web Semântica é uma proposta de estruturação de dados na *Web* "de forma que eles possam ter significado e principalmente que se tornem passíveis de interpretação por máquinas, através de agentes computacionais." (SANTAREM SEGUNDO, CONEGLIAN, 2016, p. 2). O W3C considera a Web Semântica como a própria *web* de dados conectados. Para Isotani e Bittencourt (2015), "A Web de Dados cria inúmeras oportunidades para a integração semântica dos próprios dados, motivando o desenvolvimento de novos tipos de aplicações e ferramentas, como navegadores e motores de busca".

Santarem Segundo (2015, p. 224) afirma que: "Nos últimos anos vários elementos foram surgindo e ampliando o contexto da ideia original de Web Semântica de Berners-Lee". Entre esses novos elementos está o *Linked Data* (dados ligados ou conectados) como tecnologia da Web Semântica para a representação do conhecimento. O termo refere-se essencialmente a aplicações de conexão entre conjuntos de dados e outros dados ou informações que forneçam contexto e significado a esses conjuntos de dados, de forma estruturada e seguindo padrões definidos da *web*.

O W3C (2015) define *linked data* como a coleção de conjuntos de dados inter-relacionados na *web*. Porém, há também o entendimento de que o termo resume um conjunto de melhores práticas para a publicação de dados conectados e estruturados na *web* por meio do uso de tecnologias-chave que o suportam (BERNERS-LEE, 2006; BIZER, HEATH, BERNERS-LEE, 2009; HEATH, BIZER, 2011; BAKER et al., 2011; SANTAREM SEGUNDO, CONEGLIAN, 2016). Berners-Lee (2006, não paginado, tradução nossa) afirma que "o *linked data* é essencial para, de fato, conectar a Web Semântica». Ou seja, no contexto da *Web* de Dados, o grande potencializador das conexões semânticas é o *linked data*.

A publicação de dados na *web* de acordo com os princípios de dados abertos e de dados conectados dá origem ao chamado *Linked Open Data - LOD* ou "dados abertos conectados":

> O LOD, que atualmente apresenta-se como a melhor forma de materialização dos conceitos e tecnologias da Web Semântica, é um projeto, com um conjunto de normas a serem seguidas, que usa os mesmos princípios de ligação semântica da Web de Dados, entretanto tem particularidades específicas, indicando um grau de exigência maior na constituição de sua rede de interligações. (SANTAREM SEGUNDO, 2015, p. 225).

Nem todos os conjuntos de dados conectados (*linked data*) são constituídos de dados abertos (*open data*), portanto, não constituem LOD. Baker et al. (2011, não paginado, tradução nossa) afirmam que "enquanto 'Linked Data' refere-se à interoperabilidade técnica dos dados, 'Open Data' foca na interoperabilidade legal", ou seja, dados que podem ser livremente usados, reutilizados e redistribuídos, conforme citado anteriormente. Os autores afirmam, porém, que "[...] o potencial da tecnologia é melhor percebido quando os dados são publicados como Linked Open Data" (BAKER et al., 2011, não paginado, tradução nossa).

Quando os dados são publicados na web em *linked open data* com todos os requisitos do modelo atendidos e ainda conectados a outros conjuntos de LOD considera-se que esses dados possuem padrão 5 estrelas de dados abertos. A classificação, que foi proposta por Berners-Lee em 2006, vai de 1 a 5 estrelas e considera:

- 1 estrela: disponível na Internet (em qualquer formato, como um PDF – *Portable Document Format*), desde que com licença

aberta, para que seja considerado dado aberto;
- 2 estrelas: disponível na Internet de maneira estruturada (como por exemplo em planilhas de arquivo Excel);
- 3 estrelas: disponível na Internet de maneira estruturada e em formato não proprietário (por exemplo, CSV – *Comma-separated values*, em vez de Excel);
- 4 estrelas: todas as regras anteriores, mas dentro dos padrões estabelecidos pelo W3C (RDF e SPARQL – *SPARQL Protocol and RDF Query Language*) para identificar coisas e propriedades, possibilitando que as pessoas possam relacionar seus recursos;
- 5 estrelas: todas as regras anteriores, além de conectar os dados a outros conjuntos de dados, de forma a fornecer contexto.

A estruturação dos vários conjuntos de dados com base nos padrões do *linked data* e LOD estão sendo mapeados pelo projeto *Linking Open Data*, da W3C. O objetivo do projeto é publicar os vários *datasets* abertos em RDF e estabelecendo links RDF entre itens de dados de diferentes fontes de dados. O projeto é publicado na chamada *LOD-Cloud* (http://lod-cloud.net/) o que dá visibilidade aos conjuntos de dados estruturados em RDF no padrão LOD, por serem disponibilizados para reuso de forma livre e gratuita.

A comunidade do *Linking Open Data* afirma que "para demonstrar o valor da Web Semântica é essencial ter mais dados do mundo real online" e que "o RDF é a tecnologia óbvia para interligar os dados das várias fontes". (W3C, 2017, não paginado, tradução nossa).

3 RDF

Como abordado na seção 2, o RDF é uma tecnologia essencial para estruturar os dados em *linked data*. A sigla significa *Resource Description Framework*, ou seja, estrutura para descrição de recursos. Segundo o W3C, o RDF é uma estrutura ou modelo de dados para descrição de informações na Web. (CYGANYAK, WOOD, LANTHALER, 2014). Segundo Isotani e Bittencourt (2015): "O RDF (*Resource Description Framework*) equivale a uma linguagem de representação de informação na Web, permitindo que recursos possam ser descritos formalmente e sejam acessíveis por máquinas." O W3C recomenda seu uso para estruturação de *linked data* pois facilita o processo de interoperabilidade entre os recursos. Ferreira e Santos (2013) afirmam que a criação do RDF nos anos 1990 foi influenciada por várias linguagens, e vocabulários de áreas do conhecimento, como o *Dublin Core* e o XML – *eXtensible Markup Language* e as comunidades de bibliotecas digitais (metadados) e de representação do conhecimento (ontologias). O RDF é formado pelas chamadas "triplas", estruturas compostas por sujeito, predicado e objeto também referidas como recurso, propriedade e valor:

> RDF faz o que o nome sugere, isto é, fornece uma estrutura para descrever recursos utilizando um esquema simples para expressar fatos ou declarações. A ideia por trás do RDF é clara, todo conceito seria representado por uma tripla composta por sujeito, propriedade (ou predicado) e objeto. De fato, essa combinação é familiar a todo nativo de línguas ocidentais, pois é a forma intuitiva pela qual se constroem frases simples. O sujeito refere-se ao conceito que se quer descrever; a propriedade, aos atributos relacionados ao sujeito; o objeto é algo a que se refere com a propriedade. Utilizando essa simples ideia, pode-se descrever qualquer coisa. (SCHIESSL, 2015, p. 56)

Essa estrutura permite que as máquinas consigam "compreender" o significado dos recursos na Web, tornando-a semântica. Os recursos devem ser sempre representados por URI – *Uniform Resource Identifier*, conforme regra do *linked data* definida por Berners-Lee em 2006. Já os objetos ou valores podem ser descrito por URI ou pelo próprio valor do dado, que são chamados de literais (exemplo: "2017"). Essas triplas de RDF possuem representação visual em forma de grafos, estrutura matemática baseada em vértices e arestas. Os grafos RDF formados pelas triplas RDF, porém, não são legíveis por máquina, o que torna necessário o uso de notações para representar o RDF de forma legível ao computador (SCHIESSL, 2015). Entre as notações mais utilizadas para descrever recursos em RDF destacam-se: RDF/XML *Syntax Specification*, RDFa (*Resource Description Framework in Attributes*), JSON-LD (*JavaScript*

Object Notation for Linked Data), *Turtle* e *N-triple*. As notações são também chamadas de formato de serialização de RDF.

Na descrição de recursos utilizando alguma das notações, há outro elemento que deve ser inserido para que a semântica seja compreendida pelos computadores, que são as linguagens. Para isso, recomenda-se a utilização do RDF-S, que é descrito pela W3C como uma extensão semântica do RDF. O RDF-S, ou RDF *Schema*, é uma linguagem que define o vocabulário utilizado no RDF. (SCHIESSL, 2015). Isotani e Bittencourt (2015, não paginado) afirmam que "O RDF-S é um vocabulário para modelagem de dados que amplia a expressividade do RDF para prover mecanismos de descrição de taxonomias entre recursos e suas propriedades".

Outros vocabulários e padrões que podem ser utilizados para inserir semântica nos relacionamentos das tripas são: *Dublin Core* (DC), *Friend of a Friend* (FOAF), *Simple Knowledge Organization System* (SKOS), *Ontology Web Language* (OWL), entre outros. Santarem Segundo (2015, p. 228) afirma que:

> [...] há um grupo de vocabulários que são utilizados em larga escala nas principais ontologias conhecidas e também em grande parte dos exemplos de publicação de datasets em formato semântico disponíveis na Web de Dados, por força de um reconhecimento imediato e global do significado que se pretende dar a ligação construída. Os vocabulários RDF, FOAF, RDFS, DC, e OWL são os vocabulários mais utilizados pelos datasets.

Com essas características, o RDF permite descrever recursos na web em diferentes granularidades: é possível descrever a página (*webpage*) como um todo, sendo um único recurso, mas também pode-se optar pela descrição de seus elementos individualmente, no nível desejado. Assim, o recurso pode ser um dado, que passa a ser o objeto de descrição, aumentando o nível semântico.

4 METODOLOGIA

A presente pesquisa pode ser caracterizada como pesquisa bibliográfica e pesquisa aplicada, pois, a partir da teoria encontrada na literatura, partiu-se para aplicação no ambiente pesquisado. As subseções a seguir irão detalhar a metodologia de pesquisa aplicada realizada. A modelagem foi feita com base no *dataset* disponibilizado pela Câmara dos Deputados no mês de junho de 2017.

4.1 Diagnóstico atual dos dados

Atualmente os dados legislativos da Câmara dos Deputados estão disponíveis no portal de dados abertos da instituição: https://dadosabertos.camara.leg.br/. Lançado em junho de 2017, essa é uma nova versão dos conjuntos de dados da Câmara em formato aberto e conta com as seguintes coleções de dados:

- Deputados
- Proposições
- Partidos políticos
- Blocos partidários
- Legislaturas
- Despesas parlamentares
- Órgãos
- Eventos
- Referências

Os dados estão disponíveis nos formatos XML e JSON, e para despesas parlamentares e proposições, nos formatos XML, JSON, CSV, XLSL (*MS-Excel*) e ODS (*Open Office*).

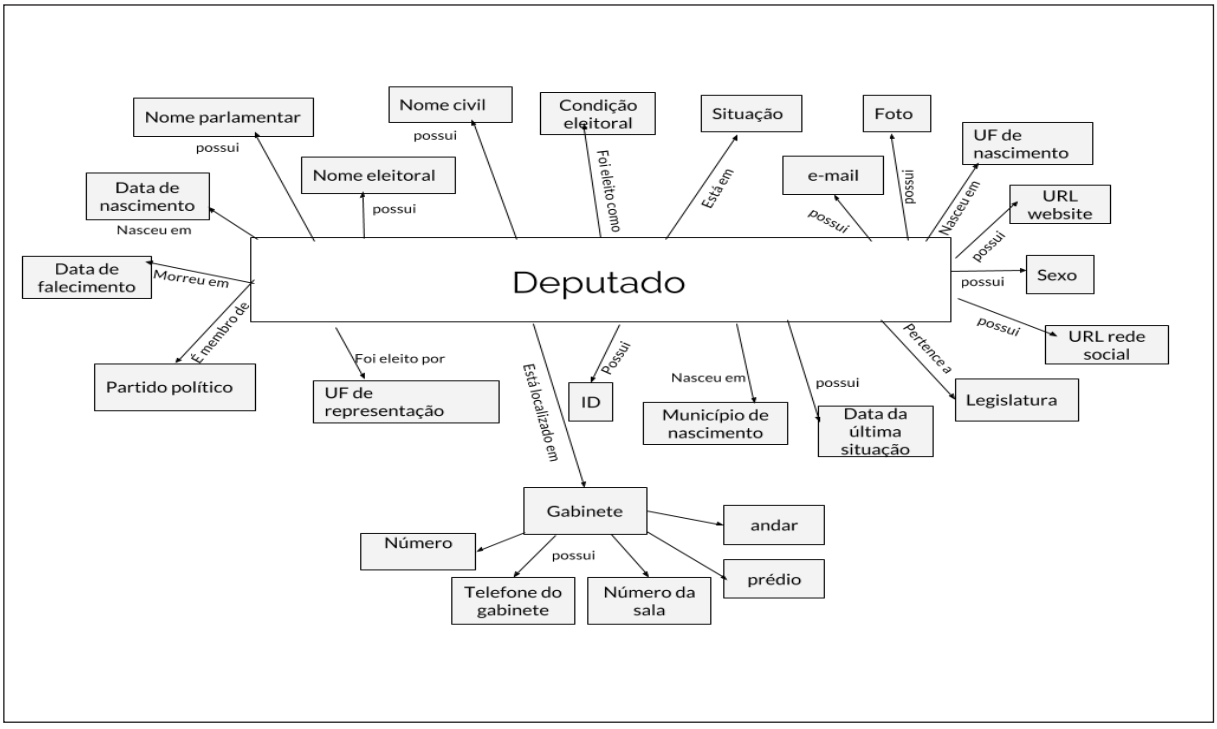
4.2 Seleção dos dados

A partir da análise dos dados disponíveis, decidiu-se optar por uma modelagem para o conjunto de dados "Deputados" e estudar os relacionamentos possíveis. A escolha desse conjunto foi feita porque os demais conjuntos de dados e relacionamentos entre eles pressupõem um estudo mais aprofundado do processo legislativo brasileiro para modelagem correta dos dados, o que não é escopo desta pesquisa. Além disso, por informação sobre parlamentares são consideradas importantes para os cidadãos que os elegeram. Ademais, por representar o agente principal do processo legislativo, esse conjunto é passível de ligação com todos os outros dados da instituição.

Os dados de deputados disponibilizados pela Câmara incluem as seguintes informações (metadados) sobre os parlamentares: ID (identificador) do deputado, URI do deputado,

Nome civil, Nome parlamentar, Sigla do partido, URI do partido, UF de representação, Número da legislatura, Foto do deputado, Data da última situação, Nome eleitoral, Número do gabinete, Prédio, Número da sala, Andar, Telefone do gabinete, e-mail do deputado, situação do deputado, condição eleitoral, CPF do deputado, Sexo do deputado, URL – *Uniform Resource Locator* do website do deputado, URL rede social, Data de nascimento, Data de falecimento, UF de nascimento, Município de nascimento e Escolaridade. Tais atributos são representados na figura 1, a seguir.

Figura 1: Relacionamentos entre os dados do conjunto "Deputados"



Fonte: elaborada pelos autores

A partir do modelo lógico do conjunto de dados "Deputados" apresentado na figura 1, identificaram-se mais dois conjuntos de dados disponibilizados pela Câmara que podem ser relacionados a eles: "partidos políticos" e "legislaturas". Dessa forma, as propriedades partido político e legislatura poderão ser representadas pelas URI dos recursos "Partidos" e "Legislaturas configurando modelo de dados abertos conectados.

4.3 Seleção dos vocabulários

Catarino e Souza (2012) afirmam que, no contexto da Web Semântica, as estruturas formadas por termos que representam conceitos, relacionamentos entre os termos e suas limitações de uso, são chamadas de vocabulários ou de ontologias. O W3C (2011) entende como 'vocabulário' as coleções de termos mais simples e 'ontologia' para as mais complexas, ou seja, estruturadas com maior grau de formalidade.

Conforme é possível observar na representação da figura 1, várias das propriedades dos recursos "Deputado" estão relacionadas com o predicado "possui", o que insere uma semântica básica, porém não tão expressiva em termos de representação do conhecimento. Dessa forma, recomenda-se o uso de ontologias por serem ferramentas mais robustas e com estruturação semântica mais avançada. Conforme afirma Santarem Segundo

(2015, p. 226), o uso de ontologias é "uma das maneiras de se construir uma relação organizada entre termos dentro de um domínio, favorecendo a possibilidade de contextualizar os dados, tornando mais eficiente e facilitando o processo de interpretação dos dados pelas ferramentas de recuperação da informação". Dessa forma, recomenda-se descrever os recursos com uso de vocabulários de representação de conhecimento padronizados e reconhecidos internacionalmente, o que fornece enriquecimento semântico dos dados e melhora a recuperação da informação. Entre as recomendações para publicação de dados na web da W3C, editadas por Lóscio, Burle e Calegari (2017), a boa prática n. 15 é o reuso de vocabulários, de preferência os padronizados, para descrever dados e metadados, pois captura e facilita o consenso nas comunidades, aumenta a interoperabilidade e reduz redundâncias, além de auxiliar a comparação e processamento automático de dados e metadados e também a evitar ambiguidade entre elementos similares. Esta prática traz os benefícios de interoperabilidade, processamento, reuso, compreensão e confiança.

O critério de seleção dos vocabulários foi a partir da avaliação dos mais utilizados e reconhecidos entre os *Linked Open Vocabularies*[1] (LOV), e, entre eles, os que possuem relacionamentos que atendem ao domínio dos dados trabalhados, conforme orientação de Heath e Bizer (2011, não paginado, tradução nossa):

> Se termos adequados forem encontrados em vocabulários existentes, estes devem ser reutilizados para descrever dados sempre que possível, ao invés de reinventar. O reuso de termos existentes é altamente desejável, pois maximiza a probabilidade de que os dados sejam consumidos por aplicações que podem ser configuradas para vocabulários conhecidos, sem necessidade de pré-processamento de dados ou modificações da aplicação.

Segundo os autores, é possível também o uso de mais de um vocabulário para descrever a mesma propriedade de um recurso quando a propriedade se trata de uma especificação de termos existentes em outro vocabulário, o que introduz um elemento de redundância que também contribui para a acessibilidade dos dados por aplicações de *linked data* que não utilizam mecanismo de inferência. Ou seja, o uso de mais vocabulários é uma estratégia para melhoria da recuperação da informação. Heath e Bizer (2011) indicam ainda quatro critérios para a escolha dos vocabulários: uso e aceitação, governança e manutenção, cobertura e expressividade, os quais foram considerados na seleção de vocabulários para este trabalho.

O primeiro passo para a seleção de vocabulários foi realizar uma busca no LOV pelos termos relacionados ao domínio trabalhado (poder legislativo, parlamento, política e temas correlatos). O LOV permite buscas por vocabulários, propriedades, classes e agentes, ou busca em todos os anteriores. A busca foi feita em todos os campos para aumentar a revocação. Foram utilizados os termos: Legislação, *Legislation*, Parlamento, *Parliament*, Política, *Politics*, Lei, *Law,* Congresso e *Congress*. Entre todas as buscas realizadas, somente 2 vocabulários foram encontrados. Porém, foram encontrados vários registros para classes e propriedades de ontologias, que levaram a ontologias relacionadas ao assunto buscado. Além disso, foram analisados os 11 vocabulários organizados pela *tag* '*Government*' no LOV. Com a análise dos resultados, foram encontrados os seguintes vocabulários relacionados, de alguma forma, ao conjunto de dados do estudo:

- *Ontologia Camera dei Deputati*
- *The European Legislation Identifier*
- *Parliament Ontology*
- Parlamento britânico - várias ontologias
- *Vocabulario de Resultados Electorales*
- *Ontologie de l'Assemblée Nationale*
- *Vocabulario de la estructura de organismos públicos*

Observou-se também que a Ontologia da DBpedia (http://dbpedia.org/ontology/) possui classes relacionadas a agentes políticos e estruturas de governo, as quais poderiam ser aproveitadas. A partir daí esses vocabulários foram analisados para buscar as propriedades existentes no conjunto de dados modelados.

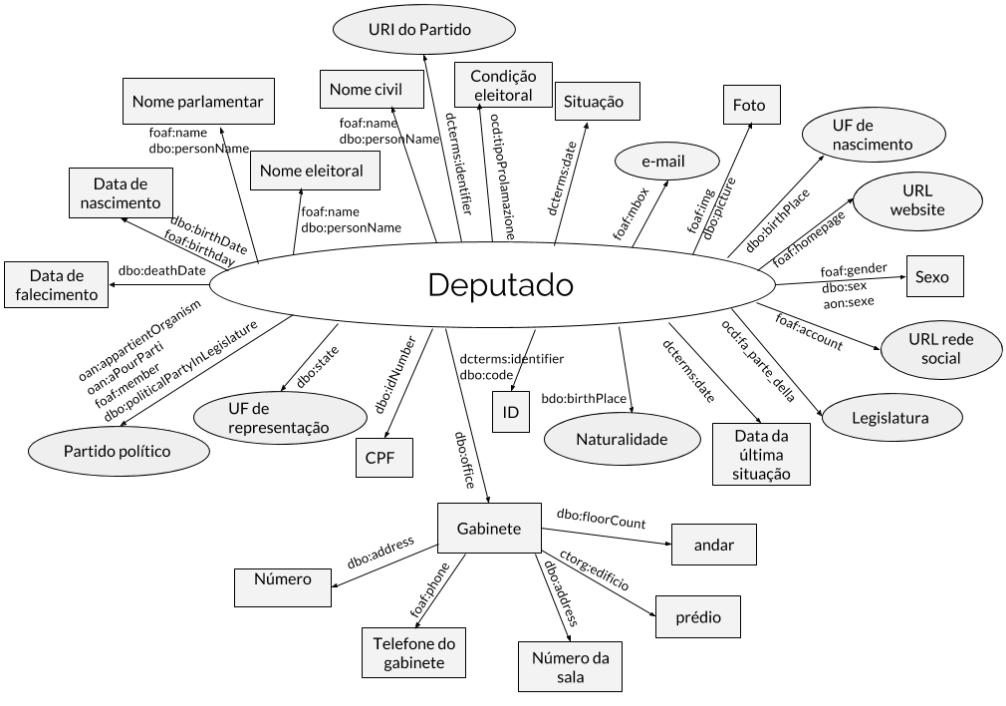
Além disso, entre os vocabulários mais utilizados (DC, RDFS, FOAF), foram identificadas classes e propriedades que possam ser utilizadas para descrever os dados do conjunto

1 http://lov.okfn.org/dataset/lov/

deste conjunto – acredita-se que a utilização desses vocabulários gere maior integração do conjunto de dados à *Web* de Dados e melhore a recuperação da informação semântica, já que esses são vocabulários bastante utilizados e mais familiares aos motores de busca. Como o domínio trabalhado refere-se a agentes políticos, que é uma especificidade da classe "pessoas", optou-se por analisar as propriedades pelo vocabulário FOAF por estar voltado para o domínio "pessoas". O FOAF é um dos mais utilizados e reconhecidos vocabulários na Web de Dados, e este critério foi considerado para sua escolha. A partir disso, os relacionamentos não encontrados no FOAF foram buscados em outros vocabulários reconhecidos e recomendados pelo W3C: Dublin Core, RDF, RDFS, além da ontologia da DBpedia. Os termos úteis para descrever relações do conjunto de dados encontrados nesses vocabulários também foram utilizados, mesmo que já descritos por outros vocabulários, para agregar semântica.

Além dos vocabulários para propriedades, os próprios valores ou objetos podem ser representados por outros conjuntos de dados, os quais possuem mais elementos e acrescentam ainda mais semântica dos recursos, conectando-os a novos dados contextualizados. Conforme definido por Berners-Lee, o quarto princípio do *Linked Data* é incluir *links* para outras URI, e segundo Heath e Bizer (2011), *links* externos em RDF são fundamentais para a *Web* de Dados, pois conectam dados isolados num espaço global interconectado, além de possibilitar que as aplicações descubram novas fontes de dados. Com base nisso, observou-se também a possibilidade de relacionar outros conjuntos de dados conectados na modelagem proposta. Foram utilizadas URI dos conjuntos de dados abertos de "Partidos" e "Legislaturas", também fornecido pela Câmara dos Deputados, para os valores referentes a partidos políticos e legislatura, e, para os outros, foram selecionados LOD que pudessem representar os dados descritos. Com isso, o modelo lógico apresentado na figura 1 pode ser refeito, utilizando os vocabulários selecionados para as propriedade e valores, conforme ilustra a figura 2:

Figura 2: Conjunto de dados "Deputados" e vocabulários selecionados



Fonte: elaborada pelos autores

Na figura 2, os relacionamentos genéricos foram substituídos por vocabulários que indicam a semântica do relacionamento entre o recurso e seus objetos, e foram adotadas as representações gráficas de retângulo para os literais e as elipses para os recursos que possuem URI.

5 RESULTADOS E DISCUSSÃO

A partir da análise dos conjuntos de dados, relacionamentos e valores dos modelos das figuras 1 e 2, foi elaborado, para fins de sistematização, o quadro 1, a seguir, com os resultados obtidos. Com isso, chegou-se à seguinte proposta de uso dos vocabulários e LOD:

Quadro 1: Conexões possíveis para o conjunto de dados abertos "Deputados"

Recurso	Propriedade (vocabulários)	Valor	LOD / URI
Deputado https://dadosabertos.camara.leg.br/api/v2/deputados/	dcterms:identifier	ID do deputado	N/A
	bdo:code		
	foaf:name	Nome civil	N/A
	dbo:personName		
	foaf:name	Nome parlamentar	N/A
	dbo:personName		
	oan:appartientOrganisme	Partido político (sigla)	http://dbpedia.org/resource/
	oan:aPourParti		
	oan:appartientOrganisme		
	foaf:member		
	dbo:politicalPartyInLegislature		
	dbo:state	UF de representação	http://www.geonames.org/
	dcterms:identifier	URI do partido	https://dadosabertos.camara.leg.br/api/v2/partidos
	ocd:fa_parte_della	Legislatura	https://dadosabertos.camara.leg.br/api/v2/legislaturas
	foaf:img	Foto	N/A
	dbo:picture		
	dcterms:date	Data da **última** situação	N/A
	foaf:name	Nome eleitoral	N/A
	dbo:personName		
	dbo:office	Gabinete	N/A
	dbo:address	Número do gabinete	N/A
	ctorg:edificio	Prédio	N/A
	dbo:address	Nome da sala ou prédio	N/A
	dbo:floorCount	Andar	N/A
	foaf:phone	Telefone do gabinete	N/A
	foaf:mbox	email	N/A
	dbo:status	situação	N/A
	OCD: tipoProclamazione	condição eleitoral	N/A
	dbo:idNumber	CPF	N/A
	foaf:gender	Sexo	N/A
	dbo:sex		
	oan:sexe		
	foaf:homepage	URL do website	N/A
	foaf:account	URL rede social	N/A
	foaf:bithday	Data de nascimento	N/A
	dbo:birthDate		
	dbo:deathDate	Data de falecimento	N/A
	dbo:birthPlace	UF de nascimento	http://www.geonames.org/
	dbo:birthPlace	Naturalidade	http://www.geonames.org/
	dcterms:educationLevel	Escolaridade	N/A

Legenda:
dcterms: Dublin Core Terms http://purl.org/dc/terms/; dbo: DBpedia Ontology http://dbpedia.org/ontology/; foaf: Friend of a Friend http://xmlns.com/foaf/0.1/; oan: Ontologie de l'Assemblée Nationale http://data.lirmm.fr/ontologies/oan; ocd: Ontologia Camera dei Deputati http://dati.camera.it/ocd/reference_document/; ctorg: Vocabulario de la estructura de organismos públicos http://purl.org/ctic/infraestructuras/organizacion; elec: Vocabulario de Resultados Electorales ; N/A: não se aplica

Fonte: elaborado pelos autores

A elaboração do quadro 1 permitiu estruturar toda a informação da pesquisa: recurso, relacionamentos, propriedades, vocabulários e ontologias e ligações possíveis com conjuntos de dados externos. A partir dele, foi realizada a modelagem dos dados em RDF, descrita na seção a seguir. Foram inseridas também as estruturas RDF (rdf:value) e RDFS (rdfs:label) para identificar, respectivamente, literais e rótulos, as quais não constam no quadro 1 por estarem presentes em todas as propriedades do recurso, não havendo necessidade de repetição em todas as linhas da tabela.

5.5 Modelagem RDF

Feitas as etapas anteriores da metodologia proposta, chegou-se a um modelo de dados abertos conectados para os dados legislativos do conjunto "Deputados" fornecido pela Câmara dos Deputados. Para modelagem em RDF, é necessário o uso de uma notação, ou formato de serialização. Devido a facilidade de modelagem do formato *Turtle*, este foi o escolhido nesta proposta. Ademais, há ferramentas de conversão que permitem transformar a modelagem em *Turtle* para outras notações. Dessa forma, foi obtida a modelagem RDF em *Turtle* para o conjunto de dados abertos "Deputados". Um pequeno trecho dos dados escritos em RDF é apresentado na figura 3 a seguir:

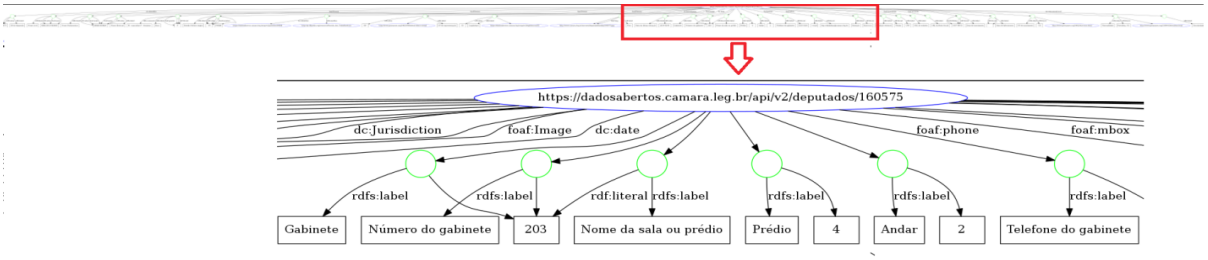
Figura 3: Modelagem RDF em *Turtle*

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix dcterms: <http://purl.org/dc/terms/>.
@prefix dbo: <http://dbpedia.org/ontology/>.
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix oan: <http://data.lirmm.fr/ontologies/oan>.
@prefix ocd: <http://dati.camera.it/ocd/reference_document/>.
@prefix ctorg: <http://purl.org/ctic/infraestructuras/organizacion>.
@prefix elec: <http://purl.org/ctic/sector-publico/elecciones#>.
@prefix dbr: <http://dbpedia.org/resource/>.

<https://dadosabertos.camara.leg.br/api/v2/deputados/160575>

dcterms:identifier [ rdfs:label "ID do deputado"; dbo:code "160575";
rdf:value "160575" ];

foaf:Person [ foaf:name "ÉRIKA JUCÁ KOKAY"; dbo:personName: "ÉRIKA
JUCÁ KOKAY"; rdf:value "ÉRIKA JUCÁ KOKAY"; rdf:label "Nome civil" ];

foaf:Person [ foaf:name "ERIKA KOKAY"; dbo:personName: "ERIKA KOKAY";
rdf:value "ERIKA KOKAY"; rdfs:label "Nome Parlamentar" ];

dbo:PoliticalParty [ rdf:value "PT"; oan:appartientOrganisme "PT";
oan:aPourParti "PT"; oan:appartientOrganisme "PT"; foaf:member "PT";
dbo:politicalPartyInLegislature "PT"; rdfs:label "Partido Político
(sigla)" ];

dcterms:Jurisdiction [ rdf:value "DF"; dbo:state
<http://www.geonames.org/3463504/federal-district.html>; rdfs:label
"UF de representação" ];

dbo:PoliticalParty [ dcterms:identifier
<https://dadosabertos.camara.leg.br/api/v2/partidos/36844>;
rdfs:seeAlso <http://pt.dbpedia.org/resource/Partido_dos_Trabalhadores> ];

dbo:Legislature [ rdf:value "55"; rdfs:label "Legislatura";
ocd:fa_parte_della
<https://dadosabertos.camara.leg.br/api/v2/legislaturas/55> ];
```

Fonte: elaborada pelos autores

A modelagem completa está disponível no Apêndice I deste artigo. O grafo RDF gerado pela modelagem é representado na figura 4:

Figura 4: Grafo RDF do conjunto de dados "Deputados"

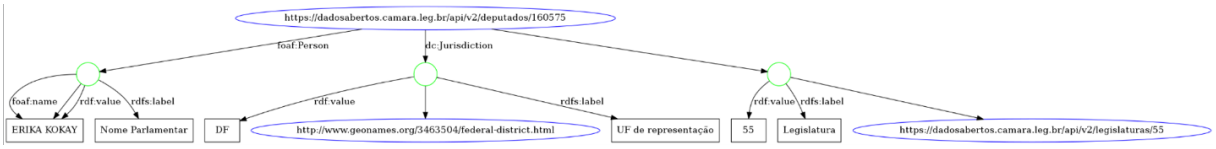


Fonte: elaborada pelos autores com a ferramenta http://www.easyrdf.org/converter

Como ilustra a figura 4, não é possível visualizar a imagem inteira do grafo devido a grande quantidade de triplas formadas. Assim, foi gerada uma imagem propositalmente ilegível do grafo completo – apenas para demonstrar sua dimensão – e um recorte ampliado que possibilita visualizar, de forma um pouco mais legível, uma pequena parte do grafo.

A seguir, outro recorte do grafo gerado pelas triplas é apresentado na figura 5 para destacar o uso de valores literais, como "DF", o uso de URI da própria instituição, referente a outro conjunto de dados (Legislaturas) e uso de LOD externo para o recurso "Distrito Federal", da *Geonames*:

Figura 5: exemplos de triplas geradas do conjunto de dados "Deputados"



Fonte: elaborada pelos autores com a ferramenta http://www.easyrdf.org/converter

Como este conjunto de dados foi disponibilizado pela instituição nos formatos JSON e XML, esta modelagem em Turtle pode ser convertida para JSON-LD e RDF/XML para real aplicação no conjunto de dados da Câmara dos Deputados e disponibilização de *Linked Open Data* Legislativo. Outra possibilidade é que a instituição forneça a estrutura base para LOD desses dados e o usuário a utilize em suas aplicações, alimentando-as com o *dataset* "Deputados" já disponibilizado.

6 CONSIDERAÇÕES FINAIS

A disponibilização de dados abertos governamentais é hoje uma realidade a ser buscada por todas as instituições públicas, tanto por motivação legal quanto por responsabilidade social. As tecnologias da Web Semântica aliadas a técnicas e conceitos de organização da informação e representação do conhecimento possibilitam a estruturação de dados e informações na *web* de forma contextualizada, precisa e rica de significados, resultando em ganhos tanto no processamento pelas máquinas quanto, e, consequentemente, para o usuário final da informação.

Acredita-se que a presente proposta de modelagem de dados abertos legislativos em RDF possa demonstrar para a instituição a viabilidade de fornecimento de dados em padrão 5 estrelas, assim como para a sociedade, a ampliação das possibilidades e o ganho de qualidade nas implementações de conjuntos de dados quando estes estão disponíveis em *linked open data*. Destaca-se a importância do modelo estrutural gerado nesta pesquisa para a melhoria da qualidade dos dados abertos governamentais e para a promoção da transparência pública. Considera-se importante ainda ressaltar a necessidade de os órgãos governamentais publicarem seus dados de acordo com métodos e técnicas mais recentes e recomendadas pelas comunidades de prática, e em especial pela instituição W3C.

Destaca-se que o modelo elaborado nesta pesquisa foi feito com base no conjunto real de dados fornecidos pela Câmara dos Deputados, de modo que pode ser transformado em uma espécie de folha de estilo ou gabarito para conversão de todo o conjunto de dados "Deputados" da instituição. Como forma de melhoria, sugere-se a criação de prefixos com os partidos políticos, UF e outros dados normalizados. Desta forma, haveria um ganho de precisão devido à especificidade do conjunto de dados. Como exemplo, a granularidade "nome" não é suficiente quando há existência de conceitos próprios como "nome parlamentar" e "nome eleitoral". Ambos foram descritos como o genérico "foaf:-name", o que caracteriza o objeto e fornece um nível de semântica, mas não na granularidade necessária para que seja diferenciado de forma unívoca dos demais objetos do recurso. O desenvolvimento de uma ontologia específica para dados legislativos brasileiros, com reuso de ontologias já existentes e criação de elementos particulares próprios, possibilitaria um nível semântico mais acurado.

Verificou-se, pela metodologia estabelecida para a criação do modelo estrutural do conjunto de dados "Deputados", a contribuição fundamental da Ciência da Informação, em especial das já citadas áreas de representação da informação e do conhecimento, com o uso de vocabulários e ontologias para representar e trazer significado aos relacionamentos entre as entidades modeladas. Assim, corrobora-se a ideia de que a Ciência da Informação é uma área do conhecimento fundamental para a implementação efetiva da Web Semântica.

Recomenda-se como estudos futuros a realização de diagnóstico do processo de publicação de dados das instituições públicas para alinhamento com as melhores práticas para publicação de dados na *web* de acordo com a W3C, como o uso de metadados, declaração das licenças de uso, informações de versionamento, proveniência dos dados, entre outros.

Artigo recebido em 18/01/2018 e aceito para publicação em 21/03/2018

LEGISLATIVE LINKED OPEN DATA:
a proposal of linked open data modelling for legislative information

ABSTRACT *This research aimed to create a linked open data (LOD) model using RDF (Resource Description Framework) for a legislative dataset from the House of Representatives. It has proceeded a literature review embracing the concepts of open data, open government data, linked data and linked open data followed by an applied research for a data modelling of legislative data in linked open data using RDF. The dataset "Deputies" was selected for this research. It covers information about political parties, state, e-mail, legislature among others, about the congressmen. It was observed that modelling the dataset using RDF is possible by reusing vocabularies and frameworks already established in Semantic Web e.g. Dublin Core, Friend of a Friend (FOAF), RDF and RDF Schema, together with correlated vocabularies like the Italian House of Representatives and the French National Assembly ontologies. As recommended by Linked Data best practices, the resources were also related to other LOD such as Geonames and DBpedia for semantic enhancement. The study allows us to conclude that making government data available, especially legislative data can be done according to the recommendations and best practices from W3C (World Wide Web Consortium). This practice integrates legislative data on the Web of Data and increase the possibilities of data reuse in transparency and oversight projects, which approaches citizens to the Congress and to their representatives.*

Keywords: *Linked Open Data. Linked Data. Semantic Web. Open Government. Legislative Data.*

REFERÊNCIAS

BAKER, T. et al. **Library Linked Data Incubator Group Final Report**. W3C Incubator Group Report, 2011. Disponível em: <http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/>. Acesso em: 10 jun. 2017.

BERNERS-LEE, T. **Linked Data**: Design Issues. 2006. Disponível em: <http://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 07 jun. 2017.

BERNERS-LEE, T. **Putting Government Data online**. 2009. Disponível em: <https://www.w3.org/DesignIssues/GovData.html>. Acesso em: 19 jun. 2017.

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: 12 mar. 2018.

CATARINO, M. E.; SOUZA, T. B. A representação descritiva no contexto da web semântica. **Transinformação**, v. 24, n. 2, p. 77-90, 2012. Disponível em: < http://dx.doi.org/10.1590/S0103-37862012000200001>. Acesso em: 12 mar. 2018.

CYGANYAK, R.; WOOD, D.; LANTHALER, M. (Eds). **RDF 1.1 Concepts and Abstract Syntax.** W3C, 2014. Disponível em: <https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/>. Acesso em: 08 jun 2017.

FERREIRA, J. A.; SANTOS, P. L. V. A. da C. O modelo de dados Resource Description Framework (RDF) e o seu papel na descrição de recursos. **Informação & Sociedade**: Est., João Pessoa, v. 23, n. 2, p. 13- 23, maio/ago. 2013. Disponível em: < https://repositorio.unesp.br/handle/11449/10557>. Acesso em: 11 jun. 2017.

LÓSCIO, B. F; BURLE, C.; CALEGARI, N. (Eds.). **Data on the Web Best Practices**. 2017. Disponível em: <https://www.w3.org/TR/dwbp>. Acesso em: 20 jun. 2017.

HEATH, T.; BIZER, C. **Linked Data**: Evolving the Web into a Global Data Space (1st edition). EUA: Morgan & Claypool, 2011. Disponível em: <http://linkeddatabook.com/editions/1.0/>. Acesso em: 09 jun. 2017.

ISOTANI, S.; BITTENCOURT, I. **Dados abertos conectados**: em busca da web do conhecimento. São Paulo: Novatec, 2015. Disponível em: <http://ceweb.br/livros/dados-abertos-conectados/>. Acesso em: 09 jun. 2017.

SANTAREM SEGUNDO, J. E. Web Semântica, dados ligados e dados abertos: uma visão dos desafios do Brasil frente as iniciativas internacionais. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v.8, p.219 – 239, 2015. Disponível em: <http://inseer.ibict.br/ancib/index.php/tpbci/article/view/207>. Acesso em: 06 jun. 2017.

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, p. 217-244, 2016. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/26417>. Acesso em: 07 jun. 2017.

SCHIESSL, M. **Lexicalização de Ontologias**: o relacionamento entre conteúdo e significado no contexto da Recuperação da Informação. Brasília, 2015. 261 p. Tese (Doutorado – Doutorado em Ciência da Informação) — Universidade de Brasília, 2015.

W3C. **Manual dos Dados Abertos**: Governo traduzido e adaptado do opendatamanual.org. 2011. Disponível em: <http://www.w3c.br/pub/Materiais/PublicacoesW3C/Manual_Dados_Abertos_WEB.pdf>. Acesso em: 06 jun. 2017.

W3C. **Web Semântica**. 2011. Disponível em: <http://www.w3c.br/Padroes/WebSemantica>. Acesso em: 07 jun 2017.

W3C. **Linked Data**. 2015. Disponível em: <https://www.w3.org/standards/semanticweb/data>. Acesso em: 07 jun 2017.

W3C. **Linking Open Data Project**. 2017. Disponível em: <https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData> Acesso em: 08 jun 2017.

Agradecimentos: Fabricio Rocha de Sousa, gerente do projeto de Dados Abertos da Câmara dos Deputados pelas informações prestadas, essenciais para a elaboração deste artigo.