

RUMO À PUBLICAÇÃO DE DADOS ABERTOS GOVERNAMENTAIS COMO DADOS
ABERTOS LIGADOS UTILIZANDO UMA ABORDAGEM BASEADA EM ONTOLOGIAS

Aíquis Rodrigues Gomes

Dissertação apresentada ao Programa de
Pós-Graduação em Ciência da Computação,
do Centro Federal de Educação Tecnológica
Celso Suckow da Fonseca, CEFET/RJ, como
parte dos requisitos necessários à obtenção
do título de mestre.

Orientador(a): Kele Teixeira Belloze

Rio de Janeiro,
Agosto 2020

RUMO À PUBLICAÇÃO DE DADOS ABERTOS GOVERNAMENTAIS COMO DADOS
ABERTOS LIGADOS UTILIZANDO UMA ABORDAGEM BASEADA EM ONTOLOGIAS

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação,
do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como
parte dos requisitos necessários à obtenção do título de mestre.

Aíquis Rodrigues Gomes

Banca Examinadora:



Presidente, Professora D.Sc. Kele Teixeira Belloze (CEFET/RJ) (Orientador(a))



Professora D.Sc. Laura Silva de Assis (CEFET/RJ)



Professora D.Sc. Maria Claudia Reis Cavalcanti (IME)

Rio de Janeiro, 12 de agosto de 2020

Agosto 2020

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

G633 Gomes, Aíquis Rodrigues
Rumo à publicação de dados abertos governamentais como
dados abertos ligados utilizando uma abordagem baseada em
ontologias / Aíquis Rodrigues Gomes — 2020.
108f : il. (algumas color.) , enc.

Dissertação (Mestrado) Centro Federal de Educação
Tecnológica Celso Suckow da Fonseca , 2020.
Bibliografia : f. 103-108
Orientadora: Kele Teixeira Belloze

1. Intercâmbio eletrônico de dados. 2. Processamento eletrônico
de dados. 3. Informações governamentais. 4. Sistemas de
transmissão de dados. 5. Ontologia. I. Belloze, Kele Teixeira
(Orient.). II. Título.

CDD 005.72

AGRADECIMENTOS

Ao CEFET/RJ, seu corpo docente e funcionários que ofereceram ambiente, estrutura e conhecimento necessário para o meu crescimento acadêmico.

À professora Kele Belloze, pela orientação brilhante, estando sempre disponível, sendo sempre solícita e sendo uma grande parceira nessa jornada do mestrado.

Aos meus pais, Clarice e José Roberto, por todo o apoio durante o percurso do mestrado e por toda a base que me permitiu chegar até aqui.

À Halanna, que sempre esteve ao meu lado me apoiando e me incentivando a continuar, mesmo nos momentos que eu mais pensei em desistir. Sem o apoio e compreensão dela durante o curso certamente eu não teria chegado ao fim.

Aos meus amigos, especialmente Eduardo e Cesar, que também sempre trouxeram doses de motivação nos momentos duros dessa caminhada.

O presente trabalho foi desenvolvido com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

RESUMO

Rumo à Publicação de Dados Abertos Governamentais como Dados Abertos Ligados utilizando uma abordagem baseada em Ontologias

Governos são grandes produtores e publicadores de dados e têm buscado utilizá-los como forma de aumentar a transparência e gerar mais valor para a sociedade. No entanto, o nível de maturidade na publicação de dados governamentais ainda é baixo, o que resulta na publicação utilizando formatos que dificultam a conexão com outros conjuntos de dados e a leitura por parte de máquinas, além de publicações que não são realmente em formatos abertos. Dados Abertos Ligados constituem um conjunto de tecnologias e padrões da web semântica que permitem a ligação entre diferentes conjuntos de dados abertos publicados na web. Por intermédio dos Dados Abertos Ligados, governos podem atingir alto grau de maturidade na publicação de dados utilizando um formato realmente aberto, que permite a leitura por máquinas e que pode potencializar o valor gerado para a sociedade com iniciativas de dados.

No entanto, existem algumas barreiras para a publicação dos dados utilizando essas tecnologias e padrões. Uma dessas barreiras refere-se a falta de um guia para sua implementação que possa direcionar, de maneira estruturada, os passos a serem seguidos para a publicação de um conjunto de dados como Dados Abertos Ligados. Nesse trabalho é apresentada uma metodologia baseada no uso de ontologia para que conjuntos de dados publicados em formatos tradicionais possam ser publicados como Dados Abertos Ligados. A metodologia é composta por quatro etapas: (i) identificação, análise e integração dos dados; (ii) desenvolvimento da ontologia; (iii) publicação dos dados como Dados Abertos Ligados; e (iv) publicação de *endpoint* SPARQL.

Dois experimentos utilizando conjuntos de dados governamentais reais dos domínios eleitoral e da saúde foram realizados seguindo a metodologia proposta. Como resultados houve a produção de duas ontologias, sobre as eleições brasileiras e sobre as Unidades Básicas de Saúde em funcionamento no Brasil, e a disponibilização dos dois conjuntos de dados referentes a essas ontologias em arquivos RDF com alguns dos recursos ligados a outros conjuntos de dados. Com os experimentos foi possível comprovar que por meio de um processo estruturado é possível evoluir na publicação de dados abertos e que os passos propostos podem ser aplicados independente do domínio dos dados.

Palavras-chave: Dados Abertos; Dados Abertos Governamentais; Dados Ligados; Dados Abertos Ligados; Ontologia.

ABSTRACT

Towards Publication of Open Government Data as Linked Open Data using an ontology-based approach

Governments are significant producers and publishers of data and have sought to use them to increase transparency and generate more value for society. However, the level of maturity in publishing government data is still low, which results in the publication using formats that make it difficult to connect to other data sets and to be read by machines, in addition to publications that are not really in open formats. Linked Open Data constitutes a set of technologies and standards of the semantic web that allow the connection between different sets of open data published on the web. Through Linked Open Data, governments can achieve a high degree of maturity in the publication of data using a truly open format, which allows machine-reading and can enhance the value generated for society with data initiatives.

However, there are some barriers to publishing data using these technologies and standards. One of these barriers is the lack of a guide for its implementation that can direct, in a structured way, the steps to be followed for the publication of a data set as Linked Open Data. This work presents an ontology-based methodology to publish traditional data sets as Linked Open Data. The methodology consists of four stages: (i) identification, analysis, and integration of data; (ii) ontology development; (iii) publication of the data as Linked Open Data; and (iv) publication of a SPARQL endpoint.

Two experiments using real government data sets from the electoral and health domains were carried out following the proposed methodology. As a result, there was the production of two ontologies, on the Brazilian elections and the Basic Health Units operating in Brazil, and the availability of the two data sets referring to these ontologies as RDF files with some resources linked to other datasets. With the experiments, it was possible to show that through a structured process, it is possible to evolve in the publication of open data and that the proposed steps can be applied regardless of the data domain.

Keywords: Open Data; Open Government Data; Linked Data; Linked Open Data; Ontology.

LISTA DE ILUSTRAÇÕES

Figura 1 –	Exemplo de relacionamento entre instâncias de classes em uma ontologia	20
Figura 2 –	Fluxo de trabalho proposto para a construção da ontologia utilizando etapas das metodologias NeOn e Ontology Development 101	23
Figura 3 –	Sub-nuvem da LOD <i>Cloud</i> referente a dados governamentais	34
Figura 4 –	Exemplo de URI com dados sobre o Dalai Lama mantido pela organização do Prêmio Nobel. Fonte: Nobel Prize Data	35
Figura 5 –	Entidades e relacionamentos do modelo ER das Eleições Brasileiras	62
Figura 6 –	Extrato do Documento de Especificações de Requisitos da Ontologia das Eleições Brasileiras	66
Figura 7 –	Classes e relacionamentos da ontologia das Eleições Brasileiras derivadas das Questões de Competência	69
Figura 8 –	Classes e propriedades de objetos e de dados da ontologia das Eleições Brasileiras na ferramenta Protégé	72
Figura 9 –	Exemplo de propriedades de objeto e de dados de um candidato. Nesse exemplo o candidato apresentado é Guilherme Boulos	74
Figura 10 –	Apresentação de como é feita a ligação com recursos externos à ontologia a nível de indivíduo. A imagem mostra que o candidato Guilherme Boulos, que é um indivíduo da ontologia, é o mesmo representado no recurso http://pt.dbpedia.org/resource/Guilherme_Boulos	75
Figura 11 –	Visualização da classe Eleicao em um navegador, acessada através de sua URI	77
Figura 12 –	Resultado da consulta SPARQL mostrando os cargos disputados para a eleição 295_ELEIÇÃOGERALFEDERAL2018	79

Figura 13 – Extrato do Documento de Especificação de Requisitos da Ontologia das UBS em Funcionamento no Brasil	84
Figura 14 – Exemplo de propriedades preenchidas de uma UBS na ontologia. Nesse exemplo são dados do Centro de Saúde Aurelio Texeira Cotta	88
Figura 15 – Representação RDF do Centro de Saúde Aurelio Texeira Cotta visualizado em um navegador	92
Figura 16 – Resultado da consulta SPARQL para saber qual o nome fantasia da UBS pesquisada	93

LISTA DE TABELAS

Tabela 1 –	Critérios de inclusão de artigos no mapeamento sistemático	40
Tabela 2 –	Critérios de exclusão de artigos do mapeamento sistemático	41
Tabela 3 –	Comparação entre os trabalhos relacionados a essa dissertação	47
Tabela 4 –	Comparação entre os trabalhos relacionados a essa dissertação (continuação)	48
Tabela 5 –	Questões de Competência sobre as eleições no Brasil	64
Tabela 6 –	Termos com maior frequência extraídos das Questões de Competência	65
Tabela 7 –	Vocabulários e recursos reutilizados ou ligados à ontologia das Eleições Brasileiras	67
Tabela 8 –	Relação de classes e instâncias da ontologia das Eleições Brasileiras de cenário hipotético	70
Tabela 9 –	Questões de Competência sobre as Unidades Básicas de Saúde em funcionamento no Brasil	83
Tabela 10 –	Termos com maior frequência extraídos das Questões de Competência para a ontologia das UBS em funcionamento no Brasil	84
Tabela 11 –	Vocabulários e recursos reutilizados ou ligados à ontologia das UBS em Funcionamento no Brasil	85

LISTA DE ABREVIATURAS E SIGLAS

DADOS ABERTOS LIGADOS	<i>Linked Open Data</i>
OGD	<i>Open Government Data</i>
ORSD	<i>Ontology Requirements Specification Document</i>
OWL	<i>Web Ontology Language</i>
RDF	<i>Resource Description Framework</i>
RDFS	<i>RDF Schema</i>
SPARQL	<i>Simple Protocol and RDF Query Language</i>
TSE	Tribunal Superior Eleitoral
URI	<i>Unique Resource Identifier</i>
W3C	<i>World Wide Web Consortium</i>

SUMÁRIO

1	Introdução	14
2	Referencial Teórico	18
2.1	Tecnologias da Web Semântica	18
2.1.1	Ontologia	18
2.1.2	RDF	20
2.1.3	SPARQL	21
2.2	Desenvolvimento de Ontologias	22
2.2.1	Metodologias para Construção de Ontologias	24
2.2.2	Etapas para Desenvolvimento da Ontologia	25
2.2.3	Reutilização de Recursos Ontológicos e Não-ontológicos	28
2.2.4	Conceitualização da Ontologia	29
2.2.5	Implementação da Ontologia	30
2.3	Dados Abertos	31
2.4	Dados Abertos Ligados	32
2.5	Dados Abertos Governamentais	34
2.6	Dados Abertos Governamentais no Brasil	38
3	Trabalhos Relacionados	40
4	Metodologia	49
4.1	Identificação, Análise e Integração dos Dados	49
4.1.1	Identificação	50
4.1.2	Análise e Integração	51
4.2	Desenvolvimento da Ontologia	52
4.2.1	Levantamento de Requisitos	53
4.2.2	Reutilização de Recursos Ontológicos e Não-Ontológicos	54

4.2.3	Conceitualização da Ontologia	56
4.3	Publicação dos Dados como Dados Abertos Ligados	56
4.4	Publicação de <i>Endpoint</i> SPARQL	58
5	Resultados	59
5.1	Dados das Eleições Brasileiras	60
5.1.1	Identificação, Análise e Integração dos Dados	60
5.1.2	Desenvolvimento da Ontologia	62
5.1.3	Publicação como Dados Abertos Ligados	76
5.1.4	Publicação de <i>endpoint</i> SPARQL	78
5.2	Dados de Distribuição das Unidades Básicas de Saúde em Funcionamento	80
5.2.1	Identificação, Análise e Integração dos Dados	80
5.2.2	Desenvolvimento da Ontologia	81
5.2.3	Publicação como Dados Abertos Ligados	91
5.2.4	Publicação de <i>endpoint</i> SPARQL	91
6	Discussões	95
6.1	Publicação de um Arquivo de Dados x Modelagem e Publicação de Dados de um Domínio	95
6.2	Evolução nos Problemas da Publicação de Dados Abertos Ligados	96
6.3	Cenário Brasileiro de Dados Governamentais Abertos Ligados	97
7	Considerações Finais	99
	Referências	102

1- Introdução

Em sua criação, a Web foi pensada e projetada para a navegação humana, guiada por documentos, que são as páginas HTML, e sem uma estrutura forte em relação aos dados nela publicados (Milić et al., 2018). Contudo, desde sua criação já se indicava a necessidade de evolução para que objetos contivessem mais informações semânticas orientadas às máquinas (Berners-Lee et al., 1994). A Web se tornou de fato um grande repositório de dados dos mais diversos tipos e formatos, conectando não somente documentos através da rede, mas também os dados de diferentes domínios, assim possibilitando a evolução da Web de Dados (em inglês, *Web of Data*) (Bizer et al., 2009). Dentro deste contexto, os governos têm um papel muito importante, pois eles são grandes produtores e publicadores de dados sobre os domínios dos quais são responsáveis.

Governos ao redor do mundo têm buscado cada vez mais por ferramentas que permitam que eles sejam mais transparentes, que ajudem a combater a corrupção e que contribuam para um governo participativo, feito em conjunto com sociedade civil. A corrupção por si só é um problema global que afeta diretamente a vida das pessoas e a democracia, levando os cidadãos a terem menos confiança em seus políticos, que, em teoria, deveriam estar trabalhando para atender às preocupações da população e respeitar as leis do país (Attard et al., 2015). Como forma de endereçar estes problemas, muitos países estão criando e desenvolvendo iniciativas de Dados Abertos Governamentais (em inglês, *Open Government Data* (OGD)), que podem ser definidos como dados produzidos e mantidos por agências governamentais que respeitam os requisitos de Dados Abertos (Kučera et al., 2013). Estes requisitos estão associados à não existência de barreiras para acesso aos dados, seja em relação à licença sob a qual eles são disponibilizados, uma restrição ou discriminação de quem pode acessá-los (Attard et al., 2015).

Para tornar ainda mais robustas as iniciativas de Dados Abertos Governamentais, os governos podem combinar seus dados com conjuntos de dados disponíveis na Web, sejam de outros órgãos governamentais ou organizações, utilizando o conceito de Dados Ligados (em inglês, *Linked Data*). Dados Ligados representam um conjunto de tecnologias e padrões da Web Semântica que buscam ligar dados de diferentes fontes por meio da Web de modo que esses dados sejam consumíveis por máquinas de maneira fácil

(*machine-readable*) e seu significado seja explicitamente definido (Bizer et al., 2009). A publicação de Dados Abertos Governamentais utilizando Dados Ligados permite a criação do que é conhecido como Dados Abertos Ligados (em inglês, *Linked Open Data* (Dados Abertos Ligados)). Para os governos existem diversas vantagens em investir em Dados Abertos Ligados e outras tecnologias e padrões da Web Semântica, como a possibilidade de agregar e interligar dados de diferentes setores de maneira integrada, o controle maior sobre os dados publicados e a flexibilidade em casos de mudanças nos dados publicados (Shadbolt et al., 2011).

Apesar dos benefícios da publicação dos dados como Dados Abertos Ligados, ainda existem poucos governos utilizando o potencial da tecnologia. Existem algumas razões que explicam esse fato, como a falta de familiaridade com as tecnologias envolvidas, a falta de um guia de implementação e o baixo interesse da comunidade de desenvolvedores nos padrões propostos pelo Dados Abertos Ligados (Sheridan and Tension, 2010). O uso de Dados Ligados denota um alto nível de maturidade na publicação de dados abertos, o que muitos governos ainda não possuem. Esse nível de maturidade pode ser medido por meio do esquema de implementação das 5 estrelas dos dados abertos proposto por Tim Berners-Lee, que é um modelo de qualidade de 5 etapas para a publicação de dados abertos da melhor maneira possível (Hausenblas and Kim, 2015). A primeira estrela consiste em tornar os dados disponíveis em qualquer formato utilizando uma licença aberta (ex.: PDF (*Portable Document Format*)); a segunda refere-se a torná-los disponíveis como dados estruturados (ex.: planilha); a terceira equivale a disponibilizá-los em um formato aberto não-proprietário (ex.: CSV (*Comma Separated Values*)); a quarta traduz em utilizar URIs (*Uniform Resource Identifier*) para denotar coisas e a quinta representa a ligação de seus dados com outros para prover contexto (Hausenblas and Kim, 2015). Para a maioria dos governos alcançar a terceira estrela é o suficiente e escalar as outras duas etapas pode representar, para eles, um esforço demasiado para a publicação em um formato que pode parecer complicado demais quando comparado com padrões como o CSV.

No entanto, atingir o grau máximo na escala de publicação de dados abertos é uma tarefa importante no processo de amadurecimento da publicação desse tipo de dado por parte dos governos. Essa tarefa vai ao encontro da busca por maior transparência ao mesmo tempo em que empodera a sociedade com uma quantidade mais ampla de ferramentas e formatos para que ela possa trabalhar com esses dados, permitindo a

criação de uma gama maior de tipos de aplicações. Dada essa importância, prover para governos um guia de implementação com passos e processos simplificados a serem seguidos para a publicação de Dados Abertos Ligados, sejam novos ou já existentes, ataca uma das principais barreiras para a adoção dessas tecnologias.

O objetivo desse trabalho é criar uma metodologia baseada no uso de ontologia, com uma sequência de passos a serem seguidos para a publicação de conjuntos de dados governamentais como Dados Abertos Ligados. Ou seja, um guia de implementação para que órgãos governamentais e outras organizações possam atingir a quinta estrela na escala de Berners-Lee, na publicação de dados abertos. Para tanto, dois experimentos utilizando conjuntos de dados reais de domínio governamental são conduzidos de forma a exemplificar a aplicação dos passos. No primeiro são utilizados os conjuntos de dados abertos disponibilizados pelo Tribunal Superior Eleitoral (TSE), órgão responsável pela realização das eleições no Brasil e, portanto, responsável pela disponibilização e manutenção dos dados referentes ao domínio eleitoral. No segundo são utilizados dados sobre as Unidades Básicas de Saúde em funcionamento no Brasil, disponibilizados pelo Ministério da Saúde.

O foco desse trabalho é promover a evolução na publicação de dados governamentais existentes e o uso da tecnologia de Dados Abertos Ligados por parte do setor governamental. No entanto, parte das etapas propostas podem ser aplicadas também para a publicação de novos conjuntos de dados, que não tenham sido ainda publicados em formatos tradicionais (planilhas, arquivos CSV). Como principais resultados obtidos com esse trabalho estão a publicação de dois conjuntos de dados, referentes aos dois experimentos realizados, em arquivos no formato RDF com alguns dos recursos descritos ligados a conjuntos de dados expostos na web, e a produção de duas ontologias, uma referente às eleições no Brasil e outra referente às Unidades Básicas de Saúde em funcionamento no país.

Na literatura existem trabalhos diversos que buscam a evolução na publicação de dados abertos para alcançar a quinta estrela e que executam etapas semelhantes às propostas neste trabalho. Essas abordagens são adotadas para diversos domínios da alçada governamental e utilizam dados produzidos pelo governo, como saúde (Freitas et al., 2017; Lopes et al., 2016), política (Araújo and de Souza, 2011; Brandt et al., 2018) e economia (de Oliveira Araújo et al., 2015), e também em domínios não-governamentais, como música (Pattuelli et al., 2015) e multimídia (Yu et al., 2012). Como novidade do

presente trabalho é destacada a criação de um guia para se alcançar o grau máximo na publicação de Dados Abertos Ligados. Desta forma, espera-se que organizações que hoje não têm o conhecimento para realização desse processo possam utilizar este guia para atingir esse nível de maturidade. Essas etapas são descritas em detalhes no Capítulo 4.

Essa dissertação está dividida em mais seis capítulos, além dessa introdução. No Capítulo 2 são apresentados os principais conceitos necessários para o entendimento desse trabalho. No Capítulo 3 são discutidos os trabalhos relacionados a este. As etapas propostas para publicação de Dados Abertos Governamentais como Dados Abertos Ligados são detalhadas do Capítulo 4. Os resultados obtidos são descritos no Capítulo 5. No Capítulo 6 são abordadas algumas discussões motivadas por essa pesquisa. O Capítulo 7 traz as considerações sobre o trabalho desenvolvido, contribuições e próximos passos.

2- Referencial Teórico

Nesse capítulo são apresentados os conceitos mais relevantes para o devido entendimento deste trabalho. As principais tecnologias relacionadas à Web Semântica são abordadas, uma descrição das metodologias para desenvolvimento de ontologias que foram seguidas, além dos conceitos de dados abertos e dados abertos governamentais. Um panorama das iniciativas brasileiras nesse sentido também é apresentado.

2.1- Tecnologias da Web Semântica

A Web Semântica surgiu para endereçar alguns problemas existentes na Web que navegamos. Dois principais problemas referem-se às limitações de acesso a dados na Web e à necessidade de inferências e junções de informações terem de ser feitas pelos usuários ao invés das máquinas. A Web Semântica se apresenta como uma extensão da Web de Documentos, sendo uma interface de navegação primária por máquinas e que traz contexto e significado aos dados (Domingue et al., 2011). Nessa seção são apresentadas algumas tecnologias que compõem a Web Semântica e que são utilizadas nas etapas propostas por este trabalho.

2.1.1 Ontologia

Ontologia é um termo cuja definição vem evoluindo ao longo do tempo e que pode ter significados diferentes dependendo da área do conhecimento para a qual se está olhando. O termo tem origem na filosofia, definido por Aristóteles como a ciência de ser como tal, buscando responder questões como “o que é um ser?” ou “quais são as características comuns de todos os seres?”. Para se diferenciar do significado de ontologia para outros domínios, quando se trata do domínio filosófico costuma-se escrever

“Ontologia” com letra maiúscula, no singular e sem acompanhamento de artigo (Guarino and Giaretta, 1995).

Para a ciência da computação, uma ontologia (com artigo e sem letra maiúscula no início) é uma maneira formal de modelar a estrutura de um sistema por meio de entidades e relacionamentos, e que as entidades podem consistir em generalizações ou especializações, permitindo assim hierarquias (Guarino and Oberle, 2009). O conceito de ontologia foi evoluindo ao longo do tempo, até chegar a um entendimento mais completo, que é o seguinte: *“uma ontologia é uma especificação formal e explícita de uma conceitualização compartilhada”* (Studer et al., 1998). Mesmo esta definição já foi evoluída para também dizer que uma ontologia deve tratar de um domínio específico de interesse (Domingue et al., 2011).

Como dito anteriormente, uma ontologia modela de maneira formal a estrutura de um sistema por meio de entidades e relacionamentos. Essas entidades são habitualmente chamadas de classes e são utilizadas para representar conceitos do domínio de interesse. Os relacionamentos entre essas classes visam formalizar interações entre elas no mundo real. Essa estrutura também conta com instâncias das classes, que representam unidades daquela classe. A Figura 1 apresenta um pequeno exemplo conceitual de uma ontologia no domínio eleitoral, que é um dos focos desse trabalho. Na figura, podem ser vistas uma instância da classe Eleitor, uma da classe Candidato e uma da classe Partido. Um candidato é filiado a um Partido e um Eleitor vota em um Candidato. João é uma instância de Eleitor, enquanto Fernanda é uma instância de Candidato. O Partido Foo é uma instância de Partido ao qual Fernanda é filiada.

Habitualmente as ontologias também são chamadas de vocabulários ou vocabulários controlados. Essa denominação faz algum sentido dado que uma ontologia representa elementos de um domínio específico, que têm seu próprio vocabulário no mundo real, mas ontologias permitem coisas que simples vocabulários não permitem, como a realização de inferências. Uma ontologia representa esse vocabulário real de forma computacional, utilizando formatos como *Web Ontology Language* (OWL) ou *RDF Schema* (RDFS). As ontologias são especialmente importantes para iniciativas de Dados Ligados, pois elas podem ser facilmente reutilizadas para representar domínios já mapeados ou que possuem ontologias de referência. A criação de uma nova ontologia é necessária apenas quando não existir de fato nenhuma que represente aquele conhecimento ou houver a necessidade de evoluir uma já existente para representar também

uma evolução no conhecimento.

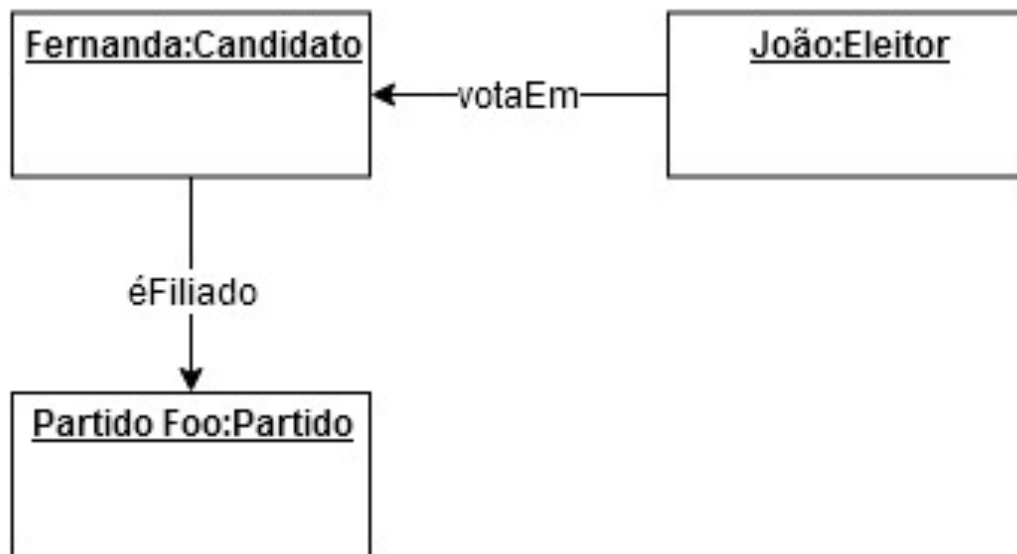


Figura 1 – Exemplo de relacionamento entre instâncias de classes em uma ontologia

2.1.2 RDF

Uma tecnologia que tem ligação direta com a estrutura de ontologias e é fundamental dentro dos padrões de Dados Ligados é o *Resource Description Framework* (RDF). O RDF é o padrão definido pelo *World Wide Web Consortium* (W3C) para o intercâmbio de dados na Web sendo o formato recomendado para troca de informação entre diferentes sistemas, sendo o responsável por prover interoperabilidade entre eles (RDF Working Group, 2014; Shadbolt et al., 2011).

É interessante entender o significado de RDF a partir dos seus termos. Em tradução livre, RDF é um *framework* de descrição de recursos. Um recurso é tudo a que se pode referenciar, como uma pessoa, uma página web, uma imagem etc. A descrição desses recursos que estão na Web se dá por meio de atributos, propriedades e relacionamentos referentes ao recurso. Ser um *framework* significa que essa descrição é feita por meio de modelos, linguagens e sintaxes definidos (Shadbolt et al., 2011).

O formato RDF busca descrever coisas do mundo real por meio de uma estrutura de sujeito-predicado-objeto chamada de tripla. Nessa estrutura tanto o sujeito quanto o objeto são recursos e o predicado é o relacionamento que liga esses recursos. Sujeito, predicado e objeto são representados por meio de *Unique Resource Identifier* (URI), que é um endereço único de identificação de um objeto. Dessa maneira se garante que a informação está atrelada a um documento que pode ser descoberto por qualquer um na Web (Berners-Lee et al., 2001). Podemos exemplificar essa estrutura com a seguinte frase: João (sujeito) é um (predicado) eleitor (objeto), sendo que cada um desses elementos possui uma URI única que pode identificá-los na Web. Dessa maneira, o RDF forma grafos de conexões na Web, ligando recursos em lugares diferentes através de suas URIs, contribuindo para a construção de uma Web de dados ligada.

Como um formato de descrição de recursos, o RDF também pode e é utilizado para representação de ontologias. O formato RDFS é geralmente utilizado para representar ontologias leves, que possuem menor grau de formalismo. O formato OWL é utilizado para representação de ontologias mais completas, que possuem maior grau de formalismo como lógicas e axiomas que dão maior capacidade de inferência (Pattuelli et al., 2015). Ambos os formatos são construídos tendo o RDF como base e são maneiras de descrever ontologias em um formato de fácil leitura por máquinas (*machine-readable*) (Shadbolt et al., 2011).

2.1.3 SPARQL

Com a publicação de dados na Web utilizando o padrão RDF, faz-se necessária a existência de uma linguagem de consulta capaz de buscar e retornar dados publicados utilizando este formato. Para resolver este problema foi criado o *Simple Protocol and RDF Query Language* (SPARQL), que é o padrão proposto pelo W3C para consulta de dados RDF na Web (SPARQL Working Group, 2013).

O SPARQL tem uma sintaxe semelhante ao SQL (*Structured Query Language*, linguagem padrão para consulta em bancos de dados relacionais), mas é uma linguagem mais expressiva por ter sido projetada para funcionar em uma Web aberta, descentralizada e fluída (Shadbolt et al., 2011). As consultas podem ser feitas diretamente em grafos

RDF com o objetivo de busca direta pelos dados ou como forma de popular aplicações (sejam visualizações ou aplicações web) com dados contidos neste formato. As consultas SPARQL geralmente são feitas por meio de *endpoints*, que são URLs pelas quais um serviço ou aplicação, nesse caso um serviço de consulta, podem ser acessadas por clientes.

2.2- Desenvolvimento de Ontologias

Muitas abordagens para a publicação de Dados Abertos Ligados passam pela criação de uma ontologia de domínio como um passo intermediário (Pattueli et al., 2015; Yu et al., 2012; de Oliveira Araújo et al., 2015). No contexto de Dados Ligados essas ontologias funcionam como um vocabulário comum ao domínio dos dados que serão publicados e costumam ser mais leves (em inglês, *Lightweight Ontologies*) do que ontologias tradicionais por possuírem menos formalismos e serem menos restritivas no aspecto semântico. Essa maior flexibilidade permite mais agilidade na construção das ontologias e está de acordo com a natureza da sua utilização, pois em uma web descentralizada, ontologias mais leves permitem um compartilhamento sem que haja necessidade de uma coordenação e concordância sobre a semântica dos elementos que estão sendo representados (Pattueli et al., 2015). Pattueli et al. (2015) diz que “*no contexto de dados ligados, ontologias são simples coleções de triplas com requisitos de modelagem leves e flexíveis*”.

Nesse trabalho, propomos que o desenvolvimento das ontologias leves que serão utilizadas como base para publicação dos dados da Web dos Dados seja feita utilizando metodologias já consolidadas junto a comunidade de web semântica. Para isso foram escolhidas duas metodologias para utilização: a metodologia NeOn (Suárez-Figueroa et al., 2012) e a Ontology Development 101 (Noy and McGuinness, 2001). A contribuição desse trabalho nesse processo se dá por meio da aplicação de etapas de metodologias diferentes ao longo do processo de desenvolvimento para que se produza uma ontologia com poucos artefatos formais, mas que permitam a sua validação e replicação por meio dos artefatos produzidos.

Nessa seção são apresentadas as duas metodologias de construção de ontologias

supracitadas, detalhando as etapas de cada metodologia que foram elencadas para o desenvolvimento deste trabalho. A Figura 2 apresenta tais etapas e os artefatos produzidos, organizadas em um fluxo de trabalho que foi proposto para a presente pesquisa. Ao longo deste fluxo de trabalho foram feitas adaptações sobre o que é definido em algumas etapas das metodologias, de modo que ficassem mais adequadas ao objetivo de publicar os dados como Dados Abertos Ligados. Essas contribuições são destacadas na Seção 4.2.

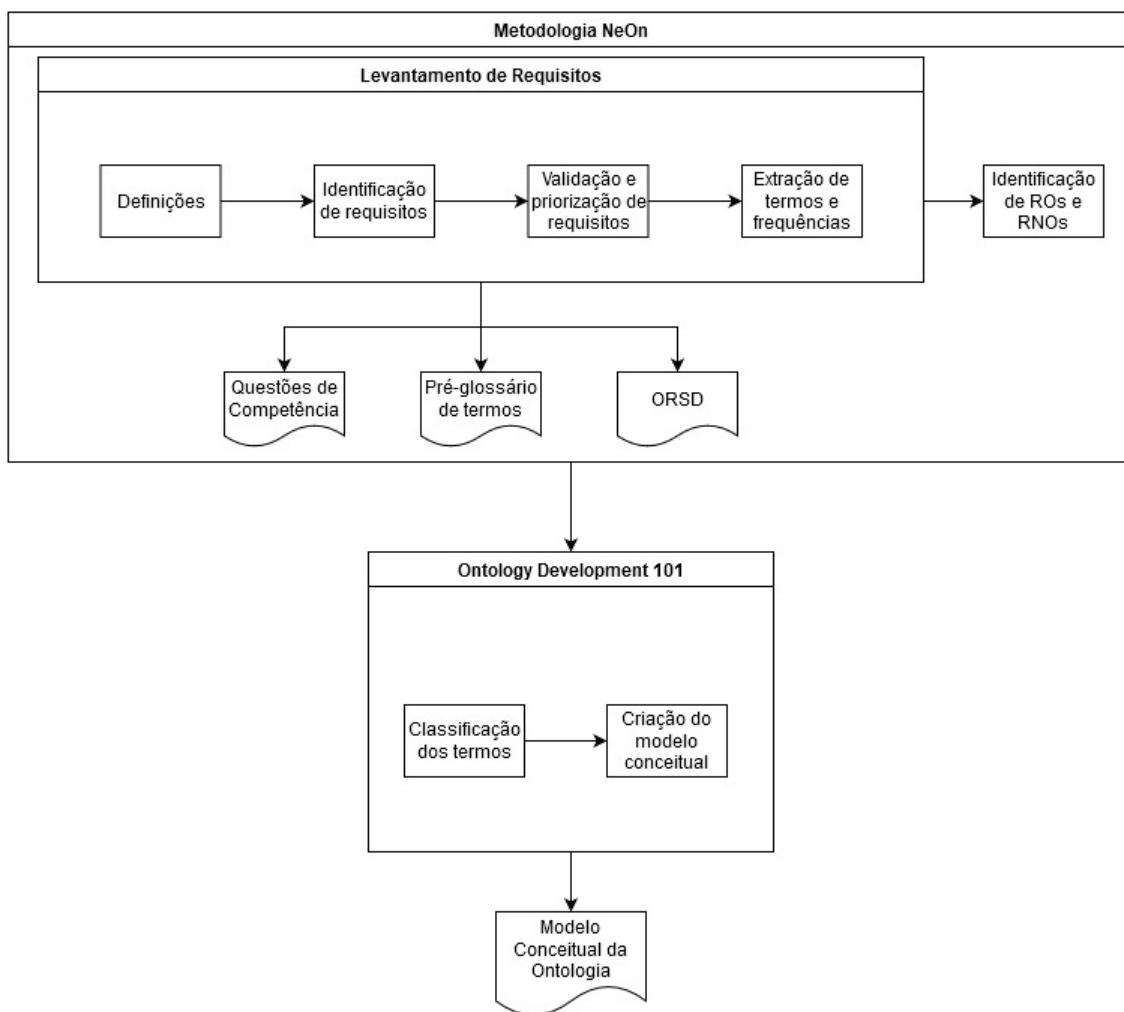


Figura 2 – Fluxo de trabalho proposto para a construção da ontologia utilizando etapas das metodologias NeOn e Ontology Development 101

2.2.1 Metodologias para Construção de Ontologias

O estudo de ontologias é bastante difundido nas áreas de representação do conhecimento e inteligência artificial. Existem várias metodologias que foram propostas ao longo dos anos para o desenvolvimento de ontologias, para diferentes domínios e de variadas maneiras. Metodologias como a METHONTOLOGY (Fernández-López et al., 1997), Ontology Development 101 (Noy and McGuinness, 2001), SABiO (Falbo, 2014) e NeOn (Suárez-Figueroa et al., 2012) são bastante conhecidas entre desenvolvedores de ontologias, e possuem, inclusive, muitos aspectos semelhantes entre elas.

A metodologia NeOn é baseada em cenários, a qual apoia diversas situações do desenvolvimento de ontologias, além de ter um foco grande em colaboração e no reuso de ontologias e recursos não-ontológicos (Suárez-Figueroa et al., 2012). As características em questão motivaram a escolha dessa metodologia como guia primário para o desenvolvimento da ontologia deste trabalho. As principais etapas aplicadas referem-se ao levantamento de requisitos e à identificação de recursos ontológicos e não-ontológicos. Contudo, a etapa de conceitualização da ontologia é pouco detalhada pela metodologia NeOn. Por este motivo, foi avaliado como as metodologias METHONTOLOGY, SABiO e a Ontology Development 101 tratavam a etapa referente. Apesar de serem três abordagens diferentes, todas elas têm como objetivo tornar claro os principais elementos da ontologia que está sendo desenvolvida, ou seja, identificar e explicitar conceitos, atributos e relações. A principal diferença entre elas está nos artefatos que são produzidos durante o processo e o grau de formalismo derivado desses artefatos.

A METHONTOLOGY é a metodologia que especifica o maior número de artefatos a serem produzidos durante a conceitualização. Esses artefatos servem como instrumentos para que os usuários da ontologia não precisem analisar o código dela para entender sua utilidade, além de tornar mais fácil a comparação com outras ontologias a respeito do escopo e completude. SABiO propõe maior grau de formalismo em comparação às outras, e faz uso da OntoUML para modelagem de conceitos e taxonomias para organização, e axiomas implementados em Lógica de Primeira Ordem para se evitar ambiguidades. A Ontology Development 101 não apresenta explicitamente uma etapa de conceitualização. Contudo, ao compará-la com as outras metodologias, nota-se que ela apresenta a descrição das etapas que são realizadas em um processo de conceitualização

de uma ontologia. Ela tem seu foco voltado para a parte prática da implementação, além de ser direcionada para um público com menos experiência em desenvolvimento de ontologias, devido ao número menor de formalismos e artefatos a serem produzidos. Por essas características e pelo nosso foco na construção de ontologias leves, com escopo e domínio bem definidos, a metodologia *Ontology Development 101* foi a utilizada neste trabalho e a que recomendamos para essa tarefa.

Assim, o processo para o desenvolvimento de ontologias proposto por esse trabalho utiliza etapas da metodologia *NeOn* para o levantamento de requisitos e identificação de recursos ontológicos e não-ontológicos para reutilização. E para a conceitualização da ontologia são utilizadas etapas da metodologia *Ontology Development 101*. A implementação da ontologia é feita utilizando como base o modelo conceitual produzido.

2.2.2 Etapas para Desenvolvimento da Ontologia

Nessa seção são detalhadas as etapas seguidas para o processo de desenvolvimento de uma ontologia leve de modo a apoiar a publicação de um conjunto de dados como *Dados Abertos Ligados*. As etapas aqui descritas foram utilizadas nos experimentos apresentados no Capítulo 5.

Levantamento de Requisitos

No processo de desenvolvimento de ontologias, uma das principais etapas é a de definição dos requisitos que a ontologia deve atender, de modo a garantir que o que for desenvolvido esteja de acordo com as definições previamente estabelecidas, de maneira semelhante aos processos da área de Engenharia de Software. O processo de levantamento de requisitos faz com que se tenha maior assertividade e agilidade no momento do desenvolvimento da ontologia em si, além de ser um instrumento para validação dela. Nesta seção são descritas as etapas realizadas para o levantamento de requisitos utilizando como base as etapas recomendadas pela metodologia *NeOn*.

Questões de Competência O desenvolvimento de Questões de Competência (*Competency Questions*) é uma técnica bastante utilizada e difundida para auxiliar no levantamento dos requisitos. Proposta por Grüninger and Fox (1995), as Questões de Competência (QCs) são parte de um procedimento para desenvolvimento e avaliação de uma ontologia e são utilizadas como requisitos no formato de perguntas escritas em linguagem natural que uma ontologia deve responder. Além de servir como forma de se levantar os requisitos funcionais, as QCs também permitem uma avaliação inicial da ontologia, ajudando a responder se as questões propostas podem ser atendidas por uma ontologia já existente ou se é necessária a extensão de uma ontologia já existente ou criação de uma nova.

De acordo com a metodologia NeOn, essa etapa tem como objetivo a geração de um Documento de Especificação de Requisitos da Ontologia (Suárez-Figueroa et al., 2012) (*Ontology Requirements Specification Document* (ORSD) na sigla em inglês). Porém, antes da elaboração das QCs é necessário que se defina qual o cenário motivador da construção da ontologia (Grüninger and Fox, 1995). A metodologia NeOn sugere também que além do cenário motivador sejam descritos o escopo da ontologia e a linguagem na qual ela será desenvolvida.

Cenário Motivador, Escopo e Linguagem Nessa tarefa deve-se descrever qual a motivação para a construção da ontologia que se está propondo, qual será o seu escopo, ou seja, que parte do conhecimento do domínio será representado por ela, e qual será a linguagem de implementação da ontologia. Com essas definições é possível identificar o objetivo buscado com o desenvolvimento da ontologia e o que será representado.

Usuários e Uso Pretendido Nessa tarefa é feita a identificação de quem serão os possíveis usuários da ontologia e qual será o uso dado por eles para a ontologia. Com esse entendimento é possível projetá-las de acordo com as necessidades de quem de fato utilizará o que for desenvolvido.

Identificação de Requisitos A etapa de identificação de requisitos busca identificar tanto os requisitos funcionais quanto os não-funcionais da ontologia que será desenvolvida. Para a definição dos requisitos funcionais são empregadas as Questões de Competência. As questões escritas em linguagem natural são conhecidas como Questões de Competência Informais (Grüninger and Fox, 1995), por não estarem representadas na

linguagem formal da ontologia.

As QCs devem ser elaboradas considerando-se os aspectos importantes que a ontologia proposta deve responder. Após a finalização da lista inicial de questões, realiza-se uma categorização delas por temas. Essa categorização serve de apoio para o processo de priorização das QCs, realizado após a sua elaboração. Ela também pode auxiliar em cenários nos quais se deseja dividir ou modularizar uma ontologia em outras menores para que se obtenha uma redução de escopo e complexidade (Suárez-Figueroa et al., 2012). Para o desenvolvimento das QCs recomenda-se a consulta a especialistas do domínio, que podem validar se os requisitos modelados em forma de questões atendem aos cenários reais daquele domínio.

Validação dos Requisitos Essa tarefa busca refinar os requisitos por meio da utilização de diferentes critérios para análise, os quais tentam identificar problemas como conflitos e contradições entre eles e requisitos faltantes (Suárez-Figueroa and Gómez-Pérez, 2012).

A metodologia NeOn recomenda uma série de critérios para se avaliar os requisitos. São eles: corretude, completude, consistência, verificabilidade, entendibilidade, não ambiguidade, concisão, realismo, modificabilidade e rastreabilidade (Suárez-Figueroa and Gómez-Pérez, 2012). Recomendamos que sejam selecionados os que fazem mais sentido para os dados com os quais se está trabalhando de forma que a validação de fato sirva para garantir que os requisitos pensados sejam corretos e façam sentido.

Priorização dos Requisitos Assim como no desenvolvimento de software, a priorização também é presente no desenvolvimento de ontologias. Dessa forma, a priorização dos requisitos da ontologia, define o que será desenvolvido. Os requisitos priorizados devem ser suficientes para atender ao escopo definido, aos usuários e ao uso pretendido previamente descritos.

Extração de Terminologia e Frequências A etapa de extração das terminologias e suas frequências têm como objetivo a obtenção de um pré-glossário de termos a partir da lista de QCs e suas repostas, no qual sejam identificados elementos no universo de discurso que vão ser transformados em elementos da ontologia que se está construindo. Esses elementos são conceitos, atributos, relacionamentos e instâncias de objetos que vão dar forma à ontologia. Obter a frequência dos termos ajuda na busca por recursos ontológicos e não-ontológicos que possam ser reaproveitados no desenvolvimento da

ontologia (Suárez-Figueroa et al., 2012).

Para identificação dos termos, deve-se buscar por substantivos, adjetivos e verbos presentes nas Questões de Competência e em suas respostas, que representarão os conceitos, propriedades, relacionamentos e instâncias na ontologia. O processo de extração deve ser feito separadamente para as questões e suas respostas pois, segundo a metodologia NeOn, a classificação dos termos extraídos é diferente entre eles. Das questões são retirados conceitos, atributos, relacionamentos e instâncias. Das respostas são retirados apenas conceitos e instâncias (Suárez-Figueroa et al., 2012).

A criação do pré-glossário de termos é a última etapa da fase de levantamento de requisitos de uma ontologia proposta pela metodologia NeOn. Todas as informações levantadas nessa etapa são utilizadas para o preenchimento do Documento de Especificação de Requisitos da Ontologia, que é um artefato utilizado nas etapas seguintes do desenvolvimento da ontologia.

2.2.3 Reutilização de Recursos Ontológicos e Não-ontológicos

Em busca de otimizar o tempo e o esforço empregado no desenvolvimento de ontologias, a busca por recursos que possam ser reaproveitados nesse processo é uma boa prática. Esses recursos podem ser tanto outras ontologias como também tesauros, léxicos e esquemas de classificação, os quais são classificados como recursos não-ontológicos. O principal artefato de apoio para que se faça a identificação de recursos ontológicos (RO) e recursos não-ontológicos (RNO) é o Documento de Especificação de Requisitos da Ontologia, elaborado na etapa de levantamento de requisitos e que apresenta o propósito, uso e usuários pretendidos, identificação dos requisitos por meio de Questões de Competência e o pré-glossário de termos retirados dos requisitos. A metodologia NeOn propõe que se faça a busca em sites confiáveis na internet, sites específicos do domínio da ontologia, em órgãos de normas e padrões, e em recursos organizacionais a partir dos termos do pré-glossário identificados com maior frequência (Suárez-Figueroa et al., 2009).

Após o levantamento dos recursos, a metodologia NeOn prevê que se faça uma avaliação deles e propõe alguns critérios para a realização dessa atividade. A seleção

desses recursos é importante como complemento da ontologia que será construída, podendo ser utilizados para criar uma ontologia mais completa do domínio, ou até mesmo criar uma rede de ontologias.

2.2.4 Conceitualização da Ontologia

A conceitualização da ontologia, que dá origem ao modelo conceitual, é uma etapa anterior à implementação de uma ontologia. A conceitualização é uma maneira de formalizar o conhecimento sobre um domínio por meio de objetos e seus relacionamentos. Nesse contexto, objetos são definidos como qualquer coisa sobre o qual se queira dizer algo e, habitualmente, representam um universo de conhecimento reduzido, chamado de universo de discurso (Genesereth and Nilsson, 1987). Como explicado na Seção 2.2.1, para o processo de conceitualização, recomenda-se a utilização da metodologia *Ontology Development 101* (Noy and McGuinness, 2001).

A primeira etapa proposta pela metodologia *Ontology Development 101* é a listagem de termos importantes da ontologia. Como nesse trabalho propomos uma abordagem de realizar um levantamento de requisitos antes, utilizando etapas da metodologia *NeOn*, aqui propõe-se utilizar o pré-glossário de termos produzido durante o levantamento, que traz os termos mais relevantes com base nas QCs. O pré-glossário apresenta ainda informações adicionais, como a classificação de cada termo (conceito, atributo, relacionamento ou instância). Dada essa classificação, as etapas seguintes são facilitadas, pois englobam a identificação das classes (mapeadas como conceitos no pré-glossário), suas hierarquias, propriedades (mapeadas como atributos e relacionamentos no pré-glossário) e indivíduos. Nesse caso, propõe-se seguir a proposta da *Ontology Development 101* como uma maneira de reavaliar o que havia sido mapeado no pré-glossário.

Uma etapa que não é proposta na *Ontology Development 101* mas que recomendamos nesse trabalho é que após reavaliar os termos e garantir que estão classificados corretamente, se formalize o que foi mapeado em um diagrama de classes. Criar esse diagrama antes de iniciar a implementação ajuda a organizar melhor a arquitetura da ontologia, tornando mais fácil visualizar se as classes estão se conectando de uma forma que faça sentido e se os relacionamentos mapeados são suficientes ou se é necessária

a criação de novos. Como o foco da Ontology Development 101 é na implementação, o que foi identificado no pré-glossário de termos poderia ser diretamente implementado em uma ferramenta como Protégé¹ utilizando uma linguagem para desenvolvimento de ontologias, como a OWL. No entanto, acreditamos que a criação do diagrama contribuí para que sejam necessárias menos iterações na implementação (correções do modelo) pois possíveis problemas serão identificados durante seu desenvolvimento.

2.2.5 Implementação da Ontologia

Para realizar a implementação da ontologia foi utilizada a ferramenta Protégé, que é a indicada na metodologia Ontology Development 101. Como dito na Seção 2.2.4, tendo um modelo conceitual desenhado torna-se mais simples fazer a transposição para a implementação. As classes do modelo são criadas como classes ou subclasses na ferramenta. Os relacionamentos entre classes são criados como propriedades de objetos. As propriedades de uma classe que não são ligações com outras classes são criadas como propriedades de dados. Instâncias de classes são criadas como indivíduos.

A implementação da ontologia é o momento no qual serão feitas as ligações com recursos externos que vão permitir que o conjunto de dados seja classificado como Dados Abertos Ligados. Os recursos ontológicos e não-ontológicos (que tenham passado por processo de reengenharia para ontologia) e recursos “avulsos” (recursos identificados em conjuntos de dados publicados como Dados Abertos Ligados mas que não estejam dentro de uma ontologia) podem ser adicionados à ontologia de duas formas: (i) sendo importados para a ontologia; ou (ii) sendo ligados a recursos existentes usando propriedades da ontologia. É importante destacar que para a publicação como Dados Abertos Ligados é necessária que sejam feitas ligações a nível de indivíduo.

Para a primeira opção, o Protégé oferece a possibilidade de importar outras ontologias diretamente de URIs ou de arquivos. As classes, propriedades e indivíduos da ontologia importadas são adicionadas à ontologia em desenvolvimento, sendo exibidas da mesma forma que as classes, propriedade e indivíduos criados manualmente.

A segunda possibilidade geralmente é feita utilizando a propriedade de anotação

¹<https://protege.stanford.edu/>

`owl:sameAs`, do vocabulário OWL² para os indivíduos da ontologia. Ao utilizar essa abordagem está sendo dito que um recurso da ontologia é o mesmo que um outro recurso externo, referenciado por meio de sua URI única. Essa propriedade é usada como o predicado na estrutura de tripla e vai indicar que um indivíduo (instância) na ontologia tem as mesmas propriedades que um outro indivíduo, externo à ontologia, identificado através de sua URI. Também existem as anotações `owl:equivalentClass`³, utilizado para equivalência entre classes, e `owl:equivalentProperty`⁴, utilizado para equivalência entre propriedades.

Depois de criadas as estruturas da ontologia, precisamos popular a mesma com os dados existentes nos conjuntos de dados a ser publicados. O modo como isso será feito deriva da estratégia de integração dos dados realizada e o formato destes dados (planilhas, arquivos CSV, bases relacionais etc.). Essa etapa é explicada em mais detalhes na Seção 4.1. A escolha da ferramenta para apoiar esta tarefa depende do formato dos dados. Para esse trabalho, pela natureza dos dados e decisão de mantê-los em seu formato original (CSV), foi utilizado um *plugin* nativo do Protégé chamado Cellfie⁵, que permite a importação de dados armazenados em planilhas no formato XLS.

2.3- Dados Abertos

Na Web de Dados, quanto maior o volume de dados, melhor. E os Dados Abertos (em inglês, *Open Data*) são os principais responsáveis por povoá-la. Para que se entenda o que são os Dados Abertos Governamentais, conceito apresentado na Seção 2.5, é preciso entender o que são Dados Abertos. Um dado ou conhecimento pode ser classificado como aberto se ele respeitar quatro princípios (Open Knowledge Foundation, 2015):

- Licença: os dados precisam ser disponibilizados utilizando uma licença aberta ou estarem em domínio público e qualquer tipo de termo disponibilizado em conjunto com os dados deve também respeitar essa licença.

²<https://www.w2.org/TR/owl-ref/#sameAs-def>

³<https://www.w3.org/TR/owl-ref/#equivalentClass-def>

⁴<https://www.w3.org/TR/owl-ref/#equivalentProperty-def>

⁵<https://github.com/protegeproject/cellfie-plugin>

- Acesso: os dados disponibilizados devem ser completos e com um custo de reprodução único razoável. Também deve ser possível baixá-los gratuitamente através da Internet.
- Leitura por máquinas: o formato no qual os dados são disponibilizados devem ser facilmente legíveis e processáveis por aplicações computacionais.
- Formato: o formato no qual os dados são disponibilizados deve ser aberto, ou seja, não apresentar restrições para seu consumo e processamento e nem utilizar formatos proprietários.

Além dos quatro princípios listados também existe uma série de requisitos que a licença sob a qual os dados são disponibilizados deve respeitar, como a gratuidade para uso e redistribuição dos dados. Isso para que se tenha a garantia de que o dado é realmente aberto para seu uso, reuso e distribuição tanto no aspecto legal quanto no tecnológico. Em linhas gerais uma licença, em conformidade com as políticas adotadas por outras licenças abertas, deve atender os requisitos necessários (Open Knowledge Foundation, 2015).

Os Dados Abertos são chamados de Dados Abertos Ligados quando eles são publicados na Web seguindo os padrões tecnológicos da Web Semântica. Além disso, devem ser ligados a outros conjuntos de dados e permitir o fácil consumo por máquinas.

2.4- Dados Abertos Ligados

A evolução da Web de Documentos para uma Web de Dados não se trata apenas da publicação de conjuntos de dados utilizando os padrões da Web Semântica como RDF e OWL. O grande valor da Web de Dados está na possibilidade de conexão entre esses dados e o potencial que essas conexões possuem. Dessa forma torna-se possível uma navegação entre conjuntos de dados que até então não conversavam entre si, permitindo que alguém que esteja os explorando possa, de maneira fluída, descobrir mais informações e significados sobre algum recurso. Essa estrutura para navegar e explorar os dados refere-se aos Dados Abertos Ligados (Dados Abertos Ligados).

De maneira mais técnica, Dados Abertos Ligados se referem a dados que res-

peitam os requisitos de dados abertos e que sejam facilmente lidos por máquinas, que tenham significado explícito, que façam ligações com outros conjuntos de dados na Web, e que recebam ligações de outros conjuntos de dados (Bizer et al., 2009). Existem quatro requisitos específicos dos Dados Abertos Ligados (Berners-Lee, 2006), além dos princípios dos Dados Abertos apresentados na seção anterior:

- Usar URIs para nomear coisas na Web;
- Usar URIs HTTP para que pessoas possam encontrar esses nomes;
- Prover informações importantes (semânticas e de contexto) quando alguém acessar as URIs utilizando as tecnologias padrão da Web Semântica;
- Incluir links para outras URIs, permitindo a descoberta de novos recursos.

No entanto, não basta apenas seguir os requisitos listados anteriormente para que um conjunto de dados publicado como Dados Abertos Ligados tenha a chancela para ser parte da LOD *Cloud*⁶, que é uma coleção de conjuntos de dados publicados como Dados Abertos Ligados na Internet que tem ligações entre si. Para fazer parte dessa nuvem são necessários alguns requisitos adicionais como ter um conjunto de dados que tenha pelo menos 1000 triplas e que seja ligado com pelo menos 50 outros recursos (por meio de URIs). A Figura 3 mostra uma sub-nuvem da Dados Abertos Ligados *Cloud* apenas relacionado a dados governamentais. A nuvem completa pode ser explorada em <https://lod-cloud.net/>.

Como pode ser observado nos princípios, a utilização dos padrões da Web Semântica como RDF e SPARQL são fundamentais para a utilização de Dados Abertos Ligados na Web. Por meio da estrutura de grafo do formato RDF os diferentes recursos espalhados pela Web Semântica são conectados. Dessa forma, a estrutura de sujeito-predicado-objeto é representada por URIs, que são únicas para cada um desses elementos. O contexto e significado necessário para entendimento do recurso que se está acessando é dado por meio da notação do formato. Na Figura 4 observa-se uma URI referente ao Dalai Lama mantida pela organização do Prêmio Nobel. Um exemplo de relação que pode ser vista entre URIs é que Dalai Lama (sujeito - <http://data.nobelprize.org/page/laureate/551>) nasceu (predicado - <http://data.nobelprize.org/page/laureate/551>)

⁶<https://lod-cloud.net/>

The 14th Dalai Lama (Tenzin Gyatso)	
Resource URI: http://data.nobelprize.org/resource/laureate/551	
Home All laureate	
Property	Value
dbpedia-owl:birthPlace	< http://data.nobelprize.org/resource/city/Taktser >
dbpedia-owl:birthPlace	< http://data.nobelprize.org/resource/country/Tibet_%28now_People%27s_Republic_of_China%29 >
foaf:birthday	1935-07-06 (xsd:date)
dbpprop:dateOfBirth	1935-07-06 (xsd:date)
foaf:gender	male
foaf:givenName	The 14th Dalai Lama (Tenzin Gyatso)
rdfs:label	The 14th Dalai Lama (Tenzin Gyatso)
is nobel:laureate of	< http://data.nobelprize.org/resource/laureateaward/546 >
is nobel:laureate of	< http://data.nobelprize.org/resource/nobelprize/Peace/1989 >
nobel:laureateAward	< http://data.nobelprize.org/resource/laureateaward/546 >
foaf:name	The 14th Dalai Lama (Tenzin Gyatso)
nobel:nobelPrize	< http://data.nobelprize.org/resource/nobelprize/Peace/1989 >
owl:sameAs	< http://dbpedia.org/resource/14th_Dalai_Lama >
owl:sameAs	freebase:m.01hw4n
owl:sameAs	< http://viaf.org/viaf/38242123 >
rdf:type	nobel:Laureate
rdf:type	foaf:Person
Metadata	
< http://data.nobelprize.org/data/laureate/551 >	
dc:date	2019-10-03T03:57:21.902Z
prv:containedBy	< http://data.nobelprize.org/dataset >
void:inDataset	< http://data.nobelprize.org/dataset >
rdf:type	prv:Dataltem
rdf:type	foaf:Document

Figura 4 – Exemplo de URI com dados sobre o Dalai Lama mantido pela organização do Prêmio Nobel. Fonte: Nobel Prize Data

e econômicos são só alguns exemplos de domínios dos quais órgãos governamentais são responsáveis e, portanto, também são os proprietários dos dados dessas áreas. Nos últimos anos, governos ao redor do mundo têm investido em iniciativas de disponibilização desses dados de maneira aberta como forma de aumentar a transparência, gerar valor social e comercial, e incentivar uma governança participativa (Attard et al., 2015). Iniciativas como a *Public Sector Information Directive* (PSI)⁷ criado em 2003 na Europa, o memorando do ex-presidente norte-americano Barack Obama sobre transparência e governo aberto em 2009⁸ e a criação do *Open Government Partnership*⁹ de 2011 são algumas das confirmações de como governos estão preocupados com a transparência e

⁷<http://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX:32003L0098>

⁸<https://obamawhitehouse.archives.gov/the-press-office/transparency-and-open-government>

⁹<https://www.opengovpartnership.org/>

como veem na abertura e disponibilização de dados o caminho para alcançá-la (Attard et al., 2015).

Posto isso, Dados Abertos Governamentais é definido como dados de governo, ou seja, que são criados por órgãos governamentais, que atendem aos requisitos de Dados Abertos (Kučera et al., 2013). Além dos princípios dos Dados Abertos que precisam ser respeitados, existem também oito princípios específicos de Dados Abertos Governamentais que podem ser utilizados para a correta identificação e classificação de iniciativas desse tipo. São eles (Open Gov Data, 2007):

- **Completos:** todos os dados públicos devem ser disponibilizados. Incentiva-se que mesmo artefatos físicos, como documentos não-digitalizados, sejam disponibilizados de maneira digital.
- **Primários:** dados coletados em sua fonte primária e disponibilizados no maior nível de granularidade possível, sem agregações ou modificações.
- **Em tempo:** os dados devem ser disponibilizados assim que possível para que seja preservado seu valor.
- **Acessível:** os dados devem ser de fácil acesso para qualquer tipo de usuário e uso.
- **Processável por máquinas:** os dados devem ser estruturados de forma a permitir seu consumo e processamento de maneira fácil por máquinas.
- **Não-discriminatório:** os dados devem ser disponibilizados a todos sem a necessidade de registro.
- **Não-proprietário:** os dados devem ser disponibilizados em um formato aberto, que não seja proprietário, permitindo controle exclusivo.
- **Livre de licenças:** os dados não devem estar sujeitos a nenhuma restrição de direito autoral, patente, marca registrada ou regulamento de segredos industriais.

Como é possível notar, alguns dos princípios são semelhantes aos quatro vistos na Seção 2.3, o que faz sentido, uma vez que os Dados Abertos Governamentais são considerados como uma vertente dos Dados Abertos. Além dos oito princípios listados, mais sete princípios foram adicionados posteriormente. Eles dizem que os dados devem ser disponibilizados online e gratuitamente; devem ser permanentes; devem ser confiáveis;

deve haver presunção de abertura; devem ser documentados; sua abertura deve ser segura; e devem ser desenhados de acordo com retorno do público (Open Gov Data, 2007).

Os conjuntos de dados que são criados e mantidos por governos são dos mais variados tipos. Dados do Censo, o orçamento da União, a utilização da cota parlamentar, a disseminação do vírus da Zika, a taxa de escolaridade da população, o número de homicídios de uma cidade e Diários Oficiais são alguns exemplos de conjuntos de dados gerados por órgãos governamentais. É importante ressaltar que nem todo tipo de dado pode ser liberado como um dado aberto. Dados que possam comprometer a segurança ou soberania nacional são um exemplo de conjunto de dados que não pode ser disponibilizado como Dados Abertos Governamentais (Kučera et al., 2013). A disponibilização dos Dados Abertos Governamentais geralmente se dá por meio de Portais de Dados, que servem como um agregador de diferentes fontes de dados governamentais, o que traz praticidade a quem deseja buscar esses dados por não ter que buscá-los em diferentes ambientes e instituições geradoras dos mesmos (Attard et al., 2015).

Existem três motivos principais para a abertura de dados governamentais: transparência, geração de valor social e comercial e governança participativa (Attard et al., 2015). Os benefícios trazidos pela abertura são grandes, principalmente no campo político e social, contribuindo com o aumento na transparência dos governos, a geração de maior confiança da sociedade, o fortalecimento da responsabilidade dos órgãos governamentais, a maior possibilidade de participação e fiscalização da sociedade (empoderamento dos cidadãos), entre outros (Janssen et al., 2012). No entanto, também existem obstáculos para a implantação de iniciativas efetivas de abertura de dados. Alguns exemplos dessas barreiras são a falta de políticas para abertura e publicação, falta de vontade em publicar dados relevantes, uma cultura de aversão ao risco, falta de ferramentas para tratar *feedback* dos usuários, falta de conhecimento dos usuários em como explorar aqueles dados e problemas com a qualidade dos dados disponibilizados (Janssen et al., 2012).

2.6- Dados Abertos Governamentais no Brasil

O Brasil vem trabalhando há alguns anos na abertura de dados governamentais, com as primeiras iniciativas datando de 2009, época em que o Comitê Organizador de Informações da Presidência do Brasil iniciou um trabalho de coleta de dados governamentais para publicação em meios digitais e que posteriormente veio a ser aberto ao público por meio do portal DadosGov (Breitman et al., 2012). Em setembro de 2011 o Brasil se tornou membro do *Open Government Partnership*, uma iniciativa internacional com participação de diversos países que tem como objetivo promover a abertura de dados governamentais (Breitman et al., 2012). Em novembro do mesmo ano foi criada a Lei de Acesso à Informação (LAI), que regulamenta o direito de qualquer pessoa poder ter acesso a informações públicas e que visa que a publicidade dos dados governamentais seja uma regra, aplicada independentemente do órgão ou poder (Brasil, 2011).

Além da Lei de Acesso à Informação, outro importante dispositivo relacionado a Dados Abertos Governamentais no Brasil é a Infraestrutura Nacional de Dados Abertos (INDA)¹⁰. A INDA é a política do governo brasileiro para Dados Abertos e que tem como seu principal projeto o portal `dados.gov.br`. Esse é considerado o principal Portal de Dados do Brasil e assim visa ser o ponto central para a publicação, busca e acesso de dados públicos do país.

Apesar de ser um país comprometido com iniciativas de dados abertos, ao observar o nível de maturidade dessas iniciativas, principalmente em relação a utilização de Dados Abertos Ligados, há ainda um longo caminho para alcançar um alto grau. Essa questão pode ser notada ao consultar a quantidade de conjuntos de dados publicados em formato RDF no Portal Brasileiro de Dados Abertos em comparação aos disponibilizados em formato CSV. Para o primeiro formato existem 24 conjuntos de dados, enquanto para o segundo existem mais de 5.000 publicações (dados de setembro de 2019).

Essa diferença de quantidade de conjuntos de dados publicados não surpreende, dado que a própria Cartilha Técnica para Publicação de Dados Abertos no Brasil¹¹ trata o formato RDF como sendo de implementação mais complexa. Apesar disso, o governo brasileiro possui iniciativas interessantes para evolução de suas políticas de

¹⁰<https://www.governodigital.gov.br/transformacao/cidadania/dados-abertos/inda-infraestrutura-nacional-de-dados-abertos>

¹¹<http://dados.gov.br/pagina/cartilha-publicacao-dados-abertos>

dados abertos como a já citada Infraestrutura Nacional de Dados (INDA); o ePING¹², que abriga os padrões de interoperabilidade do governo eletrônico e define quais tecnologias e padrões devem ser preferencialmente usadas para troca de informações do governo; e o e-VoG¹³, um repositório de vocabulários e ontologias do governo eletrônico que funciona como um conjunto de padrões, ferramentas e metodologias para habilitar a troca de informações usando vocabulários semânticos e ontologias para domínios governamentais. É válido ressaltar que os padrões sugeridos pelo ePING para organização e intercâmbio de informações estão totalmente de acordo com as tecnologias da Web Semântica, pois recomendam o uso dos formatos que constituem essas tecnologias anteriormente discutidas como RDF para descrição de recursos e especificação de vocabulários e SPARQL como linguagem de consulta dinâmica.

¹²<http://eping.governoeletronico.gov.br/>

¹³<http://vocab.e.gov.br/>

3- Trabalhos Relacionados

Para identificar os trabalhos relacionados à presente pesquisa foi conduzido um mapeamento sistemático de trabalhos sobre Dados Abertos Ligados ao domínio governamental. Para isso foi definida uma *string* de busca para ser utilizada em bases de pesquisa a fim de identificar artigos de referência sobre o assunto.

A base escolhida para pesquisa foi a Scopus¹ e a *string* de busca utilizada foi a seguinte: *TITLE-ABS-KEY(("linked data" OR "linked open data" OR "linked open government data" OR "Dados Abertos Ligados" OR "logd") AND ("open government" OR "open government data" OR "digital government" OR "electronic government" OR "e-government" OR "ogd"))*. Essa busca retornou 199 resultados, os quais foram analisados com base em três critérios de inclusão e sete critérios de exclusão. A Tabela 1 apresenta os critérios de inclusão (CI) e a Tabela 2 os critérios de exclusão (CE).

Tabela 1 – Critérios de inclusão de artigos no mapeamento sistemático

Código	Critério
CI1	Trabalhos da área de computação que explorem a utilização de <i>Linked Open Data</i> e <i>Open Government Data</i>
CI2	Trabalhos da área de Representação da Informação que explorem a utilização de <i>Linked Open Data</i> e <i>Open Government Data</i>
CI3	Trabalhos que explorem o processo de publicação de <i>Linked Open Data</i>

Dos 199 artigos resultantes da *string* de busca, 17 foram selecionados com base nos critérios para uma leitura completa e avaliação. Após nova filtragem feita com base na leitura completa, 12 artigos foram considerados os trabalhos mais relacionados ao tema desta pesquisa. Além dos trabalhos identificados por meio do mapeamento sistemático, foram também incluídos 6 trabalhos identificados por meio do processo de *snowballing*. Após a revisão dessa dissertação, outros 3 trabalhos foram incluídos. Esses trabalhos são destacados a seguir.

O governo mais notável pela adoção de Dados Abertos Ligados é o do Reino Unido. Ele possui um compromisso em aplicar os padrões da W3C para a publicação de

¹<https://www.scopus.com>

Tabela 2 – Critérios de exclusão de artigos do mapeamento sistemático

Código	Critério
CE1	A publicação é apenas um resumo (abstract)
CE2	O artigo é uma cópia ou uma versão mais antiga de outro artigo já considerado
CE3	Não foi possível ter acesso ao trabalho completo
CE4	O artigo não trata de <i>Linked Open Data</i> ou <i>Open Government Data</i>
CE5	O artigo não explora a publicação de <i>Linked Open Data</i> por governos ou órgãos governamentais
CE6	A motivação e problema apresentados na introdução não é condizente com a presente pesquisa
CE7	A conclusão do artigo não condiz com o que foi apresentado na introdução

Dados Abertos Ligados em seu portal de dados, data.gov.uk. Existem alguns artigos notáveis sobre a experiência britânica ligando dados governamentais. Sheridan and Tennison (2010) apresentam como o governo do Reino Unido trabalhou na criação de guias e recomendações para a implementação de Dados Abertos Ligados, pois apesar dos padrões robustos e maduros, a comunidade não estava pronta para o governo adotar a tecnologia em larga escala. Os principais direcionamentos foram em relação ao projeto de URIs, abordagens para versionamentos e fornecimento de informações de proveniência dos dados. Já Shadbolt et al. (2012) trata sobre o processo de levar Dados Abertos Governamentais publicados no data.gov.uk para o que eles chamam de *Linked Data Web* (LDW) seguindo as abordagens do projeto EnAKTing. Os principais desafios apresentados são relacionados à descoberta e migração dos dados, a integração de dados abertos governamentais para LDW, e a busca e exploração por dados em Dados Abertos Ligados. Os autores concluem que ainda existem muitos desafios em aberto em todas as pontas (governo, comunidade técnica e cidadãos) para uma ampla adoção desses padrões.

Existem alguns trabalhos que olham para as iniciativas de Dados Abertos Governamentais e Dados Ligados no contexto brasileiro. Ainda no início do amadurecimento das iniciativas brasileiras de dados abertos, Breitman et al. (2012) traz um histórico das iniciativas já realizadas no país a partir de 2009 e expõe um problema da publicação de dados brasileiros na *LOD Cloud*, que é a dificuldade em ligar dados brasileiros com dados internacionais devido às diferenças de língua. Para contornar esse problema é sugerida a utilização de vocabulários estabelecidos internacionalmente para mapeamento dos

conceitos presentes em conjuntos de dados nacionais e para apoiar esse mapeamento é proposto um processo e ferramenta chamado StdTrip. Matheus et al. (2015) fazem uma revisão e avaliação dos portais de dados abertos em diferentes esferas de governo (federal, estadual e municipal) utilizando como métrica os oito princípios de OGD e as cinco estrelas dos dados abertos. Mesmo sendo uma pesquisa apresentada em 2015, podemos perceber que a evolução de 2015 até o atual momento não foi alta, pois portais brasileiros ainda pecam em oferecer dados primários e completos e em atingir o maior nível de maturidade na publicação de dados abertos. Ainda dentro do cenário brasileiro, o trabalho de De Oliveira and Silveira (2018) apresenta uma revisão atualizada das iniciativas de OGD brasileiras por meio de um mapeamento sistemático de trabalhos relacionados ao assunto. Diferentemente de Breitman, esse trabalho não possui foco em Dados Abertos Ligados. Ele apresenta um panorama geral do estado da arte das iniciativas de OGD brasileiras. Os resultados encontrados mostram que boa parte dos trabalhos está voltada para a área da transparência, o que parece ser um reflexo do cenário político vivido pelo país nos últimos anos. Nesse artigo, também são encontrados trabalhos relacionados abordando outros temas como criminalidade, biodiversidade, saúde, educação e meteorologia. Por meio da revisão os autores também apontam os principais problemas encontrados pelos pesquisadores ao trabalhar com iniciativas brasileiras de OGD. O principal problema apontado é a falta de uma padronização na disponibilização dos dados, mas também existem outros problemas como falta de atualização, dados agregados e que possuem “guardiões” (*gatekeepers*), e dados publicados em formatos que não são de fácil leitura por máquinas ou proprietários.

Alguns trabalhos brasileiros adotam a abordagem de aplicar Dados Abertos Ligados a Dados Abertos Governamentais que não estão disponibilizados dessa maneira, seja através da criação de aplicações ou simplesmente da transformação de dados que estão disponíveis em CSV como RDF e com ligações a outros conjuntos de dados. de Oliveira Araújo et al. (2015) apresentam uma ontologia para o Orçamento Federal Brasileiro focado nos gastos, publicando-os como um vocabulário RDF e disponibilizando um *endpoint* SPARQL para consultas. Apesar de visarem atingir a quinta estrela no esquema de implementação dos dados ligados, esse trabalho alcança apenas até a quarta estrela, pois não executa a ligação com fontes externas. O trabalho de Araújo and de Souza (2011) conecta diferentes fontes de dados governamentais para criar um conjunto de dados mais completo sobre os políticos brasileiros, disponibilizando-os por meio de um

website. Ele permite que usuários não-técnicos se beneficiem do trabalho realizado, permitindo a consulta e visualização dos dados sem a necessidade de utilização de *end-points*, que não são triviais para o usuário final. Os autores aplicam o reuso de ontologias quando possível, utilizando algumas bastante difundidas como FOAF² e SKOS³. Na escala de publicação dos dados abertos, esse trabalho se enquadra no grau mais alto, pois além de disponibilizar os dados em RDF, também é feita a ligação com outras bases em RDF como DBPedia e GeoNames. Brandt et al. (2018) também trabalham com dados sobre políticos, com o objetivo de publicar dados referentes aos deputados brasileiros como Dados Abertos Ligados, fazendo extenso reuso de vocabulários estabelecidos como FOAF e Dublin Core⁴ para isso. Essa pesquisa tem semelhança com a proposta desta dissertação, no sentido de publicar dados abertos governamentais brasileiros como Dados Abertos Ligados. No entanto, não é seguida uma metodologia clara em relação as etapas realizadas, item que buscamos como uma das principais contribuições. Na área da saúde, os trabalhos de Freitas et al. (2017) e Lopes et al. (2016) descrevem problemas de integração de diferentes bases de dados do Sistema Único de Saúde (SUS). Os trabalhos utilizam uma abordagem baseada em ontologia e Dados Abertos Ligados com o objetivo de criar um sistema de apoio a decisão para risco de morte infantil (Freitas et al., 2017) e de propor um *framework* para criação de *Linked Data Mashup* (Lopes et al., 2016). Já o trabalho de Verona et al. (2019) busca entender as relações de poder entre pessoas e empresas condenadas pelo crime de escravidão no Brasil com políticos. Isso é feito através da publicação como Dados Abertos Ligados e ligação entre conjuntos que trazem os condenados com dados sobre doadores de campanhas políticas.

Outros trabalhos relacionados a esta pesquisa referem-se à classificação e proposta de ações e metodologias para a evolução na publicação de OGD através da utilização de Dados Abertos Ligados. Kalampokis et al. (2011) propõem um modelo de classificação que leva em consideração a tecnologia na qual o dado é publicado e o “modelo organizacional” que o publica, ou seja, se é centralizado num único órgão/portal ou se está distribuído em diferentes locais por diferentes órgãos. Os autores também propõem uma arquitetura para a publicação dos dados como Dados Abertos Ligados composta de quatro etapas. Eles fazem a utilização das tecnologias da Web Semântica para a publicação e aplicam essa arquitetura em um exemplo fictício. Apesar de ter

²<http://www.foaf-project.org/>

³<http://www.w3.org/2004/02/skos/core>

⁴<https://www.dublincore.org>

pontos em comum com nosso trabalho, como a arquitetura fazer uso de ontologias, o que é apresentado no protótipo trata da representação de alguns “objetos” especificamente (no exemplo, uma escola), enquanto em nossa abordagem buscamos trabalhar com conjuntos de dados completos, transformando-os em Dados Abertos Ligados.

Um dos ganhos da aplicação de tecnologias da Web Semântica é justamente dar um significado semântico e maior contexto aos dados. Hoxha and Brahaj (2011) aplicam essa abordagem para resolver problemas relacionados à dificuldade de integração entre dados de diferentes fontes mantendo seus significados. Uma ontologia para representação de conjuntos de dados estatísticos é criada e os dados são publicados como Dados Abertos Ligados, fazendo ligação com outros conjuntos de dados sem a perda de significado. Também é feita a disponibilização dos dados para consulta por meio de *endpoint* SPARQL e APIs REST. A abordagem apresentada segue os passos do que propomos em nosso trabalho, no entanto não há detalhamento sobre o processo de criação da ontologia ou da implementação das etapas propostas. Em uma revisão da literatura Milić et al. (2018) buscam identificar como governos oferecem dados e serviços inovadores para os cidadãos através da aplicação das tecnologias da Web Semântica.

Klein et al. (2016) também propõem uma metodologia para a publicação de dados abertos como Dados Abertos Ligados, mas com um foco no registro das etapas do processo de um projeto de publicação de dados como Dados Abertos Ligados. Tal registro é utilizado como forma de estruturar melhor esse processo e servir como um referencial para estimação de projetos futuros e de anotação sobre conhecimento do projeto. O trabalho de da Silva (2018) vai no caminho de prover uma ferramenta robusta para a realização do processo de triplificação, focado na etapa de extração, transformação e carga (ETL) dos dados. O trabalho de De Mendonça et al. (2013) se preocupa em como avaliar a qualidade e confiança em Dados Abertos Ligados e para isso propõe uma arquitetura e *framework* para a captura e geração de dados de proveniência ao longo do ciclo de publicação de um conjunto de dados como Dados Abertos Ligados, conectando os dados publicados com os dados de proveniência. Apesar de serem trabalhos que se preocupam com o processo de publicação de Dados Abertos Ligados, nenhum desses possui como objetivo o oferecimento de um passo-a-passo detalhado e a criação de ontologias como parte de suas propostas, o que os difere desta pesquisa.

O trabalho de Fleiner (2018) busca motivar governos a utilizarem Dados Abertos Ligados através da proposta de um modelo de ciclo de vida para iniciativas de Dados

Abertos Ligados Governamentais e da definição, sob o aspecto tecnológico, das etapas necessárias para que se vá do modelo “tradicional” da publicação de dados para uma efetiva abertura no mais alto grau de maturidade. Assim como nosso trabalho esse artigo busca trazer uma proposta para a publicação de OGD como Dados Abertos Ligados para servir de guia e facilitar esse trabalho por parte dos governos. Os diferenciais da nossa proposta estão na criação de um guia detalhado para cada uma dessas etapas descritas, facilitando assim sua utilização e aplicação por órgãos governamentais e aplicação em um conjunto de dados reais de um governo (no caso, do governo brasileiro).

Carbonaro (2019) sugerem a utilização de tecnologias da Web Semântica e Dados Ligados como forma de se criar um *framework* padronizado para representação de conceitos na criação de políticas e tornar possível a comparação e agregação de dados governamentais de diferentes órgãos e esferas, nacionais ou internacionais. Para ilustrar essa contribuição, o artigo apresenta três estudos de casos de situações nas quais foram aplicadas a publicação de dados utilizando Web Semântica com uma metodologia baseada em ontologias. Baseado nisso, os autores acreditam que pode se potencializar a geração de valor e melhorar a prestação de serviços para a sociedade.

É válido ressaltar que não somente os dados governamentais ou relacionados a serviços públicos podem ser publicados como Dados Abertos Ligados e enriquecer a Web Semântica. Pattuelli et al. (2015) apresentam uma perspectiva para a criação de ontologias focadas na publicação de Dados Abertos Ligados para o domínio da música por meio do projeto *Linked Jazz*, no qual é construída uma ontologia relacionada ao universo do jazz e é feita a ligação dos conceitos como Dados Abertos Ligados. Yu et al. (2012) apresentam um caso do universo de multimídia, aplicando ontologias e Dados Ligados para criação e publicação de anotações semânticas em vídeos, focado no cenário do ensino a distância (*e-learning*). Esses exemplos nos mostram que os princípios e tecnologias da Web Semântica podem ser aplicados em diferentes tipos de domínios e dados e trazer ganhos para quem os publica e para quem fará uso.

As Tabelas 3 e 4 apresentam um resumo das características dos trabalhos listados nesta seção divididos em duas categorias: Base Teórica e Aplicação. Os trabalhos classificados como Base Teórica são aqueles em que o tema central não são aplicações de Dados Abertos Ligados em si, mas sim que trazem conceitos e estudos relacionados a esse tema. Os critérios utilizados nas tabelas para comparação são os seguintes: (C1) Apresenta metodologia para publicação de Dados Abertos Ligados? (C2) Faz a

publicação de Dados Abertos Ligados? (C3) Cria ontologias para apoiar publicação de Dados Abertos Ligados? (C4) Atinge 5ª estrela dos Dados Abertos Ligados? Para trabalhos que são aplicações, a coluna “Comparação com proposta desse trabalho” traz um comentário descrevendo de forma sucinta como cada trabalho se compara com o que é proposto nessa dissertação.

Tabela 3 – Comparação entre os trabalhos relacionados a essa dissertação

Título	Categoria	C1	C2	C3	C4	Comparação com proposta desse trabalho
Linking UK Government Data (Sheridan and Tennison, 2010)	Base Teórica	Não	Não	N/A	N/A	N/A
Aumentando a Transparência do Governo por Meio da Transformação de Dados Governamentais Abertos em Dados Ligados (Araújo and de Souza, 2011)	Aplicação	Não	Sim	Sim	Sim	Faz publicação de dados do domínio eleitoral brasileiro assim como um dos nossos experimentos. No entanto não foca no processo realizado para a publicação como dados abertos ligados, não servindo como guia como é nossa proposta
A classification scheme for open government data: towards linking decentralised data (Kalampokis et al., 2011)	Aplicação	Sim	Sim	Sim	Sim	O trabalho apresenta apenas um protótipo com exemplos fictícios e parte de algumas premissas para facilitar implementação, e nós fazemos uso de dados reais. A metodologia proposta não é bem detalhada como a proposta nessa dissertação
Open Government Data on the Web: A Semantic Approach (Hoxha and Brahaj, 2011)	Aplicação	Sim	Sim	Sim	Sim	Apesar de adotar uma abordagem de etapas semelhantes, não há detalhamento sobre como conduzir as etapas, que é nossa principal contribuição
Linked Open Government Data: Lessons from Data.gov.uk (Shadbolt et al., 2012)	Base Teórica	Não	Não	N/A	N/A	N/A
Open Government Data in Brazil (Breitman et al., 2012)	Base Teórica	Não	Não	N/A	N/A	N/A
Using Linked Data to Annotate and Search Educational Video Resources for Supporting Distance Learning (Yu et al., 2012)	Aplicação	Não	Sim	Sim	Sim	O foco do trabalho é na aplicação desenvolvida para anotações em vídeo como Dados Abertos Ligados. Faz uso de ontologia para isso. No entanto ele não tem foco no processo de publicação como damos em nosso trabalho
LOP - Capturing and Linking Open Provenance on LOD cycle (De Mendonça et al., 2013)	Base Teórica	Sim	Sim	Não	Sim	N/A
Brazil Towards Government 2.0: Strategies for Adopting Open Government Data in National and Subnational Governments (Matheus et al., 2015)	Base Teórica	Não	Não	N/A	N/A	N/A
Ontology Building for Linked Open Data: A Pragmatic Perspective (Pattueli et al., 2015)	Aplicação	Sim	Sim	Sim	Sim	Foco é na aplicação desenvolvida, o Linked Jazz, diferente do nosso trabalho que foca na metodologia e nas aplicações como forma de validá-la. As etapas aplicadas são semelhantes
The Brazilian Federal Budget Ontology – A Semantic Web Case of Public Open Data (de Oliveira Araújo et al., 2015)	Aplicação	Não	Sim	Sim	Não	Publica dados abertos governamentais brasileiros mas não apresenta metodologia para o processo e não chega à quinta estrela na escala de Berners-Lee, diferente do que propomos que visa atingir esse grau

Tabela 4 – Comparação entre os trabalhos relacionados a essa dissertação (continuação)

Título	Categoria	C1	C2	C3	C4	Comparação com proposta desse trabalho
Towards a Linked Data Publishing Methodology (Klein et al., 2016)	Aplicação	Sim	Não	N/A	N/A	O foco da metodologia apresentada no trabalho é na documentação de projetos de publicação de Dados Abertos Ligados, diferente do nosso trabalho que foca no processo técnico de desenvolvimento
A Framework for Creation of Linked Data Mashups: A Case Study on Healthcare (Lopes et al., 2016)	Aplicação	Não	Não	Sim	Não	O foco desse trabalho é na publicação de Linked Data Mashups e na proposta de um framework para essa criação e visualização. Nossa proposta segue uma linha mais simples, apenas de publicação dos dados sem detalhar aplicações. Além disso o trabalho também não detalha metodologia para publicação de Dados Abertos Ligados como o nosso se propõe
Using Linked Data in the Data Integration for Maternal and Infant Death Risk of the SUS in the GISSA Project (Freitas et al., 2017)	Aplicação	Sim	Sim	Sim	Sim	O trabalho é focado na aplicação criada e não em prover um guia para publicação de Dados Abertos Ligados por governos. A ligação com outros conjuntos é feita apenas com conjuntos publicados dentro da aplicação proposta e não com outros conjuntos de dados da web
Modelo de dados abertos conectados para informação legislativa (Brandt et al., 2018)	Aplicação	Não	Sim	Não	Sim	Pesquisa muito aderente com objetivo de publicar dados abertos governamentais brasileiros como Dados Abertos Ligados. No entanto não segue uma metodologia tão clara em relação as etapas realizadas, que é onde contribuimos ao trazer um guia das etapas a serem seguidas para publicação como Dados Abertos Ligados
Semantic Technologies in e-government: Toward Openness and Transparency (Milić et al., 2018)	Base Teórica	Não	Não	N/A	N/A	N/A
Linking of open government data (Fleiner, 2018)	Base Teórica	Sim	Não	N/A	N/A	N/A
Open government data in Brazil: A systematic review of its uses and issues (De Oliveira and Silveira, 2018)	Base Teórica	Não	Não	N/A	N/A	N/A
ETL4LOD+: Evolução do Suporte ao Ciclo de Publicação de Dados Conectados (da Silva, 2018)	Aplicação	Sim	Sim	Não	Sim	É focado na apresentação da ferramenta ETL4LOD+ que pode ser usada em algumas etapas do ciclo de publicação como LOD. Diferente do nosso trabalho, a ferramenta e proposta fazem uso de ontologias mas não propõem a criação de uma para apoiar o processo
Linked data and semantic web technologies to model context information for policy-making (Carbonaro, 2019)	Base Teórica	Não	Não	N/A	N/A	N/A
Using Government Data to Uncover Political Power and Influence of Contemporary Slavery Agents in Brazil (Verona et al., 2019)	Aplicação	Não	Não	Não	Não	O trabalho faz a publicação de conjuntos de dados como Dados Abertos Ligados mas não apresenta metodologia para isso e não cria uma ontologia como parte do processo, como nós propomos

4- Metodologia

As etapas que serão apresentadas têm como objetivo servirem como um guia de como fazer a publicação de dados disponibilizados em formatos mais tradicionais (planilhas, arquivos CSV) como Dados Abertos Ligados, e assim atingir o nível mais alto de maturidade na publicação de dados abertos segundo a escala das 5 estrelas de Berners-Lee. Essas etapas são um caminho para que dados que atualmente estão no terceiro degrau da escala possam atingir o quinto degrau de maneira estruturada, permitindo assim que qualquer usuário ou desenvolvedor com acesso a esses dados possa fazê-lo.

São propostas quatro etapas para atingir o objetivo: (i) identificação, análise e integração dos conjuntos de dados; (ii) desenvolvimento de uma ontologia de domínio para representar os dados que serão publicados como Dados Abertos Ligados; (iii) publicação dos dados como RDF a partir da ontologia desenvolvida e ligação com outros recursos na web; e (iv) disponibilização de um *endpoint* SPARQL para consulta dos dados em RDF. As etapas são descritas em detalhes nas seções subsequentes.

4.1- Identificação, Análise e Integração dos Dados

A primeira etapa consiste na identificação e análise dos conjuntos de dados que se pretende publicar como Dados Abertos Ligados. Essa etapa é bastante importante para que se entenda qual o conhecimento que está sendo representado nos dados e como os dados podem se interligar (no caso da publicação de múltiplos conjuntos de dados) (Ristoski and Paulheim, 2016). O conhecimento do domínio é importante também para a etapa seguinte, de definição e criação de uma ontologia de domínio, pois a ontologia deve refletir os conceitos presentes nos dados selecionados, ser um reflexo do “mini-mundo” dos dados de trabalho ao invés de uma ontologia que se propõe a representar o domínio maior onde os dados estão inseridos.

Além de entender quais conceitos, propriedades e relacionamentos entre os

dados que serão representados na ontologia, a análise deles também apoia a criação da estratégia de integração e desenvolvimento da ontologia. Essa estratégia leva a uma tomada de decisão sobre realizar a integração dos conjuntos de dados em um só e ter uma única ontologia ou ter várias ontologias que representem separadamente cada conjunto do domínio que se está trabalhando.

As decisões tomadas na etapa de integração têm reflexo em todo o processo, pois impactam diretamente em como será feita a publicação dos dados como Dados Abertos Ligados no formato RDF. Por exemplo, é possível transformar dados de CSV diretamente para RDF, assim como também é possível transformar uma base relacional para RDF, mas as técnicas e ferramentas utilizadas para cada tipo são diferentes.

4.1.1 Identificação

A identificação dos dados é o processo de pesquisa e mapeamento de quais dados serão publicados como Dados Abertos Ligados. O ponto de partida pode ser um conjunto ou arquivo de dados específico, e o objetivo pode ser somente o de publicar aquele conjunto diretamente como Dados Abertos Ligados. Contudo, vão existir casos em que o desejo é de se representar um domínio mais amplo que será composto por vários conjuntos de dados espalhados em diferentes locais e arquivos.

Para o segundo caso, a etapa de identificação é bastante importante, pois nela devem ser feitas as buscas por todos os conjuntos de dados que possam ter relação com o domínio e que possam agregar conhecimento a ser representado. Essa busca deve ser feita em mecanismos de pesquisa e em sites confiáveis do domínio a ser representado. Para dados governamentais, que é o foco desse trabalho, recomenda-se buscar pelos dados nos portais de dados abertos dos governos e diretamente em sites das entidades responsáveis pela publicação e manutenção dos dados.

4.1.2 Análise e Integração

Após a identificação dos conjuntos de dados a serem utilizados, a etapa seguinte consiste na análise e integração deles, quando necessário. Essa etapa do processo tem um nível de esforço e complexidade variável em relação ao número de conjuntos de dados que se deseja publicar e como isso será feito.

Em um cenário com apenas um conjunto de dados ou vários conjuntos que serão publicados individualmente (sem junção das suas informações de maneira estruturada), o processo corresponde a apenas decidir como esses dados serão armazenados e, havendo necessidade, realizar o processo de integração/transformação para o formato escolhido. Como exemplo, o conjunto de dados original pode estar em um arquivo CSV, mas opta-se por armazená-lo em um banco de dados relacional. Nesse caso é necessário realizar um processo para transformar esses dados de CSV para uma estrutura de tabelas relacionais. Como as possibilidades são muito grandes, não iremos entrar nos detalhes de como executar o processo. A recomendação, dada a experiência com esse trabalho é de tentar manter o processo simples para que se ganhe velocidade.

Já em um cenário onde se deseja publicar dados de um domínio mais amplo e esse conhecimento está armazenado em diversos conjuntos de dados, existe a necessidade de integrar esses diferentes conjuntos para unificação do conhecimento do domínio, tornando o processo mais trabalhoso. É bastante importante entender como os dados que estão em diferentes arquivos se ligam entre si, ou seja, como podemos identificar que um registro presente em um arquivo é complementar a outro presente em outro arquivo. Com a experiência obtida com os arquivos de dados sobre as eleições brasileiras relatados nesse trabalho, foi identificada como uma boa prática a criação de um modelo entidade-relacionamento para mapeamento dessas relações entre os dados dos diferentes arquivos. Independente dos dados serem convertidos para um banco de dados relacional, o modelo apoia na identificação das classes do domínio e como elas se relacionam, indicando quais campos são comuns a cada arquivo e identificando os registros em cada um.

Mesmo em publicações que envolvam vários arquivos, não necessariamente é preciso fazer uma integração para que todos convirjam em um conjunto de dados único. Mais uma vez, a recomendação é fazer o que for mais simples. Para o caso de uso das eleições deste trabalho, por exemplo, os arquivos foram mantidos separados e carregados

na ontologia utilizando a ferramenta Cellfie do Protégé (que será explicada mais adiante) para leitura direta dos arquivos como planilhas.

Vale ressaltar que as escolhas feitas nessa etapa influenciam como será a publicação dos dados como Dados Abertos Ligados e quais ferramentas serão utilizadas. O processo utilizado como padrão neste trabalho foi manter os dados em seu formato original (arquivos CSV), construir a ontologia e populá-la utilizando a ferramenta Cellfie, e publicar diretamente um arquivo XML/RDF produto da ontologia em um servidor web para que fosse possível fazer o acesso.

4.2- Desenvolvimento da Ontologia

O objetivo dessa etapa é a construção de uma ontologia leve (*lightweight ontology*) que sirva como base semântica para apoiar o processo de publicação dos dados como Dados Abertos Ligados. A construção de uma ontologia como etapa para a publicação de Dados Abertos Ligados é importante para que se descreva formalmente as classes, instâncias e relacionamentos referentes aos dados que se deseja publicar, visto que apenas a utilização do padrão RDF não permite esse nível de descrição do conhecimento (Pattueli et al., 2015). Através da ontologia serão formalizadas as estruturas que as triplas RDF terão.

Como detalhado na Seção 2.2, para atingir esse objetivo, foi utilizado nesta pesquisa um fluxo de trabalho que faz a junção de etapas de duas metodologias consolidadas para o desenvolvimento de ontologias: a metodologia NeOn e a *Ontology Development 101*. As etapas realizadas para o desenvolvimento da ontologia são: levantamento de requisitos; identificação e reutilização de recursos ontológicos e não-ontológicos; conceitualização da ontologia; e implementação da ontologia. Os detalhes sobre como as metodologias utilizadas abordam a execução de cada etapa e subetapas são descritas na Seção 2.2.2. A recomendação é que essas etapas sejam seguidas para o desenvolvimento da ontologia como parte do processo de publicação de dados como Dados Abertos Ligados. Nesta seção são descritas como as etapas foram aplicadas na execução desse trabalho e também são detalhadas as adaptações que foram feitas à proposta original de algumas delas para que o desenvolvimento ficasse mais adequado ao objetivo.

4.2.1 Levantamento de Requisitos

O levantamento de requisitos da ontologia é conduzido utilizando como base a técnica das Questões de Competência (QCs). As questões foram desenvolvidas utilizando uma abordagem *middle-out* nesse trabalho, mas outras podem ser aplicadas. Nessa abordagem as questões que são consideradas importantes são escritas independente de sua complexidade e são compostas e decompostas para que se obtenha questões simples e abstratas (Suárez-Figueroa et al., 2012). Antes de iniciar o desenvolvimentos das questões é importante definir o cenário motivador, escopo, linguagem de desenvolvimento, usuário e uso pretendido da ontologia, de forma a garantir que as Questões de Competência, que determinam os requisitos da ontologia, sejam capazes de atender essas definições. Após sua elaboração, é necessário que os requisitos sejam validados e, caso haja a necessidade, priorizados para que a ontologia desenvolvida atenda aos mais importantes. A Metodologia NeOn propõe um ciclo de identificar → agrupar → validar até que se chegue a uma versão final dos requisitos. Contudo, na experiência obtida desenvolvendo esse trabalho, notou-se que esse ciclo deve ocorrer ao longo de todo o desenvolvimento da ontologia, pois podem ser descobertos novos requisitos ou a necessidade de alteração de requisitos já mapeados em etapas mais avançadas, como na própria implementação. Tais ajustes devem ser refletidos no conjunto de requisitos e nos artefatos que são produtos dele, como o pré-glossário de termos da ontologia.

A partir das Questões de Competência e suas respostas são extraídos os termos da ontologia e suas frequências para compor o pré-glossário de termos da ontologia. Esse artefato é bastante importante pois serve como base para o modelo conceitual da ontologia e apoia a busca por recursos ontológicos, não-ontológicos e conjuntos de Dados Abertos Ligados que serão reaproveitados na ontologia. Apesar da metodologia NeOn propor que sejam mapeados os termos e frequências das respostas das Questões de Competência, por questão de simplificação de escopo, nesse trabalho esses termos não foram mapeados, pois o volume possível de respostas contidas nos conjuntos de dados utilizados nos experimentos é bastante alto. Essa limitação não impede o desenvolvimento do modelo conceitual, que não precisa de detalhes a nível de indivíduos (que são os objetos que normalmente se obtém nas respostas das QCs).

Além da adaptação proposta no ciclo de validação dos requisitos da ontolo-

gia, também propomos nesse trabalho uma abordagem diferente no que diz respeito a utilização de usuários e especialistas de domínio no levantamento dos requisitos funcionais e não-funcionais da ontologia. A metodologia NeOn diz que devem ser entrevistados usuários e especialistas do domínio para que se obtenha um conjunto inicial dos requisitos que a ontologia deve atender (Suárez-Figueroa et al., 2009). A abordagem aqui utilizada propõe que os requisitos da ontologia sejam determinados através da análise dos dados que se deseja publicar como Dados Abertos Ligados, obtendo o conhecimento do domínio que será representado a partir dessa análise. Entende-se que a utilização de especialistas do domínio pode agilizar o processo e contribuir para a construção de uma ontologia mais robusta. No entanto, muitas vezes o desenvolvedor da ontologia não tem acesso a esses especialistas ou pode não ter a disponibilidade para a condução de rodadas de entrevista com eles. Nesse caso, é necessário que esse levantamento seja feito utilizando outros meios. A proposta é que isso seja feito pelo próprio responsável pelo processo de publicação dos dados, que deve ser quem identifica e analisa os dados conforme descrito na Seção 4.1.

Ao fim do levantamento de requisitos, deve-se ter preenchido o Documento de Especificação de Requisitos da Ontologia (ORSD) para que ele seja utilizado como base para as etapas seguintes do desenvolvimento, que são a busca por recursos ontológicos e não-ontológicos para reutilização e criação de um modelo conceitual da ontologia.

4.2.2 Reutilização de Recursos Ontológicos e Não-Ontológicos

A proposta da metodologia NeOn para a identificação e reutilização de recursos ontológicos tem como foco a identificação de ontologias completas que possam ser reaproveitadas para a ontologia que está sendo desenvolvida, mesmo que as ontologias identificadas não sejam usadas por completo, pois podem ser utilizados somente módulos que façam mais sentido para a aplicação. Já para os recursos não-ontológicos, o objetivo é que eles sejam identificados para que possa se eleger quais os mais adequados para passarem por um processo de reengenharia para ontologias e serem reaproveitados no desenvolvimento. Os recursos ontológicos e os não-ontológicos que passarem pelo processo de reengenharia para ontologia e identificados para reutilização devem ser

utilizados no momento da implementação da ontologia, através de sua importação ou ligação utilizando suas URIs.

Como o foco deste trabalho é na publicação de Dados Abertos Ligados, acreditamos que essa etapa do processo de desenvolvimento de uma ontologia pode ser aproveitada também para que possam ser identificados recursos que podem ser utilizados como indivíduos da ontologia ou ligados a indivíduos da ontologia através de predicados como o `owl:sameAs` para que se configure a publicação como Dados Abertos Ligados. Essa ligação a nível de indivíduo é necessária para que o conjunto de dados possa ser considerado ligado a outros conjuntos (se feita através da URI dos recursos) (Heath and Bizer, 2011).

A melhoria proposta à essa etapa da metodologia NeOn é buscar também por recursos que estejam fora de ontologias, em conjuntos de dados publicados como Dados Abertos Ligados. Para isso, propõe-se a utilização do pré-glossário de termos da ontologia, extraído das Questões de Competência, para que se busque indivíduos listados nas respostas às Questões de Competência em repositórios conhecidos de Dados Abertos Ligados como a DBPedia¹ e a Linked Open Vocabularies². Dessa forma, podem ser identificados indivíduos que não estejam dentro de ontologias mas que estão publicados em conjuntos de dados disponibilizados na Web dos Dados, e através da ligação com esses indivíduos, seja aproveitando-os diretamente na ontologia, seja fazendo ligações externas, atingir a quinta estrela dos dados abertos na escala de Berners-Lee.

Nos experimentos conduzidos nesse trabalho foi decidido, por simplificação, não fazer a busca por indivíduos para ligação com os indivíduos da ontologia por conta do aumento de escopo que isso traria. Mas para alguns indivíduos utilizados como exemplo, o processo utilizado para identificação dos recursos que poderiam ser ligados a eles foi o proposto nesta seção, o que corrobora com seu funcionamento. Vale ressaltar que para ontologias com muitos indivíduos esse processo de busca manual por conjuntos de dados para ligação (conhecido como *dataset interlinking*) pode ser muito custoso em tempo e esforço. Para otimizar esse processo podem ser utilizadas abordagens de busca e identificação sistemáticas de recursos como as propostas por Leme et al. (2013), Caraballo et al. (2014) e Arturo et al. (2015).

¹<https://wiki.dbpedia.org/>

²<https://lov.linkeddata.es/dataset/lov/>

4.2.3 Conceitualização da Ontologia

Para o processo de conceitualização da ontologia foi utilizado como guia a *Ontology Development 101*, como detalhado na Seção 2.2.1. Na Seção 2.2.4 é descrito o processo proposto para que se crie um modelo conceitual da ontologia antes de iniciar a sua implementação. Esse trabalho contribui com duas adaptações ao processo, que são apontadas quando descrevemos a etapa mas que serão destacadas nesta seção.

A primeira adaptação consiste em utilizar as etapas de listar os termos da ontologia, definir classes e hierarquias e definir as propriedades como uma forma de revalidar o pré-glossário de termos produzido utilizando etapas da metodologia NeOn. O pré-glossário já traz os termos e suas classificações, mas conduzir essa reavaliação se mostrou valioso para começar a dar forma a como será a ontologia. O que está classificado no pré-glossário como conceitos, atributos e relacionamentos devem ser reclassificados como classes, hierarquias e propriedades para a ontologia. Os indivíduos se tornarão instâncias.

A segunda adaptação ao processo foi a inclusão de algo que não é proposto na metodologia, que é a criação de um diagrama de classes antes de iniciar a implementação da ontologia. Na experiência obtida durante o desenvolvimento desse trabalho a criação do diagrama facilitou bastante o processo de implementação da ontologia, pois durante sua concepção pode-se já pensar e modelar como seria a ontologia, iterando até chegar em um modelo final. Sem essas iterações no diagrama, os ajustes necessários só seriam percebidos na implementação, o que geraria um desgaste maior devido às alterações na estrutura da ontologia enquanto se implementa ela. Com o diagrama em mãos, a implementação consistiu basicamente em transpor o que estava representado para a ferramenta utilizada, que neste caso foi a ferramenta Protégé.

4.3- Publicação dos Dados como Dados Abertos Ligados

A publicação dos dados como Dados Abertos Ligados consiste na disponibilização dos dados como triplas no formato RDF de modo que os recursos possam ser acessados por meio de suas URIs únicas e que estejam ligados com outros recursos presentes na

web. Ontologias desenvolvidas no Protégé tem como linguagem padrão a OWL. Para tê-las no formato correto deve-se salvar a ontologia como um arquivo RDF/XML sem a necessidade de nenhum tipo de conversão mais complexa. Também é possível realizar a conversão de outros formatos como CSV, XML, JSON e formatos de bases relacionais para RDF utilizando-se de ferramentas abertas voltadas para esse propósito ³.

No processo aplicado neste trabalho é proposto que a ligação com conjuntos de dados externos à ontologia sejam feitos no momento de sua implementação, utilizando os recursos identificados e fazendo as conexões entre os indivíduos da ontologia e os recursos externos através da utilização das URIs desses recursos. Dessa forma, quando a ontologia for convertida para o formato RDF/XML, as ligações entre recursos, que são necessárias para que se configura a publicação como Dados Abertos Ligados, estará realizada. O modo mais comum de realizar essa ligação é através da utilização do predicado `owl:sameAs`, que indica que um indivíduo da ontologia é o mesmo que um recurso descrito em outra URI. Mas também é possível fazer essa ligação sem que haja a criação de um indivíduo diferente no domínio da ontologia, somente reaproveitando recursos externos casos eles existam.

Tendo os dados em um arquivo no formato RDF, a maneira mais simples de publicar os dados como Dados Abertos Ligados é disponibilizar o arquivo em um servidor web que permita acesso externo à Internet. Isso garantirá o acesso aos dados publicados, mas não permitirá que um conteúdo diferente seja entregue dependendo do tipo de chamada (se é uma chamada para humanos que seria mais bem entregue em HTML ou para uma máquina, melhor entregue em RDF) (Magalhães et al., 2011). Para isso podem ser utilizadas ferramentas como o Pubby⁴. Contudo, fazer essa negociação de conteúdo não é necessária para que um conjunto de dados publicado seja considerado como Dados Abertos Ligados e por conta disso não houve a utilização do Pubby ou qualquer outra ferramenta com esse intuito neste trabalho.

Como dito na Seção 4.2.2 durante o processo de identificação de recursos para reutilização nos experimentos desse trabalho não houve uma busca extensa de recursos a nível de indivíduo e, por conta disso, na implementação da ontologia não foram feitas muitas conexões com recursos externos, tendo sido feita a ligação apenas para alguns exemplos, assim havendo poucas conexões com outros conjuntos de dados. Essa foi uma decisão de simplificação de escopo por conta do volume de indivíduos existentes

³<https://www.w3.org/2001/sw/wiki/Tools>

⁴<http://wifo5-03.informatik.uni-mannheim.de/pubby/>

nas ontologias dos experimentos conduzidos. Para que um conjunto de dados possa ser considerado como um conjunto de Dados Abertos Ligados é necessário que sejam feitas essas ligações a nível de indivíduos.

Tendo sido feita a ligação dos indivíduos da ontologia com recursos externos, ao final dessa etapa pode-se dizer que a quinta estrela dos dados abertos é alcançada, pois todos os padrões da Web Semântica requeridos para isso terão sido aplicados. No entanto, além de atender aos requisitos específicos da publicação de Dados Abertos Ligados, também é uma boa prática que se disponibilize uma maneira de consultar os dados, o que é habitualmente feito através da disponibilização de *endpoints* SPARQL.

4.4- Publicação de *Endpoint* SPARQL

Essa etapa refere-se à disponibilização dos dados, a qual engloba a exportação completa da base de dados e a disponibilização de um meio para consulta aos dados utilizando *endpoints* SPARQL (Heath and Bizer, 2011).

Disponibilizar a consulta dos dados dessa maneira é uma abordagem técnica, mas trata-se de um padrão para consultas em RDF e é uma linguagem bastante conhecida na comunidade de Web Semântica (de Oliveira Araújo et al., 2015). Disponibilizando essa forma de consulta, passa a ser oferecida uma opção que está de acordo com os padrões da web semântica. Como uma melhoria e forma de dar maior flexibilidade, é desejado também a disponibilização de uma forma de consulta amigável, permitindo que usuários não técnicos possam se beneficiar da total abertura dos dados.

Para a criação dos *endpoints* deste trabalho, foi utilizada a ferramenta Apache Jena Fuseki⁵. A ferramenta possui uma boa documentação de configuração e de uso. Nela é possível fazer o *upload* da ontologia criada utilizando o arquivo OWL ou RDF/XML. As triplas são carregadas e é possível realizar consultas tanto por meio da interface visual da ferramenta quanto por meio de *endpoints* fazendo chamadas HTTP para o endereço do servidor web no qual a ferramenta está em execução. Como dito anteriormente, a linguagem SPARQL é a utilizada para se realizar as consultas às triplas.

⁵<https://jena.apache.org/documentation/fuseki2/index.html>

5- Resultados

Após a elaboração das etapas propostas para a publicação de dados abertos como Dados Abertos Ligados, houve a necessidade de se aplicar o que foi proposto como forma de validação e adequação das etapas. Por meio da implementação da metodologia proposta utilizando conjuntos de dados reais seria possível identificar melhorias nas etapas e evolução no processo como um todo. Para atingir esse objetivo, foram conduzidos dois experimentos com conjuntos de dados governamentais distintos.

Para o primeiro experimento foram utilizados dados pertencentes ao domínio eleitoral brasileiro, disponibilizados pelo TSE no seu Repositório de Dados Eleitorais¹. Com essa primeira aplicação pôde-se ter maior entendimento sobre dados abertos governamentais, principalmente em relação à forma como eles são disponibilizados. Por ser o primeiro experimento conduzido, ele apoiou a adequação das etapas seguidas, de acordo com os desafios encontrados durante a implementação.

O segundo experimento foi conduzido utilizando dados do domínio da saúde. As etapas para publicação dos dados como Dados Abertos Ligados foram aplicadas ao conjunto de dados de distribuição geográfica das Unidades Básicas de Saúde (UBS) em funcionamento no Brasil. Esses dados são publicados pelo Ministério da Saúde no Portal Brasileiro de Dados Abertos². Esse experimento permitiu que o processo fosse validado utilizando outro tipo de dado governamental, diferente do que foi utilizado para a adequação das etapas no primeiro experimento.

Ambos os experimentos seguiram as etapas propostas no Capítulo 4. As etapas descritas servem como um guia estruturado de passos a serem seguidos para que, a partir de um conjunto de dados publicado em formato tradicional, se chegue a um conjunto de dados publicado como Dados Abertos Ligados. A metodologia proposta visa mitigar o problema da não utilização de Dados Abertos Ligados por falta de um guia, pois ela justamente traz a proposta de ter etapas replicáveis e que facilitem esse processo. Contudo, os experimentos conduzidos não chegaram ao nível de poderem ser considerados publicados como Dados Abertos Ligados pois não foi feita uma ligação

¹<http://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais-1/repositorio-de-dados-eleitorais>

²http://www.dados.gov.br/dataset/ubs_funcionamento

extensa dos indivíduos com recursos externos presentes em outros conjuntos de dados e também não foi feita a publicação em um domínio exposto para a internet para acesso externo. Os experimentos seguiram esse caminho para simplificação de escopo.

Além dos conjuntos de dados diferentes, os experimentos apresentam abordagens e estratégias distintas que podem ser utilizadas dentro do processo para atingir o mesmo fim: a publicação como dados abertos ligados. Eles apontam como o processo proposto pode ser adaptado dependendo da natureza dos dados e da estratégia de integração utilizada. A utilização de conjuntos de dados diferentes para cada aplicação permite também mostrar que o processo é replicável independente do domínio.

Com esses experimentos espera-se evidenciar como órgãos brasileiros podem vir a aplicar os passos apresentados e assim ter seus dados totalmente de acordo com os padrões da Web Semântica e do governo eletrônico brasileiro³.

5.1- Dados das Eleições Brasileiras

5.1.1 Identificação, Análise e Integração dos Dados

Como dito anteriormente, os dados do TSE utilizados nesse trabalho estão disponíveis no Repositório de Dados Eleitorais, espaço que fica no próprio site do Tribunal, e encontram-se no formato CSV. No repositório é possível encontrar dados sobre os seguintes assuntos relacionados às eleições:

- Candidatos
- Comparecimento e abstenção
- Eleitorado
- Partidos
- Pesquisas eleitorais
- Prestação de contas eleitorais

³<http://eping.governoeletronico.gov.br/>

- Prestação de contas partidárias
- Processual
- Resultados

Para este trabalho foi definido que o escopo a ser representado seria o de uma eleição no Brasil. Dessa forma, a análise dos dados foi feita para entender, de acordo com os assuntos supracitados, quais seriam os conjuntos de dados que poderiam produzir conhecimento referente a esse escopo. Foram selecionados os conjuntos de dados dos assuntos Candidatos, Partidos, Prestação de contas eleitorais e Resultados. O assunto Eleitorado por si só representa uma porção significativa e isolada de estudo. Dessa forma, este assunto não foi inserido no escopo definido.

Os conjuntos de dados selecionados são compostos por múltiplos arquivos CSV que possuem dados de diferentes níveis (nacionais, estaduais e municipais). Para fins de instanciação desta pesquisa, apenas um subconjunto dos dados foi usado para a publicação como Dados Abertos Ligados, mas o trabalho desenvolvido é válido para todos os dados disponíveis, independentemente de seu nível. Especificamente, foi escolhida a representação das eleições presidenciais de 2018.

Como os dados necessários para representar uma eleição estão espalhados por diferentes arquivos, foi criado um modelo entidade-relacionamento (modelo ER) para fazer uma modelagem inicial das classes e seus relacionamentos, com o objetivo de entender e identificar como elas se conectavam entre si. A Figura 5 apresenta as entidades (sem seus atributos) e relacionamentos do modelo construído, identificadas nos conjuntos de dados. O modelo foi projetado utilizando como base a forma como os dados são disponibilizados nos arquivos, que podem apresentar uma diferença de como eles são interpretados conceitualmente. Um exemplo disso é na relação entre CANDIDATOS e ELEICOES, pois para cada eleição um candidato recebe um identificador único. Caso ele concorra em mais de uma eleição, ele terá identificadores diferentes em cada uma. No entanto, essa é uma relação 1 para muitos pois um candidato pode concorrer em mais de uma eleição com o mesmo identificador em casos onde há primeiro e segundo turno, pois cada turno possui identificadores distintos. A partir do modelo ER é possível ter um melhor entendimento das entidades e relacionamentos, facilitando a posterior modelagem de uma ontologia de modo a representar todo o conhecimento contido nos dados. O modelo completo pode ser visto em <https://git.io/JU2JR>.

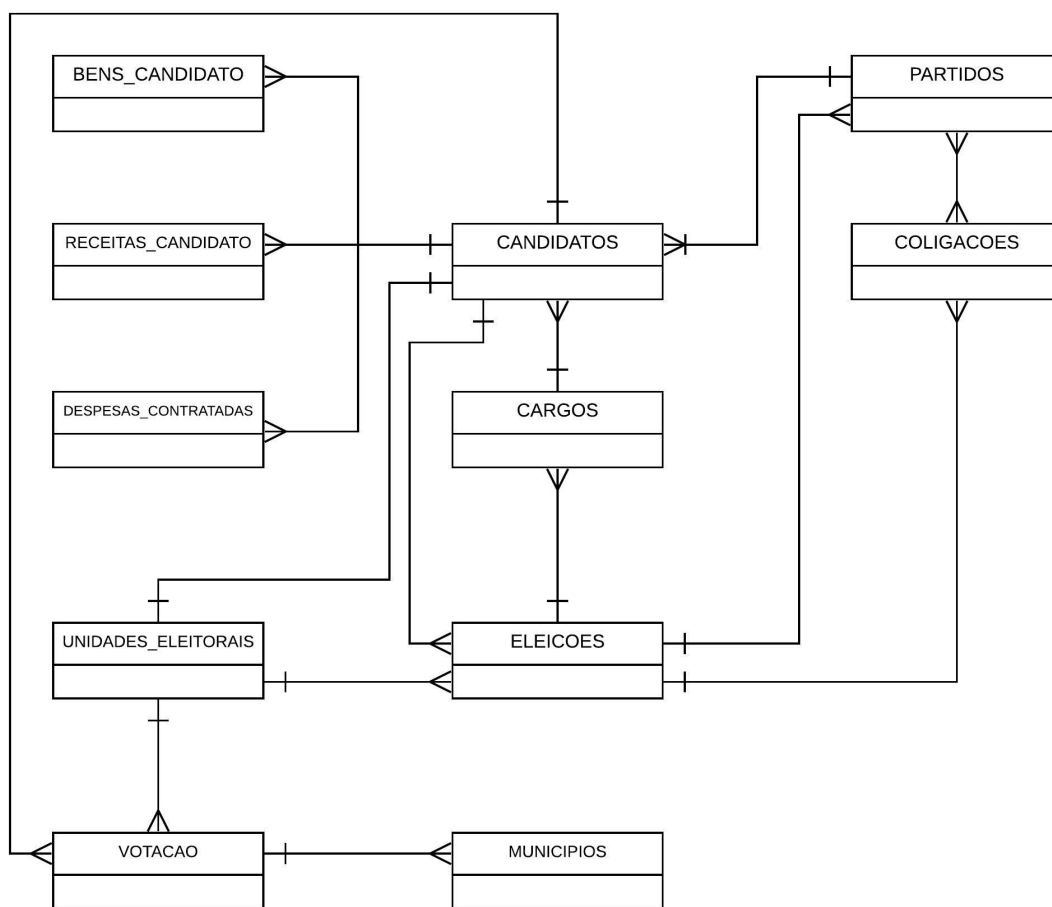


Figura 5 – Entidades e relacionamentos do modelo ER das Eleições Brasileiras

Como decisão de projeto, apesar da construção do modelo ER, foi decidido não realizar a integração dos dados usando um banco de dados relacional. A escolha foi manter os dados em arquivos CSV separados e popular a ontologia utilizando-os. Essa escolha foi feita para que se tenha maior agilidade no desenvolvimento, visto que a integração dos dados para um banco relacional demandaria uma etapa adicional no desenvolvimento e temos como objetivo manter o processo o mais simples possível.

5.1.2 Desenvolvimento da Ontologia

Nessa seção serão detalhadas as etapas seguidas para a construção da Ontologia das Eleições Brasileiras.

Levantamento de Requisitos

Cenário Motivador, Escopo e Linguagem A motivação para o desenvolvimento de uma ontologia das eleições no Brasil é prover um modelo semântico que possa ser usado como base para a interligação dos diferentes conjuntos de dados abertos relacionados às eleições, os quais são disponibilizados pelo TSE⁴. Atualmente, esses dados, encontram-se em bases separadas, o que dificulta sua integração e geração de conhecimento mais rico. Como descrito anteriormente, o objetivo é que a ontologia criada seja utilizada como base para o processo de publicação dos dados como Dados Abertos Ligados. A ontologia das eleições brasileiras deve prover um modelo que reflita a realidade dos dados disponíveis de maneira aberta sobre as eleições no Brasil, com foco nas próprias eleições (o processo em si) e nos candidatos a cargos políticos públicos. A ontologia é implementada utilizando a interface visual da ferramenta Protégé. Embora a ferramenta utilize o formato OWL (W3C OWL Working Group, 2012) como padrão, o resultado é exportado para um arquivo no formato RDF/XML.

Usuários e Uso Pretendido A proposta para a construção dessa ontologia é que ela possa ser utilizada por outros pesquisadores e usuários que tenham interesse em trabalhar com o aspecto semântico do domínio eleitoral brasileiro, dando a eles um recurso ontológico que possa ser aplicado tanto em iniciativas relacionadas a Dados Abertos Ligados quanto outras que possam ter ontologias como sua base. O próprio governo brasileiro pode fazer uso da ontologia desenvolvida nesse trabalho em iniciativas como o Repositório de Vocabulários e Ontologias do Governo Eletrônico⁵, utilizando-a como um padrão ou base para o assunto eleições. Cabe salientar que, o cidadão que não se encaixe em um dos perfis citados também pode se beneficiar dessa ontologia por meio de aplicações que a utilizem, pois sua criação visa apoiar e facilitar a publicação de dados abertos de maneira integrada, o que permite maior transparência e possibilidade de descoberta e interligação entre diferentes informações governamentais.

⁴<http://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais-1/repositorio-de-dados-eleitorais>

⁵<http://vocab.e.gov.br/>

Identificação de Requisitos Para a ontologia das eleições brasileiras, foi identificado um requisito não-funcional: a ontologia deve se referir a eleições ocorridas após a redemocratização (que ocorreu no Brasil em 1988). Esse requisito foi posto, pois a partir desse período que se passou a ter a estrutura eleitoral empregada ainda hoje no país. Para definição dos requisitos funcionais foram empregadas as Questões de Competência com uma abordagem *middle out*.

Apoiando-se na abordagem *middle out*, as QCs foram elaboradas considerando-se os aspectos importantes que a ontologia proposta deveria responder. As QCs desenvolvidas foram classificadas em três temas: eleição, candidato e partido. A Tabela 5 apresenta alguns exemplos de questões desenvolvidas. Ao todo foram desenvolvidas 25 Questões de Competência, divididas entre os três temas citados. Para o desenvolvimento das QCs não foram consultados especialistas do domínio (eleições). O conhecimento para elaboração das questões se deu por meio de pesquisa e análise sobre o domínio, principalmente sobre os conjuntos de dados abertos disponibilizados pelo TSE. A lista completa de questões está disponível de maneira aberta em <https://git.io/JU2Uf>.

Tabela 5 – Questões de Competência sobre as eleições no Brasil

ID	Categoria	Questão
CQ1	Eleição	Em que ano a eleição foi realizada?
CQ8	Candidato	Qual o nome do candidato?
CQ19	Candidato	Quais as fontes de receitas declaradas pelo candidato para sua campanha?
CQ23	Partido	Quais partidos disputaram a eleição?

Validação dos Requisitos O processo de validação dos requisitos, criados em forma de Questões de Competência, foi realizado em paralelo à elaboração das questões. Foram necessárias algumas iterações do conjunto de requisitos para validação dos mesmos e obtenção da lista de QCs.

Os critérios escolhidos para a validação dos requisitos foram: i) corretude, quando todos os requisitos de um conjunto representam funcionalidades da ontologia que está sendo desenvolvida, ii) consistência, obtida quando não há conflito entre os requisitos de um conjunto, iii) não-ambiguidade, para garantir que cada requisito tenha um significado único, que não cause confusão ou permita outros entendimentos para quem o ler e iv) realismo, que avalia se os significados dos requisitos fazem sentido no domínio que eles estão representando (Suárez-Figueroa et al., 2012).

De 35 Questões de Competência inicialmente elaborados, 10 foram excluídas após validação. Um exemplo de Questão de Competência que foi eliminada após validação foi a seguinte: “Quais as fontes de receita de um partido?”. Ao avaliá-la utilizando os critérios mencionados, foi notado que não eram respeitados os critérios de corretude e realismo. Ela não respeita o critério de realismo, pois essa questão não endereça um aspecto das eleições, mas sim do domínio político partidário, afinal, existe declaração de receita de um partido independentemente de haver eleição ou não. Com isso, não faria sentido ter essa representação na ontologia, que visa modelar a eleição em si. Logo, ela deixa de atender também ao critério de corretude, pois não mais representa uma funcionalidade da ontologia.

Priorização dos Requisitos Durante a criação das QCs houve uma preocupação em questões que fossem focadas no escopo da ontologia, a saber, o processo eleitoral e os candidatos. Após a categorização das questões, elas foram priorizadas de acordo com esse foco. Dessa maneira, não foram descartadas perguntas, pois a ontologia desenvolvida deve conseguir representar todas as questões elaboradas, atendendo todos os requisitos funcionais pensados na forma de QCs.

Extração de Terminologia e Frequências Para a ontologia das eleições foram considerados apenas os termos identificados nas perguntas das questões de competência, pois havia uma variação existente nas respostas de eleição para eleição. Foram identificados 30 termos. Estes foram ranqueados (maior para menor frequência) para formar o pré-glossário de termos da ontologia. A Tabela 6 apresenta os três termos de maior frequência. A relação completa dos termos pode ser vista em <https://git.io/JU2Uk>. A Figura 6 apresenta um extrato do Documento de Especificações de Requisitos da Ontologia produzido para esse experimento e preenchido ao longo do levantamento de requisitos.

Tabela 6 – Termos com maior frequência extraídos das Questões de Competência

Origem	Termo	Tipo	Frequência
Questão	candidato	Conceito	16
Questão	eleição	Conceito	11
Questão	ser	Relacionamento	6

Documento de Especificação de Requisitos da Ontologia	
1 Propósito	Prover um modelo semântico que possa ser usado como base para a interligação dos diferentes conjuntos de dados abertos relacionadas às eleições disponibilizados pelo Tribunal Superior Eleitoral e geração de conhecimento mais rico. O objetivo é de que a ontologia criada seja utilizada como base para o processo de publicação dos dados como Dados Aberto Ligados, servindo principalmente para a criação de triplas no formato RDF.
2 Escopo	A ontologia das eleições brasileiras deve prover um modelo que reflita a realidade dos dados disponíveis de maneira aberta sobre as eleições no Brasil, com foco nas próprias eleições (o processo) e nos candidatos a cargos políticos públicos.
3 Language de Implementação	RDF/XML
4 Usuários	Usuário 1 - Pesquisadores e pessoas com interesse no domínio eleitoral brasileiro e desejam fazer uso de recursos ontológicos Usuário 2 - Governo brasileiro
5 Usos pretendidos	Uso pretendido 1 - Servir como base semântica para o processo de geração de triplas RDF para o domínio das eleições brasileiras
6 Requisitos	
a) Não-funcionais	RNF1 - A ontologia só se refere a eleições ocorridas após a redemocratização (que ocorreu no Brasil em 1988)
b) Funcionais	
Eleição (7 QCs)	Candidato (15 QCs)
CQ1 - Em que ano a eleição foi realizada? CQ2 - Quais cargos políticos foram disputados na eleição? CQ3 - Quem foram os candidatos que disputaram cargos políticos na eleição? CQ4 - Quais as zonas eleitorais existentes para a eleição? CQ5 - Quais as seções eleitorais existentes para a eleição? CQ6 - Qual a porcentagem de participação de eleitores na eleição? CQ7 - Qual a porcentagem de abstenção de eleitores na eleição?	CQ8 - Qual o nome do candidato? CQ9 - Qual o partido do candidato? CQ10 - Qual a coligação do candidato? CQ11 - Qual o número do candidato? CQ12 - Qual cargo político o candidato disputou? CQ13 - Por qual cidade e/ou estado o candidato disputou a eleição? CQ14 - Quantos votos o candidato recebeu? CQ15 - O candidato foi eleito? CQ16 - O candidato foi caçado? CQ17 - Quais os bens declarados pelo candidato? CQ18 - Qual o valor de cada bem declarado pelo candidato? CQ19 - Quais as fontes de receitas declaradas pelo candidato para sua campanha? CQ20 - Quais as fontes de despesas declaradas pelo candidato para sua campanha? CQ21 - Qual o valor recebido de cada fonte de receita do candidato em sua campanha? CQ22 - Qual o valor gasto em cada fonte de despesa pelo candidato em sua campanha?
Partido (3 QCs)	
CQ23 - Quais partidos disputaram a eleição? CQ24 - Quais coligações foram formadas para disputar a eleição? CQ25 - Quais partidos integram cada coligação formada para a	

Figura 6 – Extrato do Documento de Especificações de Requisitos da Ontologia das Eleições Brasileiras

Reutilização de Recursos Ontológicos e Não-ontológicos

Para a ontologia das eleições brasileiras foram encontrados 9 recursos passíveis de reutilização, sendo 5 Recursos Ontológicos e 4 Recursos Não-Ontológicos. Os ROs são provenientes de dois vocabulários diferentes: Friend of a Friend (foaf)⁶ e DBPedia⁷, o principal repositório de recursos em Dados Abertos Ligados do mundo. Todos os ROs mapeados para reutilização estão listados na Tabela 7. Tal tabela apresenta qual foi o recurso utilizado e a qual vocabulário ele pertence. Os recursos são acompanhados de seus prefixos, utilizados para referenciá-los na ontologia.

Os RNOs identificados foram o Tesouro da Justiça Eleitoral (Tribunal Superior Eleitoral, 2010); o Tesouro do Senado Federal (Senado Federal, 2019); o Vocabulário Controlado do Governo Eletrônico (ePing, 2011) e o Glossário Eleitoral (Tribunal Superior

⁶<http://xmlns.com/foaf/spec/>

⁷dbpedia.org/

Tabela 7 – Vocabulários e recursos reutilizados ou ligados à ontologia das Eleições Brasileiras

Vocabulário	Recurso
foaf	foaf:Person
foaf	foaf:name
foaf	foaf:Group
DBPedia	dbo:PoliticalParty
DBPedia	dbo:Election

Eleitoral, 2019). Foi feita a escolha de não realizar um processo de reengenharia para serem transformados em ontologias e assim poderem ser agregados à ontologia pois esse processo seria custoso em comparação a desenvolver a ontologia com base nos requisitos levantados. Mas apesar não-utilização como recursos da ontologia, eles foram referenciados na ontologia em forma de anotações utilizando a propriedade `rdfs:seeAlso`. Com isso, foi procurado mostrar para os utilizadores da ontologia que esses recursos existem e são fontes confiáveis de conhecimento semântico sobre o domínio.

Os RNOs podem ser utilizados futuramente como complemento da ontologia das eleições. O propósito seria cobrir lacunas e assim criar uma ontologia mais completa do domínio, ou até mesmo criar uma rede de ontologias (*ontology network*) que representem, cada uma, diferentes aspectos do domínio das eleições brasileiras.

Conforme dito no Capítulo 4, não houve busca por indivíduos em conjuntos de dados publicados como Dados Abertos Ligados na web para esse experimento, visto que não foram extraídos os termos e frequências das respostas das Questões de Competência. Durante a implementação foram buscados alguns dos indivíduos da ontologia na DBPedia⁸ para exemplificar como seria a representação deles como Dados Abertos Ligados. Para esse experimento acreditamos que de fato a melhor fonte para ligação é a DBPedia, por não terem sido identificados recursos ontológicos específicos sobre o domínio eleitoral brasileiro em outros vocabulários.

⁸dbpedia.org/

Conceitualização da Ontologia

O pré-glossário produzido foi utilizado como a lista de termos importantes, conforme sugerida pela Ontology Development 101. Foi conduzida uma reavaliação dos termos listados no pré-glossário para identificar se estavam classificados corretamente. Nessa reavaliação alguns termos tiveram sua classificação alterada de maneira a refletirem melhor o conhecimento do domínio. A Figura 7 apresenta as classes e relacionamentos propostos para a ontologia das eleições brasileiras, que foram derivadas a partir das QCs e refinadas na etapa de levantamento de requisitos. Ao todo foram projetadas 12 classes e 12 relacionamentos distintos entre elas. Existem relacionamentos que se repetem entre classes diferentes, mas que têm o mesmo significado, e por isso não são separados. Um exemplo é o relacionamento `disputaEleicao`, que é compartilhado entre as classes `Candidato` e `Partido` para ligação com a classe `Eleicao`. Por uma decisão de projeto não foram incluídas as propriedades exclusivas de cada classe (suas propriedades de dados).

Como conceito/classe principal da ontologia há a `Eleicao`, que representa um processo eleitoral realizado no Brasil. Uma `Eleicao` e um `Municipio` tem `ZonaEleitoral`. Uma `ZonaEleitoral` tem uma ou mais `SecaoEleitoral`. Uma `Eleicao` é disputada por vários candidatos (`Candidato`) que disputam `CargoPolitico` por uma `UnidadeEleitoral`. Os candidatos precisam ser associados a um `Partido`, o qual geralmente (mas nem sempre) forma `Coligacao` para disputar a `Eleicao`. Para fazer parte da disputa, um candidato precisa declarar os bens (`Bem`) que possui. Além disso, ele também precisa declarar as fontes de receita (`FonteDeReceitaCandidato`) e as fontes de despesa (`FontesDeDespesaCandidato`) da sua campanha, ou seja, qual a origem e destino do dinheiro gasto.

O exemplo a seguir, apresenta algumas dessas classes e relacionamentos instanciados de acordo com dados das Eleições de 2018 do Brasil. A Eleição Geral Federal de 2018 teve vários cargos em disputa, entre eles o de presidente do país. O segundo turno foi disputado por dois candidatos: Fernando Haddad e Jair Bolsonaro. Seja o candidato vencedor como exemplo: Jair Bolsonaro fazia parte do partido PSL, que fez uma coligação com o partido PRTB para as eleições presidenciais. Para concorrer, o candidato precisou declarar seus bens, como um carro Toyota Corolla XEI, 2014/2015,

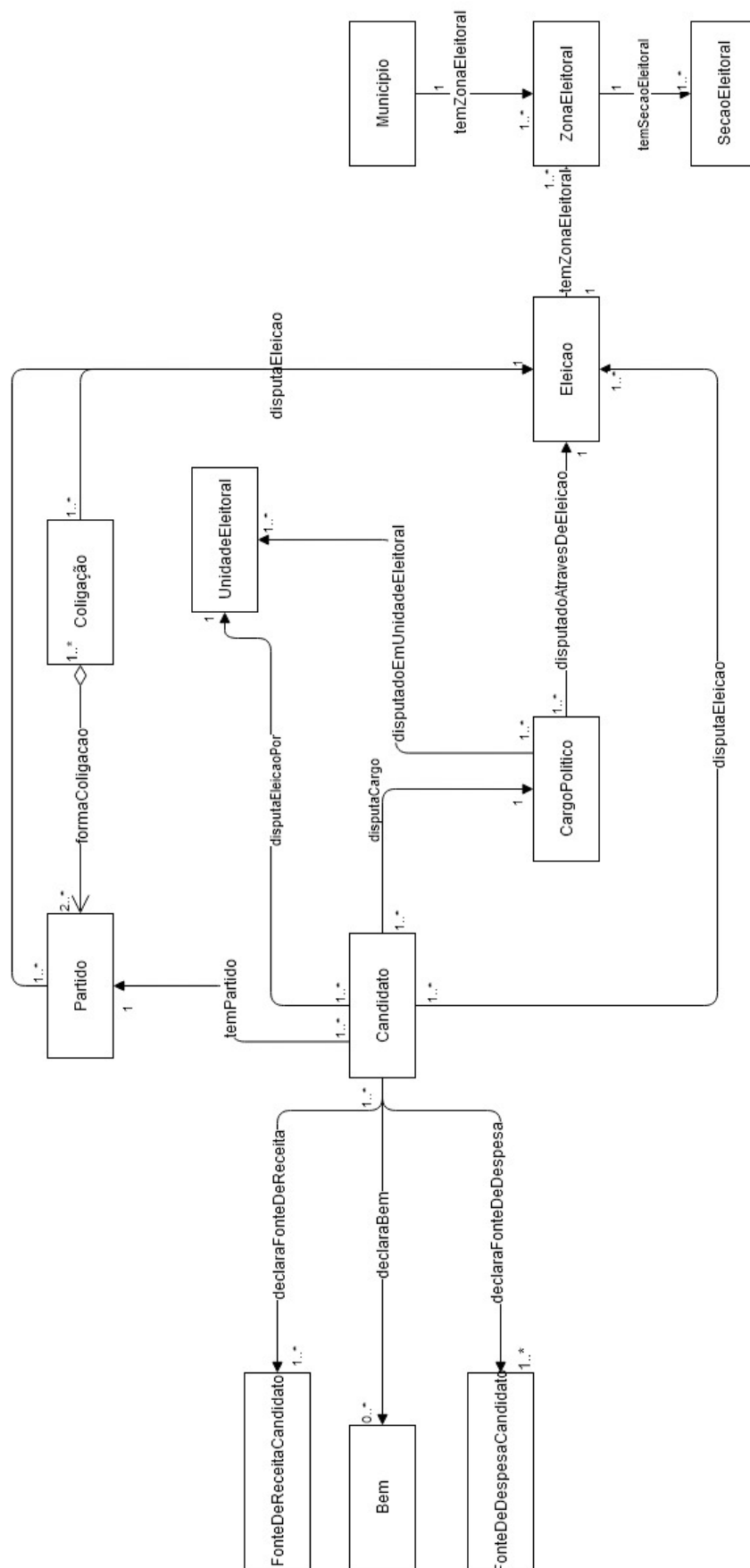


Figura 7 – Classes e relacionamentos da ontologia das Eleições Brasileiras derivadas das Questões de Competência

placa PAB3671 no valor de R\$ 50.000 (Tribunal Superior Eleitoral, 2018). Além disso, o candidato também teve de declarar receitas e despesas de sua campanha, como R\$ 4,377 milhões recebidos via financiamento coletivo (receita) e R\$ 660 mil gastos com a produção de programas de rádio, televisão ou vídeo (despesa) (Pontes, 2018). A Tabela 8 apresenta a relação de classes e instâncias descritas nesse exemplo.

Tabela 8 – Relação de classes e instâncias da ontologia das Eleições Brasileiras de cenário hipotético

Classe	Instância
Eleição	Eleição Geral Federal de 2018
Cargo político	Presidente da república
Candidato	Jair Bolsonaro, Fernando Haddad
Partido	PSL, PRTB
Bem	Toyota Corolla XEI, 2014/2015, placa PAB3671
Fonte de receita	Financiamento coletivo
Fonte de despesa	Produção de programa de rádio, televisão e vídeo

Ao mapear a hierarquia das classes foi percebido que o modelo apresentava poucas subclasses. Ao analisar pela ótica da metodologia Ontology Development 101, isso deve-se ao fato de a modelagem proposta para a ontologia ter um escopo no qual não existem muitas especializações, tratando o domínio de maneira mais generalizada. Isso faz com que elementos que poderiam ser classificados como classes em outras ontologias sejam identificados como instâncias na presente ontologia, pois são os conceitos mais específicos retratados (Noy and McGuinness, 2001). Um exemplo dessa questão é a representação de partidos políticos. Foi mapeada uma classe *Partido* para identificar esse conhecimento. Em uma ontologia mais robusta, poderia haver subclasses de *Partido* referentes a orientação ideológica deles (extrema-esquerda, esquerda, centro-esquerda, centro, centro-direita, direita, extrema-direita) e cada partido em si seria mapeado também como subclasses dessas, formando níveis de especialização. Como não havia o interesse em tratar as linhas ideológicas dos partidos, cada um deles foi mapeado como uma instância da classe *Partido*.

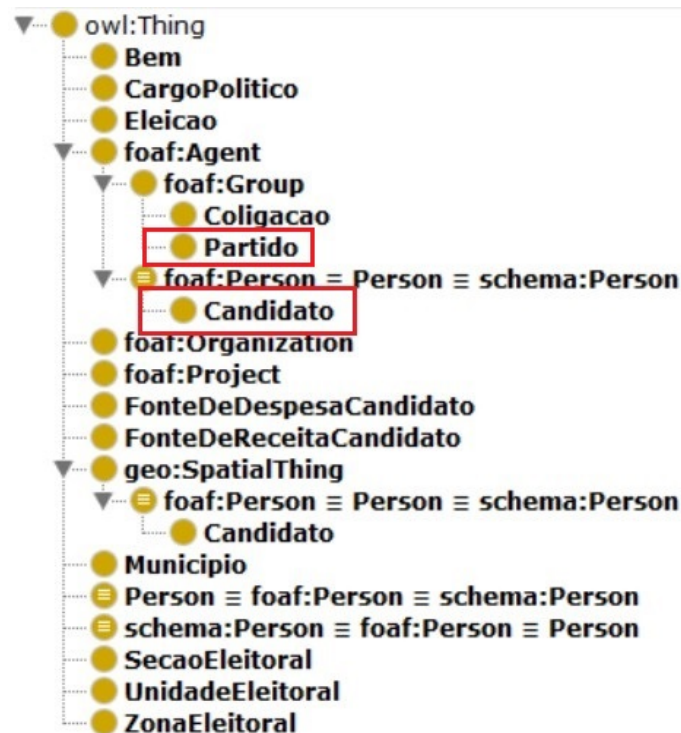
Implementação da Ontologia

Como a *Ontology Development 101* é focada na implementação da ontologia, ao adotá-la, o modelo conceitual concebido para a ontologia pôde ser diretamente transposto para a linguagem de implementação. Para isso foi utilizada a ferramenta Protégé, através de sua interface visual. As classes do modelo conceitual foram criadas como classes na ferramenta. Os relacionamentos entre as classes foram criados como propriedades de objeto. As propriedades de cada classe, que não foram representadas no modelo conceitual, foram criadas como propriedades de dados.

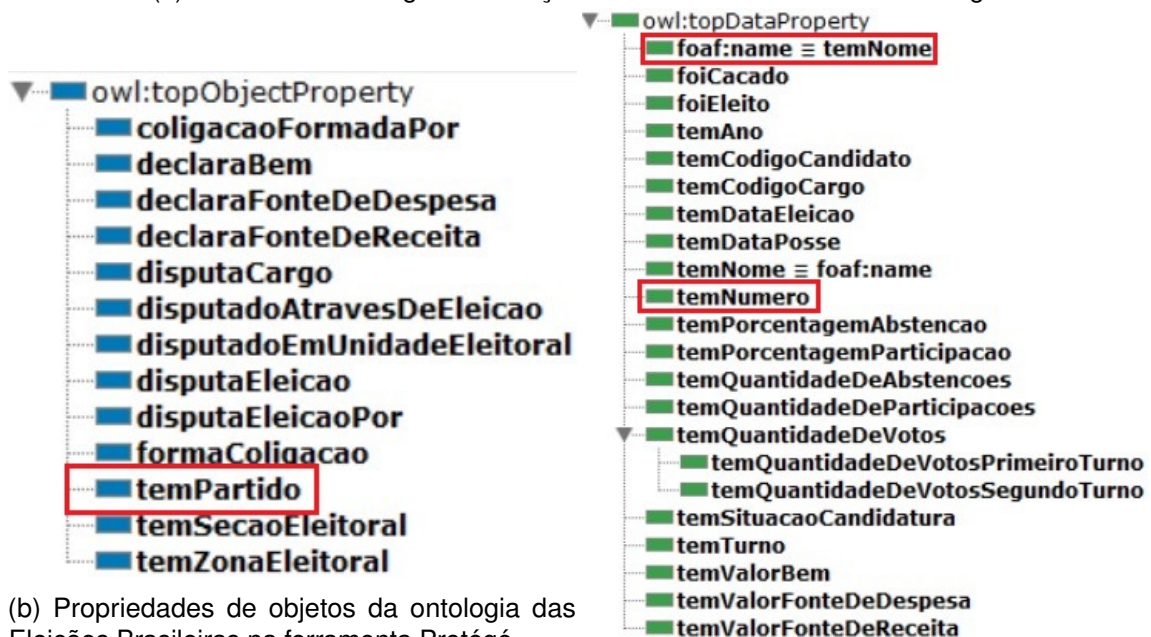
A Figura 8 mostra as classes, propriedades de objetos e de dados da Ontologia das Eleições Brasileiras criadas no Protégé. Na Figura 8a duas classes são destacadas: *Candidato* e *Partido*. Essas classes possuem o relacionamento *temPartido* entre elas. Essa propriedade está destacada na Figura 8b. Na Figura 8c são destacadas propriedades do *Candidato*. A propriedade *temNome* possui o nome do candidato como valor. A propriedade *temNumero* possui o número do candidato na eleição como valor.

Como descrito na Seção 5.1.2, foram encontrados cinco recursos ontológicos para serem reutilizados. Estes ROs foram utilizados de maneiras diferentes na ontologia. Parte deles foi importada para a ontologia e outra parte referenciada através de suas URIs, conforme descrito a seguir. Para os recursos pertencentes a ontologia FOAF, foi realizada a importação para a ontologia das eleições de modo a serem utilizadas as classes e propriedades relevantes na construção da ontologia. Para os recursos pertencentes à DBPedia, a referência foi feita através da utilização do predicado *owl:equivalentClass*, para confirmar que as classes *Eleicao* e *Partido* são equivalentes às classes *Election*, representada pela URI <http://dbpedia.org/ontology/Election> e *Party*, representada pela URI <http://dbpedia.org/ontology/PoliticalParty>.

Após a criação da estrutura da ontologia, a etapa seguinte consiste na carga dos dados que serão publicados como Dados Abertos Ligados. Como dito ao explicar a estratégia de integração, neste trabalho foi decidido manter os dados em arquivos separados, mantendo seu formato original, e utilizar o *plugin* Cellfie para realizar a carga, transformando registros dos arquivos CSV em indivíduos da ontologia. Para utilização do *plugin* foi necessário transformar os arquivos CSV em uma planilha do Microsoft Excel (formato .xls ou .xlsx), que é o formato de entrada aceito pelo Cellfie. Para isso basta



(a) Classes da ontologia das Eleições Brasileiras na ferramenta Protégé



(b) Propriedades de objetos da ontologia das Eleições Brasileiras na ferramenta Protégé

(c) Propriedades de dados da ontologia das Eleições Brasileiras na ferramenta Protégé

Figura 8 – Classes e propriedades de objetos e de dados da ontologia das Eleições Brasileiras na ferramenta Protégé

utilizar qualquer programa de edição de planilhas e salvar o arquivo no formato adequado.

Para definir corretamente as classes e propriedades de cada indivíduo é necessária a criação de regras de transformação (*transformation rules*) utilizando a sintaxe Manchester (Horridge and Patel-Schneider, 2012). No Código 1, é apresentado um exemplo de uma regra utilizada. A regra se aplica a todas as linhas na coluna AC, que é a coluna que traz os dados de partidos na planilha, definindo suas classes (*Types*) como Partido. Os indivíduos da coluna AC possuem um relacionamento disputaEleicao com indivíduos da coluna H e um relacionamento formaColigacao com indivíduos da coluna AF.

```
Individual: @AC*
Facts: disputaEleicao @H*(mm:toUpperCase),
formaColigacao @AF*(mm:toUpperCase)
Types: Partido
```

Código 1 – Regra de transformação do Cellfie para criação de instâncias do tipo Partido

Para este trabalho, dados parciais das eleições federais (presidente) e estaduais de 2018 foram carregados. A nível estadual, apenas a eleição (instância) e os cargos disputados foram carregados. No nível federal, com exceção dos dados de receitas e despesas dos candidatos e as zonas eleitorais, os demais dados foram carregados. Essa escolha se deu por dois motivos: problemas de desempenho do *plugin* Cellfie para lidar com arquivos muito grandes; e a não necessidade de ter os dados completos para validação da ontologia. Com isso, a ontologia com os dados gerou um total de 6.821 triplas. A Figura 9 traz um exemplo de um indivíduo instanciado na ontologia, com suas propriedades de objeto e dados. O candidato apresentado é Guilherme Boulos (*temNome* "GUILHERME BOULOS"). Como pode ser observado, ele é do partido PSOL (*temPartido* PSOL), não foi eleito (*foiEleito* "NÃO ELEITO") e teve 617.122 votos (*temQuatidadeDeVotosPrimeiroTurno* 617122).

Alguns dados tiveram que ser preenchidos manualmente na ontologia pela necessidade de agregação deles para que fosse possível responder a todas as questões de competência. Os dados preenchidos manualmente foram a quantidade de votos de cada candidato, a participação na eleição e a abstenção na eleição. Vale ressaltar que os dados são obteníveis através dos arquivos, mas seria necessária sua agregação e não é

Property assertions: 280000601016					
	declaraFonteDeReceita 280000601016_12085394				
	declaraFonteDeReceita 280000601016_12085370				
	declaraFonteDeReceita 280000601016_12085386				
	declaraFonteDeReceita 280000601016_12085398				
	declaraFonteDeReceita 280000601016_12085362				
	declaraFonteDeReceita 280000601016_12085374				
	declaraFonteDeReceita 280000601016_12085390				
	temPartido PSOL				
	declaraBem 280000601016_21				
	declaraFonteDeReceita 280000601016_12085357				
	declaraFonteDeReceita 280000601016_12085369				
	declaraFonteDeReceita 280000601016_12085377				
	declaraFonteDeReceita 280000601016_12085389				
	declaraFonteDeReceita 280000601016_12085393				
	declaraFonteDeReceita 280000601016_12085361				
	declaraFonteDeReceita 280000601016_12085381				
	declaraFonteDeReceita 280000601016_12085397				
	declaraFonteDeReceita 280000601016_12085365				
	declaraFonteDeReceita 280000601016_12085373				
	declaraFonteDeReceita 280000601016_12085385				
Data property assertions					
	temQuantidadeDeVotosPrimeiroTurno 617122				
	temSituacaoCandidatura "APTO"^^xsd:string				
	foiEleito "NÃO ELEITO"^^xsd:string				
	foiCacado "NÃO"				
	temCodigoCandidato "280000601016"^^xsd:string				
	temNome "GUILHERME BOULOS"^^xsd:string				
	temNumero "50"^^xsd:string				

Figura 9 – Exemplo de propriedades de objeto e de dados de um candidato. Nesse exemplo o candidato apresentado é Guilherme Boulos

possível fazê-la através das regras de transformação do *plugin* Cellfie.

Tendo os indivíduos criados na ontologia, deve ser feita a ligação deles com recursos externos presentes em outros conjuntos de dados para que se tenha de fatos Dados Ligados. Pelo volume grande de indivíduos da ontologia, optou-se por não fazer a busca e ligação de todos os indivíduos carregados, que são das mais variadas classes (Candidato, Partido, Bem etc). Contudo, foi feita a ligação para alguns indivíduos como forma de exemplificar o processo. A Figura 10 mostra a declaração de que o candidato Guilherme Boulos, representado na ontologia pela URI http://localhost:8080/dissertacao_mestrado/ontologia_eleicoes_brasileiras_dissertacao.rdf#280000601016 é o mesmo recurso (owl:sameAs) que o recurso representado pela URI http://pt.dbpedia.org/resource/Guilherme_Boulos, que é a página dele na DBPedia. Foram feitas ligações para alguns indivíduos de outras classes também, que podem ser vistos no arquivo disponibilizado em <https://git.io/JU2U6>.

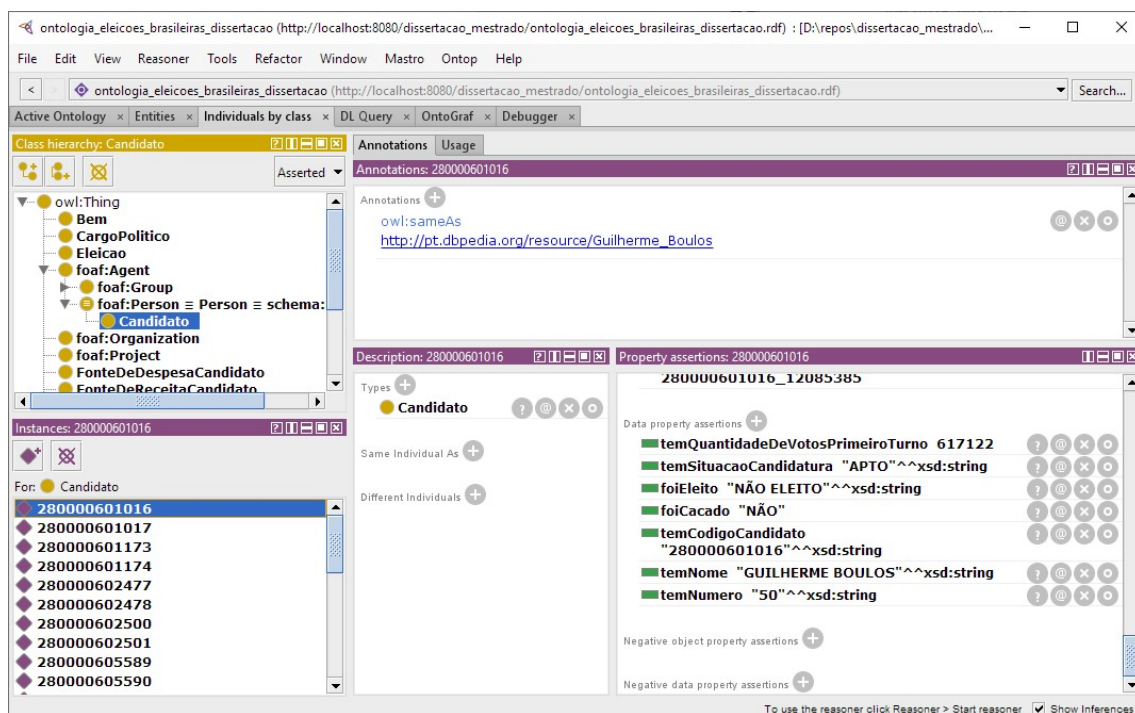


Figura 10 – Apresentação de como é feita a ligação com recursos externos à ontologia a nível de indivíduo. A imagem mostra que o candidato Guilherme Boulos, que é um indivíduo da ontologia, é o mesmo representado no recurso http://pt.dbpedia.org/resource/Guilherme_Boulos

5.1.3 Publicação como Dados Abertos Ligados

A implementação da ontologia na ferramenta Protégé facilita a tarefa da publicação dos dados como Dados Abertos Ligados. Isto se deve à facilidade da ferramenta para gerar, a partir do arquivo OWL da ontologia, outro arquivo no formato RDF/XML. O arquivo RDF gerado nesta tarefa já se encontrará de acordo com os padrões necessários para sua publicação como Dados Ligados, visto que a ligação com recursos externos é feita durante o desenvolvimento da ontologia.

Devido ao arquivo RDF ser estático, ou seja, que não tem mudanças dinâmicas em suas triplas, optou-se por seguir o caminho mais simples de publicação, que é disponibilizar o arquivo em um servidor web. Para a finalidade dessa pesquisa não foi realizada a publicação em um servidor exposto para a internet, tarefa essa elencada para evolução do trabalho. Um servidor local executando a aplicação Apache Tomcat⁹ foi utilizado para validação do processo. A Figura 11 mostra o resultado de um acesso direto a URI que representa o recurso `Eleicao`. Utilizando uma *Hash URI* (Sauermann and Cyganiak, 2008) para o recurso torna-se possível acessá-lo diretamente. Para conseguir visualizar em um navegador da web, foi utilizada uma extensão do navegador Mozilla Firefox chamada RDF Browser¹⁰, pois navegadores da web tradicional não conseguem tratar arquivos RDF naturalmente.

A abordagem utilizada não permite a negociação de conteúdo ou redirecionamentos 303. Isso significa que não é possível receber um conteúdo com base no tipo da requisição (se está pedindo por uma página HTML ou pelo arquivo RDF) e nem que esse redirecionamento seja feito automaticamente (Magalhães et al., 2011). Apesar disso optou-se por seguir com essa abordagem por ser mais simples e atender aos critérios necessários para considerar um conjunto de dados publicados como Dados Ligados.

⁹<http://tomcat.apache.org/>

¹⁰<https://github.com/kianschmalenbach/rdf-browser>



Figura 11 – Visualização da classe Eleicao em um navegador, acessada através de sua URI

5.1.4 Publicação de *endpoint* SPARQL

Mesmo não sendo algo mandatório para considerar como Dados Abertos Ligados o conjunto de dados publicado, é boa prática disponibilizar uma maneira de consultá-lo. Essa disponibilização é feita por meio de consultas SPARQL.

Para permitir a consulta aos dados foi utilizado o servidor de SPARQL Apache Jena Fuseki, também hospedado em um servidor web local (assim como os dados). O Fuseki permite tanto consultas através de uma interface visual acessível pelo navegador quanto através de *endpoints*, nos quais podem ser feitas requisições HTTP das consultas. O Código 2 traz um exemplo de consulta escrita em SPARQL para responder a QC2 - “Quais cargos políticos foram disputados na eleição?”.

A Figura 12 apresenta o resultado da consulta do Código 2 na interface do Fuseki. Como a consulta se refere a dados específicos da Eleição Geral Federal de 2018, o resultado traz o único cargo que foi disputado nessa eleição, que é o de presidente.

```
PREFIX eleicoes: <http://localhost:8080/dissertacao_mestrado/
    ontologia_eleicoes_brasileiras_dissertacao.rdf#>
SELECT ?eleicao ?cargo
WHERE {
    ?cargo eleicoes:disputadoAtravesDeEleicao ?eleicao.
    FILTER (?eleicao=eleicoes:295_ELEIÇÃOGERALFEDERAL2018)
}
```

Código 2 – Consulta SPARQL para responder à QC2 - “Quais cargos políticos foram disputados na eleição?”

O Fuseki também disponibiliza *endpoints* que podem ser adotados para consulta dos dados utilizando requisições HTTP. No Código 3 é mostrado um exemplo de requisição para a mesma consulta SPARQL apresentada no Código 2 e seu retorno em formato JSON.

QUERY RESULTS

Table Raw Response

Showing 1 to 1 of 1 entries

Search

	eleicao	cargo
1	eleicoes:295_ELEIÇÃOGERALFEDERAL2018	eleicoes:PRESIDENTE

Showing 1 to 1 of 1 entries

Figura 12 – Resultado da consulta SPARQL mostrando os cargos disputados para a eleição 295_ELEIÇÃOGERALFEDERAL2018

```
http://localhost:8080/fuseki/eleicoesbrasil/?query=PREFIX%20
  eleicoes%3A%20%3Chttp%3A%2F%2Flocalhost%3A8080%2
  Fdissertacao_mestrado%2
  Fontologia_eleicoes_brasileiras_dissertacao.rdf%23%3E%0A%0
  ASELECT%20%3Feleicao%20%3Fcargo%0AWHERE%20%7B%0A%20%20%3
  Fcargo%20eleicoes%3AdisputadoAtravesDeEleicao%20%3Feleicao
  .%0A%20%20FILTER%20%28%3Feleicao%3Deleicoes%3A295_ELEI%C3
  %87%C3%83OGERALFEDERAL2018%29%0A%7D

{
  "head": {
    "vars": [
      "eleicao",
      "cargo"
    ]
  },
  "results": {
    "bindings": [
      {
        "eleicao": {
          "type": "uri",
          "value": "http://localhost:8080/
            dissertacao_mestrado/
```

```

        ontologia_eleicoes_brasileiras_dissertacao.rdf
        #295_ELEIÇÃOGERALFEDERAL2018"
    },
    "cargo": {
        "type": "uri",
        "value": "http://localhost:8080/
        dissertacao_mestrado/
        ontologia_eleicoes_brasileiras_dissertacao.rdf#
        PRESIDENTE"
    }
}
]
}
}

```

Código 3 – Exemplo de consulta feita através de requisição HTTP para um endpoint SPARQL e a resposta retornada

Como pode ser visto no Código 3, a requisição HTTP nada mais é do que a consulta escrita em SPARQL utilizando uma codificação UTF-8. As consultas realizadas através de requisições são o formato mais adequado para popular aplicações que irão fazer utilização dos dados para construir visualizações dos dados publicados em RDF.

A ontologia, arquivo RDF, regras de transformação e documentos relevantes relacionados a implementação apresentada nessa seção podem ser encontrados em https://github.com/aiquis/dissertacao_mestrado/tree/master/eleicoes_brasileiras.

5.2- Dados de Distribuição das Unidades Básicas de Saúde em Funcionamento

5.2.1 Identificação, Análise e Integração dos Dados

Os dados sobre Unidades Básicas de Saúde (UBS) em funcionamento no Brasil são publicados pelo Ministério da Saúde no Portal Brasileiro de Dados Abertos. Os dados

estão disponíveis em formato CSV, JSON, ZIP, KML e HTML. Para a aplicação realizada neste trabalho foi selecionado o arquivo em formato CSV, composto por 40.612 registros.

Os dados sobre as UBS em funcionamento possuem um menor nível de complexidade de entendimento em relação à aplicação anterior, sobre os dados eleitorais. Isto se deve, pois os dados não se apresentam espalhados por múltiplos arquivos, logo não é necessário buscar como os dados espalhados por diferentes arquivos se conectam entre si. Diferente da aplicação dos dados eleitorais na qual foi representado um domínio maior (de uma eleição brasileira), para esse conjunto de dados, o delineamento foi apenas publicar os dados presentes no arquivo como Dados Abertos Ligados, sem o propósito de representar um domínio maior da área da saúde. Essa escolha foi feita para mostrar a possibilidade de aplicação do processo aqui proposto não só na representação de um domínio completo, mas também quando se deseja apenas publicar um único conjunto de dados como Dados Abertos Ligados.

O arquivo de dados das UBS contém um identificador de cada registro, o identificador da UBS no Cadastro Nacional de Estabelecimentos de Saúde (CNES), localização (por meio de latitude e longitude), endereço, telefone e colunas de ano e mês. Para essas duas últimas colunas não foi possível identificar seu significado por falta de uma documentação do conjunto de dados. Após análise, verificou-se não ser necessária uma modelagem completa dos dados, separando-os em diferentes classes pois todas as propriedades são referentes a uma classe única, a saber, a Unidade Básica de Saúde. Como decisão para o desenvolvimento da ontologia, não foram utilizados os dados de mês devido a falta de um dicionário de dados e por consequência, o não entendimento do domínio deles. Outra decisão se refere a manter os dados em formato CSV para popular a ontologia, com o objetivo de conferir mais agilidade ao processo e mantê-lo simples.

5.2.2 Desenvolvimento da Ontologia

Nessa etapa serão detalhadas as etapas seguidas para construção da Ontologia das Unidades Básicas de Saúde em funcionamento no Brasil.

Levantamento de Requisitos

Cenário Motivador, Escopo e Linguagem A motivação para a criação dessa ontologia é gerar uma representação semântica do conjunto de dados referente às Unidades Básicas de Saúde em funcionamento no Brasil, com o objetivo de publicar estes dados como Dados Abertos Ligados. A ontologia deve ser capaz de representar o conhecimento presente nos dados sobre Unidades Básicas de Saúde disponibilizados de forma aberta pelo Ministério da Saúde. Assim como a ontologia das Eleições Brasileiras, a implementação também será feita utilizando a interface visual da ferramenta Protégé, que utiliza o formato OWL como padrão. No entanto, os dados serão exportados para um arquivo no formato RDF/XML, o qual será utilizado para a publicação como Dados Abertos Ligados.

Usuários e Uso Pretendido A ontologia construída pode ser utilizada por pesquisadores, profissionais da área de saúde e pessoas que desejam utilizar dados das Unidades Básicas de Saúde em funcionamento em um formato semântico por meio de uma ontologia que represente esses dados disponíveis. Seu uso pretendido é servir como base semântica para a publicação dos dados das Unidades Básicas de Saúde em funcionamento como Dados Abertos Ligados. Assim como a ontologia das Eleições Brasileiras, governo e população em geral também podem se beneficiar. O primeiro por meio da disponibilização dos dados semânticos junto a outros formatos. A população pode se beneficiar por meio da utilização de aplicações que sejam baseadas na ontologia produzida.

Identificação de Requisitos Para a ontologia das UBS em funcionamento no Brasil não foram identificados requisitos não-funcionais. Para definição dos requisitos funcionais mais uma vez foi empregado o uso de Questões de Competência utilizando uma abordagem *middle out* com base nos dados existentes para as UBS. As QCs foram classificadas em duas categorias: unidade e localização. Sete questões que a ontologia deve ser capaz de responder foram elaboradas. A Tabela 9 apresenta todas as questões de competência.

Validação de Requisitos Para validar os requisitos dessa ontologia os mesmos critérios utilizados para validação da ontologia das eleições foram empregados. São eles: correteza, consistência, não-ambiguidade e realismo. Por ser um conjunto de dados mais

Tabela 9 – Questões de Competência sobre as Unidades Básicas de Saúde em funcionamento no Brasil

ID	Categoria	Questão
QC1	Unidade	Qual é o identificador único da Unidade Básica de Saúde?
QC2	Unidade	Qual é o número da Unidade Básica de Saúde no CNES?
QC3	Localização	Qual é a latitude da Unidade Básica de Saúde está localizada?
QC4	Localização	Qual é a longitude da Unidade Básica de Saúde está localizada?
QC5	Unidade	Qual é o nome fantasia da Unidade Básica de Saúde?
QC6	Localização	Qual é o endereço da Unidade Básica de Saúde?
QC7	Unidade	Qual é o número de telefone da Unidade Básica de Saúde?

simples e com poucos requisitos não foram necessárias múltiplas iterações para se chegar aos requisitos finais. A primeira versão dos requisitos foi validada e não houve novas alterações durante o desenvolvimento da ontologia.

Priorização de Requisitos Dada a simplicidade do modelo e o baixo número de requisitos, que estão aderentes aos dados apresentados, não foi necessário um esforço de priorização. Portanto, a ontologia deve ser capaz de responder a todas às Questões de Competência propostas, atendendo a todos os requisitos.

Extração de Terminologia e Frequências Seguindo os moldes da ontologia anteriormente desenvolvida, para a ontologia das UBS foram considerados apenas os termos identificados nas perguntas das QCs devido às respostas serem únicas para cada uma das unidades. Nove termos foram identificados e ranqueados da maior para a menor frequência para formar o pré-glossário de termos da ontologia. A Tabela 10 exibe todos os termos mapeados nas perguntas. A Figura 13 traz um extrato do Documento de Especificações de Requisitos da Ontologia produzido para esse experimento e preenchido ao longo do levantamento de requisitos, que também contém os termos e sua frequências.

Tabela 10 – Termos com maior frequência extraídos das Questões de Competência para a ontologia das UBS em funcionamento no Brasil

Origem	Tipo	Termo	Frequência
Questão	Conceito	Unidade Básica de Saúde	7
Questão	Relacionamento	ser	7
Questão	Atributo	CNES	1
Questão	Atributo	latitude	1
Questão	Atributo	longitude	1
Questão	Atributo	endereço	1
Questão	Atributo	número de telefone	1
Questão	Atributo	nome fantasia	1
Questão	Atributo	identificador único	1

Documento de Especificação de Requisitos da Ontologia		
1	Propósito	
		Criar uma representação semântica do conjunto de dados referente às Unidades Básicas de Saúde em funcionamento no Brasil, com o objetivo de publicar esses dados como Dados Abertos Ligados.
2	Escopo	
		A ontologia deve ser capaz de representar os dados sobre Unidades Básicas de Saúde disponíveis de maneira aberta em arquivos CSV pelo Governo Brasileiro.
3	Linguagem de Implementação	
		RDF/XML
4	Usuários	
		Pesquisadores e pessoas que desejam utilizar dados das Unidades Básicas de Saúde em funcionamento em um formato semântico
5	Usos pretendidos	
		Servir como base semântica para a publicação dos dados das Unidades Básicas de Saúde em funcionamento como Dados Abertos Ligados
6	Requisitos	
	a) Não-funcionais	
	b) Funcionais	
		CQ1 - Qual é o identificador único da Unidade Básica de Saúde? CQ2 - Qual é o número da Unidade Básica de Saúde no CNES? CQ3 - Qual é a latitude aonde a Unidade Básica de Saúde está localizada? CQ4 - Qual é a longitude aonde a Unidade Básica de Saúde está localizada? CQ5 - Qual é o nome fantasia da Unidade Básica de Saúde? CQ6 - Qual é o endereço da Unidade Básica de Saúde? CQ7 - Qual é o número de telefone da Unidade Básica de Saúde?
7	Pré-glossário de termos	
	a) Termos das Questões de Competência	
		Unidade Básica de Saúde - 7 ser - 7 CNES - 1 latitude - 1 longitude - 1 endereço - 1 número de telefone - 1 nome fantasia - 1 identificador único - 1
	b) Termos das respostas	
	c) Objetos	

Figura 13 – Extrato do Documento de Especificação de Requisitos da Ontologia das UBS em Funcionamento no Brasil

Reutilização de Recursos Ontológicos e Não-ontológicos

Para a Ontologia das UBS em funcionamento foram identificados nove recursos ontológicos passíveis de reutilização. Os ROs são provenientes de quatro vocabulários di-

ferentes e estão listados na Tabela 11. Não foram identificados Recursos Não-Ontológicos para referência na ontologia.

Os vocabulários utilizados foram o vcard¹¹, que traz conceitos para descrição de pessoas e organizações; o International Contact Ontology¹², que é uma ontologia para representação de contatos internacionais; Geonames¹³, que é uma ontologia para representações geográficas; e o VCGE¹⁴, que é o Vocabulário Controlado do Governo Eletrônico Brasileiro.

Tabela 11 – Vocabulários e recursos reutilizados ou ligados à ontologia das UBS em Funcionamento no Brasil

Vocabulário	Recurso
vcard	vcard:hasLocality
vcard	vcard:hasRegion
vcard	vcard:postal-code
vcard	vcard:street-address
International Contact Ontology	:hasPhoneNumber
Geonames	:lat
Geonames	:long
Geonames	:postalCode
VCGE	vcge:unidade-basica-saude

Conceitualização da Ontologia

Pela natureza simples do conjunto de dados, no qual, ao analisar os dados foi identificada apenas uma classe relevante, não foi identificada a necessidade da criação de um modelo conceitual da ontologia. Contudo, ao analisar a classe referente a Unidades Básicas de Saúde percebeu-se que tendo apenas ela na ontologia não seria possível realizar a ligação a nível de indivíduos com conjuntos de dados externos para que seja possível a publicação como Dados Abertos Ligados. Ao observar essa limitação, foi decidido pela criação de duas novas classes, Cidade e Estado, para representarem indivíduos relacionados a localização das UBS que poderiam ser ligados a recursos externos. Adotar essa abordagem foi necessário pois não foram identificados vocabulários

¹¹<https://www.w3.org/TR/vcard-rdf/>

¹²<http://ontology.eil.utoronto.ca/icontact.html>

¹³<https://www.geonames.org/ontology>

¹⁴<https://www.gov.br/governodigital/pt-br/governanca-de-dados/vocabulario-controlado-do-governo-eletronico>

que possuíssem Unidades Básicas de Saúde como seus indivíduos.

Implementação da Ontologia

Sem um modelo conceitual desenvolvido, o direcionamento para a implementação da ontologia veio da análise dos dados e das Questões de Competência. Na análise dos dados, foi identificada que a única classe relevante contida neles era a Unidade Básica de Saúde, que foi criada na ontologia como `UnidadeBasicaDeSaude`. No entanto, conforme dito na Seção 5.2.2 (Conceitualização da Ontologia), para que fosse possível realizar uma ligação com outros conjuntos de dados da web a nível de indivíduo, decidiu-se pela criação de duas outras classes: `Cidade` e `Estado`.

Para representar os relacionamentos entre as classes da ontologia foram reutilizados duas propriedades de objeto do vocabulário `vcard`, para indicar o estado onde a UBS está situado (`vcard:hasRegion`) e a cidade onde ela está localizada (`vcard:hasLocality`).

Neste experimento foi feita uma reutilização direta na ontologia para a maior parte dos recursos reutilizados, usando os recursos mapeados como uma propriedade dentro da ontologia. Propriedades foram criadas apenas para os que não tiveram equivalentes para reutilização. O único recurso que não foi utilizado da maneira descrita foi o `vcge:unidade-basica-saude`, para o qual foi utilizada a propriedade `owl:equivalentClass` para representar que esse recurso é o mesmo que a classe da ontologia `UnidadeBasicaDeSaude`. No Código 4 a classe `UnidadeBasicaDeSaude` é apresentada e um extrato das propriedades de dados dela. O arquivo completo pode ser visto em <https://git.io/JU2UH>.

```
<!-- http://localhost:8080/dissertacao_mestrado/
      ubsefuncionamento.rdf#UnidadeBasicaDeSaude -->

<owl:Class rdf:about="http://localhost:8080/
      dissertacao_mestrado/ubsefuncionamento.rdf#
      UnidadeBasicaDeSaude">
      <owl:equivalentClass rdf:resource="http://vocab.e.gov.br
```

```

        /2011/03/vcge#unidade-basica-saude"/>
</owl:Class>

<!-- http://localhost:8080/dissertacao_mestrado/
      ubsefuncionamento.rdf#nomeFantasia -->

<owl:DatatypeProperty rdf:about="http://localhost:8080/
      dissertacao_mestrado/ubsefuncionamento.rdf#nomeFantasia
      "/>

<!-- http://www.w3.org/2006/vcard/ns#street-address -->

<owl:DatatypeProperty rdf:about="http://www.w3.org/2006/vcard
      /ns#street-address"/>

<!-- https://www.geonames.org/ontology#lat -->

<owl:DatatypeProperty rdf:about="https://www.geonames.org/
      ontology#lat">
    <rdfs:range rdf:resource="http://www.w3.org/2001/
        XMLSchema#float"/>
</owl:DatatypeProperty>

<!-- https://www.geonames.org/ontology#long -->

<owl:DatatypeProperty rdf:about="https://www.geonames.org/
      ontology#long">
    <rdfs:range rdf:resource="http://www.w3.org/2001/
        XMLSchema#float"/>
</owl:DatatypeProperty>

```

Código 4 – Extrato da representação RDF das propriedades de uma UnidadeBasicaDeSaude

Apesar de o arquivo original de dados conter 40.612 registros, carregamos na

ontologia apenas 500 indivíduos devido aos problemas de desempenho com o *plugin* Cellfie. Novamente, esse número reduzido é suficiente para validar a ontologia e o processo, mas para ser uma ontologia completa que represente os dados seria necessário fazer a carga completa de todos os registros. Os dados carregados geraram um total de 5.901 triplas.

A Figura 14 apresenta um exemplo de uma UBS com seus dados preenchidos. O nome fantasia (*nomeFantasia*) da unidade é Centro de Saúde Aurelio Texeira Cotta. Seu identificador único (*identificadorUnico*) é o 2378 e seu número de identificação no CNES (*numeroCNES*) é o 2170051. O endereço da unidade (*street-address*) é Rua Maristela Braga, S/N, Centro, Bela Vista de Minas, MG. A cidade (*locality*) é Bela Vista de Minas. O CEP (*postalCode*) é o 35938000 e a unidade fica no estado (*region*) de Minas Gerais. Suas coordenadas geográficas (*lat* e *long*) são -19.8258467639902 e -43.0969322664076. O telefone (*hasPhoneNumber*) da unidade é o (31)38531271.

Property assertions: 2170051_CENTRODESAUDEAURELIOTEXEIRACOTTA	
Object property assertions +	
vcard:hasRegion MG	? @ X O
vcard:hasLocality BelaVistaDeMinas	? @ X O
Data property assertions +	
vcard:street-address "RUA MARISTELA BRAGA, S/N, CENTRO, Bela Vista de Minas, MG"^^xsd:string	? @ X O
nomeFantasia "CENTRO DE SAUDE AURELIO TEXEIRA COTTA"^^xsd:string	? @ X O
long -43.09693f	? @ X O
hasPhoneNumber "(31)38531271"^^xsd:string	? @ X O
identificadorUnico "2378"^^xsd:string	? @ X O
numeroCNES "2170051"^^xsd:string	? @ X O
postalCode "35938000"^^xsd:string	? @ X O
lat -19.825848f	? @ X O

Figura 14 – Exemplo de propriedades preenchidas de uma UBS na ontologia. Nesse exemplo são dados do Centro de Saúde Aurelio Texeira Cotta

O Código 5 exibe as regras de transformação que são utilizadas para carregar os dados na ontologia. Foram criadas 3 regras. A primeira consiste em criar um indivíduo da classe *UnidadeBasicaDeSaude* a partir da junção das colunas B, que contém o identificador único da unidade, e E, que contém o nome fantasia da unidade, da planilha. As outras colunas da planilha são mapeadas para cada propriedade. Por exemplo, a

coluna J contém os dados da propriedade `postalCode` (CEP). Além dessa foram criadas duas outras regras para criação de indivíduos da classe `Cidade` (indivíduos da coluna L) e da classe `Estado` (indivíduos da coluna K).

Após a criação dos indivíduos na ontologia deve ser feita a ligação com recursos externos para que seja possível a publicação como Dados Aberto Ligados. Assim como no experimento com dados das eleições, nesse experimento também foi feita a opção por não buscar recursos para a ligação de todos os indivíduos da ontologia. Foram feitas as ligações para alguns indivíduos das classes `Estado` e `Cidade` como forma de exemplificar a aplicação. O Código 6 traz os exemplos dos estado de SP e da cidade de Garanhuns, que estão conectadas à recursos do vocabulário Geonames que descrevem essas localidades. O exemplo traz menos cidades dentro do estado de SP do que existem na ontologia para facilitar a visualização.

```

Individual: @B*(mm:append("_", @E*))
Types: UnidadeBasicaDeSaude
Facts: lat @C* (xsd:float),
long @D* (xsd:float),
hasPhoneNumber @I*,
postalCode @J*,
vcard:hasLocality @L*,
vcard:hasRegion @K*,
vcard:street-address @F*(mm:append(", ", @G*(mm:append(", ",
    @H*(mm:append(", ", @L*(mm:append(", ", @K*)))))))),
numeroCNES @B*,
identificadorUnico @A*,
nomeFantasia @E*

Individual: @L*
Types: Cidade
Facts: vcard:hasRegion @K*

Individual: @K*
Types: Estado

```



```
Facts: vcard:hasLocality @L*
```

Código 5 – Regras de transformação utilizadas para carregar os dados das UBS na ontologia

```
<!-- http://localhost:8080/dissertacao_mestrado/
      ubsefuncionamento.rdf#SP -->
<owl:NamedIndividual rdf:about="http://localhost:8080/
dissertacao_mestrado/ubsefuncionamento.rdf#SP">
  <rdf:type rdf:resource="http://localhost:8080/
dissertacao_mestrado/ubsefuncionamento.rdf#Estado"/>
  <ns:hasLocality rdf:resource="http://localhost:8080/
dissertacao_mestrado/ubsefuncionamento.rdf#Adamantina
"/>
  <ns:hasLocality rdf:resource="http://localhost:8080/
dissertacao_mestrado/ubsefuncionamento.rdf#Osasco"/>
  <ns:hasLocality rdf:resource="http://localhost:8080/
dissertacao_mestrado/ubsefuncionamento.rdf#OuroVerde
"/>
  <owl:sameAs rdf:resource="https://www.geonames.org
/3448433"/>
</owl:NamedIndividual>

<!-- http://localhost:8080/dissertacao_mestrado/
      ubsefuncionamento.rdf#Garanhuns -->
<owl:NamedIndividual rdf:about="http://localhost:8080/
dissertacao_mestrado/ubsefuncionamento.rdf#Garanhuns">
  <rdf:type rdf:resource="http://localhost:8080/
dissertacao_mestrado/ubsefuncionamento.rdf#Cidade"/>
  <ns:hasRegion rdf:resource="http://localhost:8080/
dissertacao_mestrado/ubsefuncionamento.rdf#PE"/>
  <owl:sameAs rdf:resource="https://sws.geonames.org
/3399058"/>
</owl:NamedIndividual>
```

Código 6 – Representação RDF dos indivíduos referentes ao estado de São Paulo e à cidade de Garanhuns na ontologia. Esses recursos estão ligados a recursos do vocabulário Geonames através do predicado `owl:sameAs`

5.2.3 Publicação como Dados Abertos Ligados

Para publicação dos dados das Unidades Básicas de Saúde em funcionamento como Dados Abertos Ligados foi adotada a mesma abordagem feita para os dados eleitorais, ou seja, publicar um arquivo RDF/XML estático em um servidor web. A estrutura utilizada seguiu novamente com um servidor local executando o servidor Apache Tomcat. A Figura 15 exibe como exemplo o Centro de Saúde Aurelio Texeira Cotta sendo acessado por meio de sua URI.

5.2.4 Publicação de *endpoint* SPARQL

Para a disponibilização de uma forma de consulta dos dados, plataforma do Apache Jena Fuseki foi utilizada e hospedada no servidor local. O Código 7 apresenta um exemplo de consulta SPARQL para responder à QC5 - “Qual é o nome fantasia da Unidade Básica de Saúde?”. A Figura 16 apresenta o resultado da consulta.

```
PREFIX ubs:<http://localhost:8080/dissertacao_mestrado/
       ubsementracamento.rdf#>

SELECT ?ubs ?nome
WHERE {
  ?ubs ubs:nomeFantasia ?nome.
  FILTER (?ubs=ubs:2170051_CENTRODESAUDEAURELIOTEXEIRACOTTA)
}
```

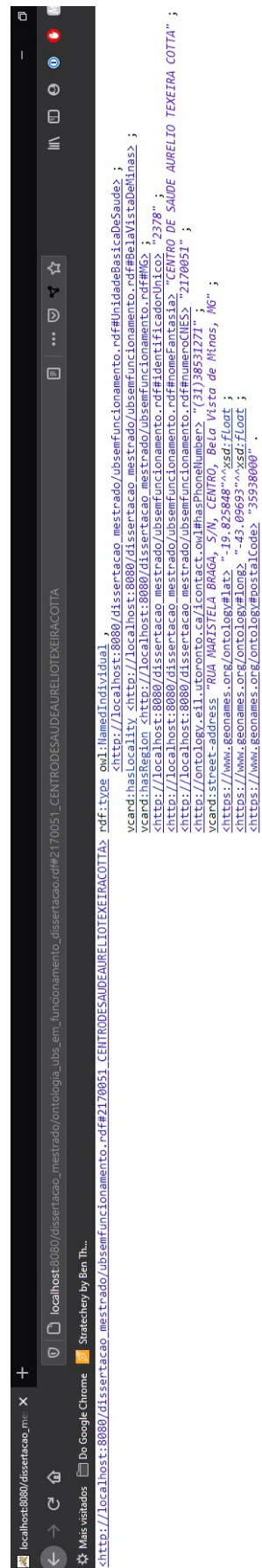


Figura 15 – Representação RDF do Centro de Saúde Aurelio Texeira Cotta visualizado em um navegador

Código 7 – Consulta SPARQL para responder à CQ5 - “Qual é o nome fantasia da Unidade Básica de Saúde?”



QUERY RESULTS

Table Raw Response

Showing 1 to 1 of 1 entries

Search:

	ubs	nome
1	ubs:2170051_CENTRODESAUDEAURELIOTEXEIRACOTTA	"CENTRO DE SAUDE AURELIO TEXEIRA COTTA"

Showing 1 to 1 of 1 entries

Figura 16 – Resultado da consulta SPARQL para saber qual o nome fantasia da UBS pesquisada

O Código 8 traz um exemplo da mesma consulta escrita no Código 7 sendo feita através de uma requisição HTTP e a resposta retornada no formato JSON.

```
http://localhost:8080/fuseki/ubs-em-funcionamento/?query=
PREFIX%20ubs%3A%3Chttp%3A%2F%2Flocalhost%3A8080%2
F dissertacao_mestrado%2Fubsementimento.rdf%23%3E%0A%0
ASELECT%20%3Fubs%20%3Fnome%0AWHERE%20%7B%0A%20%20%3Fubs%20
ubs%3AnomeFantasia%20%3Fnome.%0A%20%20FILTER%20%28%3Fubs%3
Dubs%3A2170051_CENTRODESAUDEAURELIOTEXEIRACOTTA%29%0A%7D

{
  "head": {
    "vars": [
      "ubs",
      "nome"
    ]
  },
  "results": {
    "bindings": [
      {
        "ubs": {
          "type": "uri",
```

```

        "value": "http://localhost:8080/
            dissertacao_mestrado/ubsemfuncionamento.rdf
            #2170051_CENTRODESAUDEAURELIOTEXEIRACOTTA"
    },
    "nome": {
        "type": "literal",
        "value": "CENTRO DE SAUDE AURELIO TEXEIRA COTTA"
    }
}
]
}
}

```

Código 8 – Consulta SPARQL feita através de *endpoint* SPARQL para responder à CQ5 - “Qual é o nome fantasia da Unidade Básica de Saúde?”

A ontologia, arquivo RDF, regras de transformação e documentos relevantes relacionados a implementação apresentada nessa seção podem ser encontrados em <https://git.io/JU2UF>.

6- Discussões

Nessa seção são levantados pontos de discussão identificados durante o desenvolvimento dessa pesquisa, relevantes aos temas de Dados Abertos Ligados e Dados Abertos Governamentais.

6.1- Publicação de um Arquivo de Dados x Modelagem e Publicação de Dados de um Domínio

Nesse trabalho foram apresentadas duas abordagens distintas de como evoluir na publicação de dados abertos para Dados Abertos Ligados. Para os dados das eleições, a decisão decorreu no desenvolvimento de uma modelagem mais extensa do domínio eleitoral, que necessitava da utilização de diferentes conjuntos de dados para sua representação. Para os dados das UBS em funcionamento, um único conjunto de dados, mais simples, foi utilizado e publicado como Dados Abertos Ligados.

Não existe abordagem certa e errada neste caso, sendo uma decisão de projeto dos responsáveis pela publicação em relação ao que eles desejam representar. Durante o desenvolvimento desse trabalho foi possível notar que a abordagem de publicar um único arquivo é mais simples e direta que a de modelar um domínio completo. Contudo, há valor em adotar a abordagem que demanda uma modelagem do domínio. Ao modelar um domínio completo representa-se uma parte maior do conhecimento e, por consequência, expande-se as conexões entre os dados e as possibilidades de navegação entre eles. Isso torna a publicação como dados ligados ainda mais efetiva, pois terão mais recursos e informações sobre eles sendo representados.

Ao decidir representar o conhecimento presente em um domínio mais amplo, os desenvolvedores e responsáveis pela publicação dos dados terão de buscar por todos os conjuntos de dados que compõem o que se deseja representar. Neste sentido, realizar um processo de integração entre eles se faz necessário, seja a nível conceitual, como realizado neste trabalho por meio da criação do modelo ER, ou diretamente integrando

todos os dados em um único conjunto ou estrutura (como um banco de dados relacional). Em nossa visão, esta integração traz um ganho de confiabilidade e consistência entre os dados, pois realizando esse processo se obterá uma garantia maior de que os recursos que estão sendo tratados são únicos e representam a mesma coisa para o domínio inteiro.

Um exemplo para tornar a discussão supracitada mais tangível refere-se a redundância de dados. Nos conjuntos de dados sobre as eleições existiam vários arquivos que continham dados repetidos sobre os candidatos. Ao analisar e modelar todos os dados de forma integrada foi possível identificar essa repetição e em nossa modelagem buscar os dados mais relevantes para compor aquela classe/conceito, independente do arquivo de origem do dado. Com uma abordagem baseada em ontologia, cada candidato estava representado de maneira única, tendo uma URI única que diz respeito a ele, atendendo aos requisitos dos Dados Abertos Ligados. Caso a publicação fosse realizada para cada arquivo de dados de forma separada, possivelmente um mesmo recurso seria publicado em diferentes domínios e URIs. Na aplicação do conjunto de dados eleitoral, por exemplo, um candidato seria representado por várias URIs diferentes. Essa questão minimiza um dos principais efeitos positivos da publicação como Dados Abertos Ligados, que é a possibilidade de, ao acessar um recurso, saber que ele tem todas as informações e ligações relevantes com outros recursos.

No entanto, mesmo utilizando a abordagem de publicação direta de um único arquivo, é possível minimizar esse efeito. Para isso é necessário um esforço dos responsáveis para mapear que um recurso em uma URI é o mesmo que outro em outra URI, utilizando a propriedade `owl:sameAs`. Assim poderia ser dito que uma URI de candidato que está em um domínio se refere a esse mesmo candidato em uma URI que está em outro domínio, permitindo aos usuários do recurso navegarem pelos dados e identificar todo o conhecimento representado.

6.2- Evolução nos Problemas da Publicação de Dados Abertos Ligados

Como ressaltado na introdução deste trabalho, a publicação de Dados Abertos Ligados possui uma série de barreiras que fazem com que o padrão e seus formatos relacionados tenham baixa adesão. O que vemos como preocupante em relação a isso é

a baixa evolução em soluções para esse problema ao longo dos anos.

Ao pesquisar a literatura verificou-se que os problemas listados já apareciam quando as iniciativas de dados abertos começaram a se popularizar, no final da década de 2000 e início da década de 2010. Mesmo assim, em 2020 ainda há empecilhos como a falta de um guia, a baixa adesão da comunidade de desenvolvimento e usuários, e principalmente o alto nível de conhecimento técnico necessário para que se faça uma publicação de Dados Abertos Ligados. Além de complexidade técnica alta, é necessário também um conhecimento multidisciplinar de diferentes áreas da computação e representação do conhecimento.

Enquanto a barreira de entrada para esse tipo de iniciativa for elevada, dificilmente ela decolará. Uma maneira de tentar endereçar esses problemas é por meio da criação de ferramentas que retirem essa necessidade de conhecimento muito especializado e que seja de fácil utilização (*user-friendly*) não só para a comunidade acadêmica, mas também para público externo, como servidores públicos que possam utilizar esse tipo de ferramenta no processo de publicação de dados dos órgãos em que trabalham. Sem essa facilitação é difícil que existam mais conjuntos de dados seguindo esses padrões e, por consequência, interesse da comunidade para trabalhar com esses dados.

6.3- Cenário Brasileiro de Dados Governamentais Abertos Ligados

Assim como o cenário internacional, no Brasil não há diferenças quando se olha para a adesão e utilização dos padrões da Web Semântica. Mesmo tendo toda uma estrutura de suporte e padronização recomendando a utilização, a realidade é que a publicação de Dados Abertos Ligados encontra-se distante da realidade da maior parte dos órgãos governamentais brasileiros.

Além das barreiras já apontadas, existem outras particulares no Brasil. Um exemplo disso é a diferença da língua. Como publicar dados que se liguem com conjuntos de dados do restante do mundo e que tenham significado explícito estando num país em que se fala português? Como se conectar com todo o conhecimento e dados produzidos e publicados na língua inglesa, que são dominantes na LOD *Cloud*? A Web Semântica endereça esse tipo de problema com o uso do predicado `owl:sameAs`, para

relacionar recursos que significam a mesma coisa. Porém, mesmo assim ainda existe a complexidade de representar conceitos que são únicos ou só fazem sentido para o contexto brasileiro e fazer com que eles sejam entendidos por meio de metadados.

Olhando para o cenário atual, há problemas com dados abertos publicados utilizando formatos tradicionais como CSV. Um destes problemas é a qualidade das documentações produzidas, que dificulta sobremaneira quem queira trabalhar com os dados a entender o significado publicado nos conjuntos de dados. Entendemos isso como uma maneira de tornar o dado inacessível, pois não é possível compreender o que está representado, dada a falta de uma documentação que explique os dados.

Para o Brasil, além de se buscar uma evolução na escala de publicação dos dados é necessário também que se busque um amadurecimento em relação ao conteúdo existente atualmente, tornando os conjuntos de dados disponíveis realmente abertos, respeitando os requisitos dos dados aberto. Além de respeitar os requisitos, o amadurecimento também deve passar por uma evolução na maneira com que os dados são gerenciados e administrados, seguindo por exemplo os princípios FAIR, que buscam guiar o gerenciamento e administração de dados com foco em permitir que máquinas possam encontrar, identificar e utilizar esses dados, além de suportar sua reutilização. Para isso ele define uma série de critérios para que os dados sejam encontráveis (em inglês, *findable*), acessíveis (em inglês, *accessible*), interoperáveis (em inglês, *interoperable*) e reutilizáveis (em inglês, *reusable*) (Wilkinson et al., 2016). Também é adequado que haja uma promoção da expansão da abertura dos dados, algo fundamental para evolução e amadurecimento da nossa própria democracia, que pode se dar por meio do empoderamento da sociedade.

7- Considerações Finais

Este trabalho apresenta uma metodologia baseada em ontologia com etapas a serem seguidas para a publicação de Dados Abertos Governamentais como Dados Abertos Ligados como forma de integrar dados disponibilizados por governos à Web Semântica. O desenvolvimento de ontologias é realizado utilizando duas metodologias consolidadas, que são a NeOn e a Ontology Development 101. As etapas propostas foram aplicadas em dois experimentos utilizando conjuntos de dados relacionados ao domínio eleitoral brasileiro e ao domínio da saúde no Brasil, como forma de instanciar a aplicação do processo com casos reais. As etapas propostas vão desde a identificação e análise dos conjuntos de dados disponíveis até o consumo deles como Dados Abertos Ligados, passando pela criação e utilização de ontologias e ligação com outros conjuntos de dados publicados em RDF na Web.

Como forma de aplicação da proposta foi conduzido o processo para a publicação de conjuntos de dados governamentais reais como Dados Abertos Ligados utilizando abordagens distintas. Para os dados do domínio eleitoral, a publicação representou uma eleição no Brasil, que é composta por dados espalhados por diferentes conjuntos de dados. Para o domínio da saúde, a publicação partiu de um único conjunto contendo dados sobre as Unidades Básicas de Saúde em funcionamento no País. Com essas aplicações foi possível atestar que o guia proposto pode ser aplicado tanto para uma abordagem mais extensa, de publicação de dados de um domínio completo, quanto para uma abordagem mais direta, de publicação de um único arquivo de dados.

Os resultados da primeira aplicação, que utilizou dados eleitorais, foram: (i) um modelo entidade-relacionamento sobre os dados eleitorais utilizados; (ii) a criação de uma ontologia que representa as eleições brasileiras com um escopo reduzido de 1.618 indivíduos, contendo dados da eleição presidencial de 2018; (iii) a disponibilização dos dados em um arquivo RDF/XML; (iv) a disponibilização de consultas SPARQL que podem ser utilizadas para responder aos requisitos da ontologia, escritos em forma de Questões de Competência.

Essa primeira aplicação foi fundamental para a evolução das etapas propostas, pois através dela foi possível identificar pontos de melhoria e entender os desafios na

conversão de dados de um formato tradicional para dados ligados. Houveram dois principais desafios encontrados para esse conjunto. O primeiro diz respeito ao entendimento de como dados espalhados por múltiplos arquivos se conectam. O segundo foi a complexidade maior em modelar semanticamente recursos que não são únicos e vão se alterando ao longo do tempo. Como exemplo desse segundo caso temos candidatos que concorrem em diferentes eleições e podem ter mudado de partido e mantido seus bens de uma eleição para outra.

Para a segunda aplicação, na qual foram utilizados dados de distribuição das Unidades Básicas de Saúde em funcionamento no Brasil, houveram os seguintes resultados: (i) criação de uma ontologia para representação das informações de uma UBS contendo dados de 500 unidades; (ii) a disponibilização dos dados em um arquivo RDF/XML; (iii) a disponibilização de consultas SPARQL que podem ser utilizadas para responder aos requisitos da ontologia, escritos em forma de Questões de Competência.

Com a segunda aplicação houve a validação que o processo proposto é genérico o suficiente para funcionar com dados de diferentes domínios e com diferentes estratégias de integração. Isso porque ao invés de modelar um domínio completo como na primeira aplicação, optou-se por fazer uma publicação direta de um arquivo de dados CSV como dados ligados. Um empecilho observado nessa aplicação foi a falta de uma documentação sobre os dados, sendo necessário excluir alguns dados do modelo semântico por não ser possível identificar seu significado.

Vale ressaltar que os produtos desse trabalho estão em conformidade com os princípios dos dados abertos citados na Seção 2.3. Os arquivos RDF, ontologia (no formato OWL) e todos os outros documentos de apoio estão disponíveis de maneira aberta no Github (respeitando o requisito de acesso) em um repositório que usa uma licença aberta (a MIT License¹). Por estarem publicados no formato RDF, naturalmente se respeita os requisitos de facilitar a leitura por máquinas e o de utilizar um formato aberto, pois é o que o padrão oferece.

Outros artefatos produzidos para ambas as aplicações foram os Documentos de Especificação de Requisitos das ontologias e as regras de transformação utilizadas para carregar os dados presentes nos arquivos para as ontologias. Todos os artefatos produzidos podem ser acessados em https://github.com/aiquis/dissertacao_mestrado. A disponibilização de todo o material produzido nesse trabalho permite que outros pesqui-

¹<https://opensource.org/licenses/MIT>

sadores e pessoas interessadas no assunto possam replicar os experimentos realizados e estendê-los, evoluindo o que foi produzido.

O trabalho desenvolvido possui algumas limitações técnicas devido às ferramentas escolhidas para a execução dos experimentos e limitações por decisões de projeto. A principal limitação técnica está relacionada a utilização da ferramenta Cellfie. Durante o desenvolvimento das aplicações foi notado que a ferramenta apresenta problemas de desempenho ao lidar com arquivos com muitos registros. Houve a tentativa de contornar esse problema com o aumento na memória disponível para uso, mas que não foi suficiente para lidar com alguns dos arquivos utilizados. Devido a essa questão foi necessário diminuir o tamanho dos conjuntos de dados utilizados. Outra limitação da ferramenta foi não permitir a realização de agregação dos dados (em somas por exemplo), o que fez com que fosse necessária a inserção de alguns dados manualmente na ontologia.

As limitações existentes por decisões de projeto do processo foram duas. A primeira foi a opção por não consultar especialistas dos domínios para a criação das ontologias. Essa escolha foi feita pois, para fins desse trabalho no qual o foco era no processo e não na ontologia em si, acreditamos que uma representação com base no conhecimento adquirido através dos dados disponíveis e de pesquisa sobre os domínios fosse o suficiente. A segunda foi a priorização da simplicidade do processo em relação às melhores práticas para publicação de dados ligados. Isso se reflete na opção por publicar os dados utilizando apenas um arquivo RDF/XML estático. As limitações apresentadas podem representar oportunidades de melhorias nos produtos desse trabalho, evoluindo a contribuição feita.

Este trabalho tem contribuição para a área de Web Semântica e dos Dados Abertos Ligados, pois uma das principais barreiras para a publicação de Dados Abertos Ligados, e por consequência, o crescimento em sua adoção, é a falta de um guia que auxilie em todas as etapas do processo. Espera-se diminuir esta barreira com a proposta deste trabalho, que poderá ser utilizada por órgãos governamentais para que consigam atingir níveis mais altos de maturidade na publicação de dados abertos.

Os resultados obtidos com esse trabalho podem ser evoluídos e expandidos em trabalhos futuros. Entre as possibilidades estão: (i) publicação dos conjuntos de dados completos para os dois experimentos realizados; (ii) ligação extensa dos indivíduos dos experimentos com recursos de outros conjuntos de dados; (iii) validação e evolução das

ontologias com especialistas dos domínios; (iv) publicação dos dados e disponibilização de *endpoint* SPARQL em um servidor exposto para a web para acesso externo; (v) aplicação das melhores práticas da publicação de dados ligados para permitir negociação de conteúdo e redirecionamentos; (vi) criação de uma interface web amigável para acesso aos dados por usuários não-técnicos; (vii) evolução das ontologias e conexão com mais conjuntos de dados ligados para permitir que os dados possam ser publicados na Dados Abertos Ligados *Cloud*.

Esse trabalho também resultou no desenvolvimento do artigo “Em direção a uma ontologia das Eleições Brasileiras” que foi publicado no XII Seminário de Pesquisa em Ontologias no Brasil (ONTOBRAS). Também foi produzido o artigo “*An Ontology-Based Approach to Publish Brazilian Election Data as Linked Open Data*”, em fase de revisão final, e que será submetido para a revista *International Journal of Metadata, Semantics and Ontologies*. Os estudos sobre as Unidades Básicas de Saúde geraram o artigo “Metodologia para Análise de Dados Abertos de Saúde: um Estudo sobre as Unidades Básicas de Saúde do Rio de Janeiro”, também em fase de revisão final e que será submetido para a revista *Cadernos de Saúde Pública*.

Referências Bibliográficas

- Araújo, L. d. R. and de Souza, J. F. (2011). Aumentando a transparência do governo por meio da transformação de dados governamentais abertos em dados ligados. *Revista Eletrônica de Sistemas de Informação*, 10(1):165–167.
- Arturo, A., Caraballo, M., Nunes, B. P., and Casanova, M. A. (2015). DRX: A LOD dataset interlinking recommendation tool. *Undefined*, 1:1–5.
- Attard, J., Orlandi, F., Scerri, S., and Auer, S. (2015). A systematic review of open government data initiatives. *Government Information Quarterly*, 32(4):399–418.
- Berners-Lee, T. (2006). Linked Data - Design Issues. Acessado em 08/09/2019.
- Berners-Lee, T., Cailliau, R., Luotonen, A., Nielsen, H. F., and Secret, A. (1994). The World-Wide Web. *Communications of the ACM*, 37(8):76–82.
- Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American*, 284:28–37.
- Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked Data - The Story So Far. *International journal on Semantic Web and Information Systems*, 5(3):1–22.
- Brandt, M. B., Vidotti, S. A. B. G., and Segundo, J. E. S. (2018). Modelo de dados abertos conectados para informação legislativa. *Informação & Sociedade: Estudos*, 28(2):149–161.
- Brasil (2011). Lei nº 12.527, de 18 de novembro de 2011 - regula o acesso a informações previsto no inciso xxxiii do art. 5º, no inciso ii do § 3º do art. 37 e no § 2º do art. 216 da constituição federal; altera a lei no 8.112, de 11 de dezembro de 1990; revoga a lei no 11.111, de 5 de maio de 2005, e dispositivos da lei no 8.159, de 8 de janeiro de 1991; e dá outras providências.
- Breitman, K., Salas, P., Casanova, M. A., Saraiva, D., Gama, V., Viterbo, J., Magalhães, R. P., Franzosi, E., and Chaves, M. (2012). Open government data in Brazil. *IEEE Intelligent Systems*, 27(3):45–49.

- Caraballo, A. A. M., Arruda, N. M., Nunes, B. P., Lopes, G. R., and Casanova, M. A. (2014). TRTML - A Tripleset Recommendation Tool Based on Supervised Learning Algorithms. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8798, pages 413–417.
- Carbonaro, A. (2019). Linked data and semantic web technologies to model context information for policy-making. *Journal of Ambient Intelligence and Humanized Computing*, (0123456789).
- da Silva, J. F. C. (2018). Etl4lod+: evolução do suporte ao ciclo de publicação de dados conectados.
- De Mendonça, R. R., Da Cruz, S. M. S., De La Cerda, J. F., Cavalcanti, M. C., Cordeiro, K. F., and Campos, M. L. M. (2013). LOP - Capturing and Linking Open Provenance on LOD cycle. *Proceedings of the 5th Workshop on Semantic Web Information Management, SWIM 2013*, (March 2017).
- De Oliveira, E. F. and Silveira, M. S. (2018). Open government data in Brazil: A systematic review of its uses and issues. *ACM International Conference Proceeding Series*.
- de Oliveira Araújo, L. S., Santos, M. T., and Silva, D. A. (2015). The Brazilian federal budget ontology. In *Proceedings of the 7th International Conference on Management of computational and collective intelligence in Digital EcoSystems - MEDES '15*, number October, pages 85–89, New York, New York, USA. ACM Press.
- Domingue, J., Fensel, D., and Hendler, J. A. (2011). *Handbook of Semantic Web Technologies*. Springer Berlin Heidelberg, Berlin, Heidelberg.
- ePing (2011). Vocabulário Controlado do Governo Eletrônico. Acessado em 21/02/2019.
- Falbo, R. D. A. (2014). SABiO : Systematic Approach for Building Ontologies. *1st Joint Workshop Onto.Com/ODISE on Ontologies in Conceptual Modeling and Information Systems Engineering*.
- Fernández-López, M., Gómez-Pérez, A., and Juristo, N. (1997). Methontology: from ontological art towards ontological engineering.
- Fleiner, R. (2018). Linking of open government data. *SACI 2018 - IEEE 12th International Symposium on Applied Computational Intelligence and Informatics, Proceedings*, pages 479–483.

- Freitas, R., Rocha, C., Braga, O., Lopes, G., Monteiro, O., and Oliveira, M. (2017). Using Linked Data in the Data Integration for Maternal and Infant Death Risk of the SUS in the GISSA Project. *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web - WebMedia '17*, pages 193–196.
- Genesereth, M. R. and Nilsson, N. J. (1987). *Logical Foundations of Artificial Intelligence*. Elsevier.
- Grüninger, M. and Fox, M. S. (1995). Methodology for the Design and Evaluation of Ontologies.
- Guarino, N. and Giaretta, P. (1995). Ontologies and knowledge bases: Towards a terminological clarification. In *Towards very Large Knowledge bases: Knowledge Building and Knowledge sharing*, pages 25–32. IOS Press.
- Guarino, N. and Oberle, D. (2009). Handbook on Ontologies. pages 1–17.
- Hausenblas, M. and Kim, J. (2015). 5-star open data.
- Heath, T. and Bizer, C. (2011). Linked Data: Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1(1):1–136.
- Horridge, M. and Patel-Schneider, P. F. (2012). OWL 2 Web Ontology Language Manchester Syntax (Second Edition). Acessado em 27/03/2020).
- Hoxha, J. and Brahaj, A. (2011). Open Government Data on the Web: A Semantic Approach. In *2011 International Conference on Emerging Intelligent Data and Web Technologies*, pages 107–113. IEEE.
- Janssen, M., Charalabidis, Y., and Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, 29(4):258–268.
- Kalampokis, E., Tambouris, E., and Tarabanis, K. (2011). A classification scheme for open government data: towards linking decentralised data. *International Journal of Web Engineering and Technology*, 6(3):266.
- Klein, E., Gschwend, A., and Neuron, A. C. (2016). Towards a linked data publishing methodology. *Proceedings of the 6th International Conference for E-Democracy and Open Government, CeDEM 2016*, pages 188–196.

- Kučera, J., Chlapek, D., and Nečaský, M. (2013). Open government data catalogs: Current approaches and quality perspective. In Kő, A., Leitner, C., Leitold, H., and Prosser, A., editors, *Technology-Enabled Innovation for Democracy, Government and Governance*, pages 152–166, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Leme, L. A. P. P., Lopes, G. R., Nunes, B. P., Casanova, M. A., and Dietze, S. (2013). Identifying Candidate Datasets for Data Interlinking. pages 354–366.
- Lopes, G., Vidal, V., and Oliveira, M. (2016). A Framework for Creation of Linked Data Mashups. *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web - Webmedia '16*, (Figura 1):327–330.
- Magalhães, R. P., de Macêdo, J. A. F., and Vidal, V. M. P. (2011). Linked Data: Construindo um Espaço de Dados Global na Web. c:61–89.
- Matheus, R., Ribeiro, M. M., and Vaz, J. C. (2015). Brazil Towards Government 2.0: Strategies for Adopting Open Government Data in National and Subnational Governments. In *Case Studies in e-Government 2.0*, volume 55, pages 121–138. Springer International Publishing, Cham.
- Milić, P., Veljković, N., and Stoimenov, L. (2018). Semantic Technologies in e-government: Toward Openness and Transparency. volume 24 of *Public Administration and Information Technology*, pages 55–66. Springer International Publishing, Cham.
- Noy, N. F. and McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology. pages 1–25.
- Open Gov Data (2007). The 8 Principles of Open Government Data. Acessado em 10/04/2020.
- Open Knowledge Foundation (2015). Open Definition 2.1. Acessado em 25/08/2019.
- Pattuelli, M. C., Provo, A., and Thorsen, H. (2015). Ontology Building for Linked Open Data: A Pragmatic Perspective. *Journal of Library Metadata*, 15(3-4):265–294.
- Pontes, F. (2018). Campanha de Bolsonaro entrega prestação de contas final ao TSE. *Agência Brasil*. Acessado em 05/05/2019.
- RDF Working Group (2014). Resource Description Framework (RDF).

- Ristoski, P. and Paulheim, H. (2016). Semantic Web in data mining and knowledge discovery: A comprehensive survey. *Journal of Web Semantics*, 36:1–22.
- Sauermann, L. and Cyganiak, R. (2008). Cool URIs for the Semantic Web. Acessado em 26/04/2020.
- Senado Federal (2019). Tesouro do Senado Federal. Acessado em 21/02/2019.
- Shadbolt, N., O'Hara, K., Berners-Lee, T., Gibbins, N., Glaser, H., Hall, W., and Schraefel, M. (2012). Linked Open Government Data: Lessons from Data.gov.uk. *IEEE Intelligent Systems*, 27(3):16–24.
- Shadbolt, N., O'Hara, K., Salvadores, M., and Alani, H. (2011). *eGovernment*, pages 849–910. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Sheridan, J. and Tennison, J. (2010). Linking UK Government Data. *Ldow 2010*, pages 1–4.
- SPARQL Working Group (2013). SPARQL Query Language for RDF.
- Studer, R., Benjamins, V., and Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, 25(1):161 – 197.
- Suárez-Figueroa, M. C. and Gómez-Pérez, A. (2012). Ontology Requirements Specification. In *Ontology Engineering in a Networked World*, pages 93–106. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Suárez-Figueroa, M. C., Gómez-Pérez, A., and Fernández-López, M. (2012). The NeOn Methodology for Ontology Engineering. In *Ontology Engineering in a Networked World*, pages 9–34. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Suárez-Figueroa, M. C., Gómez-Pérez, A., and Villazón-Terrazas, B. (2009). How to write and use the ontology requirements specification document. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5871 LNCS(PART 2):966–982.
- Tribunal Superior Eleitoral (2010). Tesouro da Justiça Eleitoral.
- Tribunal Superior Eleitoral (2018). Divulgação de Candidaturas e Contas Eleitorais. Acessado em 05/05/2019.

Tribunal Superior Eleitoral (2019). Glossário Eleitoral. Acessado em 20/02/2019.

Verona, L. D., Lopes, G. R., and Campos, M. L. M. (2019). Using Government Data to Uncover Political Power and Influence of Contemporary Slavery Agents in Brazil. volume 926, pages 123–138.

W3C OWL Working Group (2012). OWL - Semantic Web Standards. Acessado em 05/01/2019.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018.

Yu, H. Q., Pedrinaci, C., Dietze, S., and Domingue, J. (2012). Using linked data to annotate and search educational video resources for supporting distance learning. *IEEE Transactions on Learning Technologies*, 5(2):130–142.