

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

**Instituto de Matemática
Instituto Tércio Pacitti de Aplicações e Pesquisas
Computacionais**

Programa de Pós-Graduação em Informática

GLAUCIA BOTELHO DE FIGUEIREDO

**LigADOS: INTERLIGAÇÃO DE CONJUNTOS DE DADOS EM
PLATAFORMAS DE DADOS ABERTOS NA WEB SEMÂNTICA**

Rio de Janeiro
2020

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE MATEMÁTICA
INSTITUTO TÉRCIO PACITTI DE APLICAÇÕES E PESQUISAS COMPUTACIONAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

GLAUCIA BOTELHO DE FIGUEIREDO

LigADOS: INTERLIGAÇÃO DE CONJUNTOS DE DADOS EM
PLATAFORMAS DE DADOS ABERTOS NA WEB SEMÂNTICA

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Informática do Instituto de Matemática e do Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários para obtenção do título de Mestre em Informática.

Orientadora: Profa. Maria Luiza Machado Campos, Ph.D
Coorientadora: Profa. Kelli de Faria Cordeiro, DSc.

Rio de Janeiro
2020

CIP - Catalogação na Publicação

F475l Figueiredo, Glaucia Botelho de
LigADOS: interligação de conjuntos de dados em
plataformas de dados abertos na web semântica /
Glaucia Botelho de Figueiredo. -- Rio de Janeiro,
2020.
138 f.

Orientadora: Maria Luiza Machado Campos.
Coorientadora: Kelli de Faria Cordeiro.
Dissertação (mestrado) - Universidade Federal do
Rio de Janeiro, Instituto Tércio Pacitti de
Aplicações e Pesquisas Computacionais, Programa de
Pós-Graduação em informática, 2020.

1. Interligação de conjuntos de dados. 2.
Interligação semântica de metadados. 3. Plataformas
de dados abertos. 4. Dados abertos conectados. 5.
FAIR Data Point. I. Campos, Maria Luiza Machado,
orient. II. Cordeiro, Kelli de Faria , coorient.
III. Título.

LigADOS: Interligação de Conjuntos de Dados em Plataformas de Dados Abertos na Web Semântica

Glaucia Botelho de Figueiredo

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Informática do Instituto de Matemática, e Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários para obtenção do título de Mestre em Informática.

Aprovado em: Rio de Janeiro, 11 de dezembro de 2020 por:

Maria Luiza Campos

Profa. Maria Luiza Machado Campos, Ph.D., UFRJ (Presidente)

PARTICIPAÇÃO POR VIDEOCONFERÊNCIA

Profa. Kelli de Faria Cordeiro, D.Sc., MB

PARTICIPAÇÃO POR VIDEOCONFERÊNCIA

Profa. Giseli Rabello Lopes, D.Sc., UFRJ

PARTICIPAÇÃO POR VIDEOCONFERÊNCIA

Profa. Maria Cláudia Reis Cavalcanti, D.Sc., IME/RJ

PARTICIPAÇÃO POR VIDEOCONFERÊNCIA

Profa. Bernadette Farias Lóscio, D.Sc., UFPE

PARTICIPAÇÃO POR VIDEOCONFERÊNCIA

Prof. João Luiz Rebelo Moreira, Ph.D., UTwente (Holanda)

*Ao meu esposo Renato e aos meus pequenos Nicolás e Bernardo,
meus amores.*

AGRADECIMENTO

A Deus Pai Todo-Poderoso, pelo novo desafio, pela oportunidade de aprendizado muito além do científico, por ter colocado no meu caminho as professoras Maria Luiza e Kelli que me orientaram na descoberta de novos conhecimentos, pela compreensão e parceria daqueles que diretamente compartilharam a labuta diária comigo, pelas pessoas que estiveram ao meu lado mesmo que fisicamente distantes, pelas amizades fortalecidas, pelas novas amizades, pelos “Braços Amigos” que me deram forças e sustentação para concluir a missão.

Obrigada Deus Pai!

Se queremos melhorar o futuro, devemos perturbar o presente.
(Catherine Booth)

RESUMO

O movimento de abertura de dados tem sido pautado na necessidade de atendimento de leis de acesso à informação para reforçar responsabilidade e transparência, que estabelecem a necessidade de disponibilizar dados sobre atividades públicas e privadas aos cidadãos e à sociedade em geral, bem como resultados de pesquisas científicas. A disponibilização de dados ao público pode variar desde um simples *site* na Internet até aplicativos sofisticados de acesso aos dados. Neste contexto, uma das opções é a construção de um portal de dados abertos utilizando *softwares* de plataformas de dados abertos, como o CKAN. Nos últimos anos, tem havido uma rápida proliferação desses portais. Nas plataformas de dados abertos, os conjuntos de dados normalmente são descritos por metadados com pouca semântica. Em função disso, os conjuntos de dados muitas vezes permanecem como “silos de dados”, sem conexão explícita ou mecanismo de integração de dados, tornando difícil para os usuários localizar e inter-relacionar fontes de dados relevantes. Em contraste, a Web Semântica apoia-se em uma forma de modelar e representar dados com o intuito de estabelecer inter-relacionamentos entre os dados, acompanhados de descritores mais ricos. Com base neste cenário, esta dissertação apresenta LigADOS, uma abordagem para estabelecer interligações entre conjuntos de dados considerando seu conteúdo e os metadados associados. LigADOS é baseada nos princípios da Web Semântica, associada a soluções de dados conectados e suas tecnologias, para criar interligações entre conjuntos de dados abertos, modelados como dados conectados e publicados em plataformas de dados abertos como o CKAN e outras. A viabilidade da proposta é avaliada através de um protótipo usando conjuntos de dados extraídos de instâncias de portais de dados reais. Sem o uso de LigADOS, são observáveis somente relações geradas entre conjuntos de dados, a partir de elementos de metadados coincidentes, que usualmente não estão associados a vocabulários, ontologias ou tesouros. LigADOS suporta a geração de diversas interligações entre os conjuntos de dados, que se baseiam em significados definidos em vocabulários, ontologias ou tesouros, fornecendo suporte para resultados mais expressivos em resposta a buscas e consultas, além do suporte à navegação pelas interligações de dados. Os resultados do uso de LigADOS sugerem que há interligações entre conjuntos de dados que não estão explícitas ao público em geral; assim como há conjuntos de dados que, inicialmente, parecem ter inter-relações, mas ao se aprofundar na semântica dos dados, não se verificam as supostas interligações.

Palavras-chave: Interligação de Conjuntos de Dados. Interligação Semântica de Metadados. Plataformas de Dados Abertos. Dados Abertos Conectados. FAIR Data Point. Web de Dados. Web Semântica.

ABSTRACT

The Open Data movement has been guided by the need to comply with laws to enforce transparency and accountability, which establish the need to make public and private sector activities data available to citizens and the society in general, as well as results of scientific research. The availability of data to the public can vary from a simple website on the Internet to sophisticated data access applications. In this context, one of the options is the construction of an open data portal using open data platforms software, such as CKAN. In recent years, there has been a rapid proliferation of these portals. On open data platforms, datasets are usually described by metadata with little semantics. As a result, datasets often remain as “data silos”, with no explicit connection or data integration mechanism, making it difficult for users to locate and interrelate relevant data sources. In contrast, the Semantic Web focuses on a way of modeling and representing data in an easier manner to establish interrelationships between data, accompanied by richer descriptors. Based on this scenario, this dissertation presents LigADOS, an approach to establish interconnections between datasets considering their content and associated metadata. LigADOS is based on the principles of the Semantic Web and associated linked data solutions and technologies, to create interconnections between open datasets, modeled as linked data and published on data portal platforms such as CKAN and others. The feasibility of the proposal is evaluated through a prototype using datasets extracted from instances of real data portals. Without the use of LigADOS, only relations generated between datasets are observable, from elements of coincident metadata, which are not usually associated with vocabularies, ontologies or thesaurus. LigADOS supports the generation of several interlinkages between datasets, which are based on meanings defined in vocabularies, ontologies or thesaurus, providing support for more expressive results in responses to searches and queries, in addition to support for navigation through data interlinkages. The results of using LigADOS suggest that there are interlinkages between datasets that are not explicit to the public; just as there are datasets that initially seem to have interlinkages, but when going deeper into the semantics of the data, the supposed interlinkages are not verified.

Keywords: Dataset Interlinking. Semantic Metadata Interlinking. Open Data Platforms. Open Linked Data. FAIR Data Point. Web of Data. Semantic Web.

LISTA DE FIGURAS

Figura 1: Esquema de organização dos conjuntos de dados e suas distribuições em uma instância CKAN.....	32
Figura 2: Conjunto de dados com oito distribuições associadas: dois em XLS, cinco em PDF e um em CSV	33
Figura 3: Uma estrutura e uma instância de grafo RDF com um nó sujeito conectado por dois arcos direcionados e rotulados a dois objetos distintos	42
Figura 4: Uma tripla RDF serializada em N-Triples	42
Figura 5: Representação de tripla embutida serializada em Turtle*	46
Figura 6: Resumo dos princípios FAIR.....	49
Figura 7: Fluxo de atividades do processo para tornar dados e metadados FAIR.....	50
Figura 8: Esquema de metadados em camadas para um FDP	52
Figura 9: Propriedades da <i>FDP Ontology</i>	52
Figura 10: Exemplo de descrição de recurso	54
Figura 11: Exemplo de interligação por meio de uma propriedade.....	54
Figura 12: Exemplo de descoberta de novas informações por meio da navegação	55
Figura 13: Esquema de interligação semântica indireta gerada por LigADOS para uma instância de portal.....	67
Figura 14: Processo de interligação e disponibilização das interligações de dados de LigADOS	69
Figura 15: Identificação de conceito que categoriza termos utilizados por duas distribuições de conjuntos de dados.....	71
Figura 16: Esquema de metadados no nível de registro de dados	71
Figura 17: Utilização do vocabulário FOAF	72
Figura 18: Utilização do vocabulário DCAT	73
Figura 19: Utilização do vocabulário DCMi Metadata Terms	74
Figura 20: Utilização do vocabulário VOAF	74
Figura 21: Utilização do vocabulário VAAN	75
Figura 22: Estrutura das triplas do <i>Grafo de Interligações Semânticas</i> para cada conjunto de dados.....	77
Figura 23: Apresentação do <i>Grafo de Interligações Semânticas</i>	81
Figura 24: Tesouro “eionet-gemet” e os conceitos que foram utilizados para anotação semântica da distribuição	82

Figura 25: Estruturas das triplas geradas por LigADOS para identificadores de camadas de metadados.....	84
Figura 26: Estruturas das triplas geradas por LigADOS para representação dos relacionamentos hierárquicos entre as camadas de metadados	84
Figura 27: Estruturas das triplas para tipificar as camadas de metadados	85
Figura 28: Estrutura da tripla que caracteriza o metadado no nível de registro de dados	86
Figura 29: Estrutura da tripla que caracteriza a camada de metadados no nível de registro de dados, hierarquização entre as camadas de registro de dados e distribuição e a tipificação de um recurso que representa a camada de metadados no nível de registro de dados.....	87
Figura 30: Estrutura das triplas que caracterizam área temática do catálogo e conceitos dos conjuntos de dados.....	88
Figura 31: Estrutura das triplas da camada de Registro de Dados do <i>Grafo de Metadados FDP</i>	91
Figura 32: Esquema exemplificativo de interligações de dados de diferentes instâncias de portais em um FDP	92
Figura 33: Processo de suporte ontológico fornecido por LigADOS.....	93
Figura 34: Arquitetura de suporte à abordagem LigADOS	95
Figura 35: Extrato da instância CKAN mostrando os resultados gerados por LigADOS.....	96
Figura 36: Publicação do conjunto de dados “Vegetação de Mangue Brasileira” e sua distribuição “CSV - Vegetação de mangue” entre outras.....	104
Figura 37: Publicação original do conjunto de dados “Sistema Nacional de Informações Florestais - SNIF”	105
Figura 38: Interligações entre distribuições de conjuntos de dados por meio do uso de tesauro em comum.....	107
Figura 39: Interligações entre distribuições de conjuntos de dados que podem sinalizar possíveis pontos de interesse de análises.....	108
Figura 40: Possibilidades de análise pela compreensão dos assuntos contidos nas distribuições dos conjuntos de dados.....	109
Figura 41: Utilização da ontologia “vocab-loa”, de domínio específico, alinhada a um processo de negócio em particular	110
Figura 42: Ontologias “vocab-loa”, “dbpedia-ontology” e os tesouros “aims-agrovoc” e “eionet-gemet” e as classes/conceitos utilizadas pelas distribuições de dados	111
Figura 43: Publicação do conjunto de dados “Comercialização IBAMA 2017”.	112
Figura 44: Publicação do conjunto de dados “Comercialização IBAMA 2017” com destaque para suas distribuições e os conceitos do tesauro “aims-agrovoc” utilizados em comum por ambas	113

Figura 45: Duas distribuições de dados que utilizam “ChEBI Ontology” com destaque para as classes/conceitos que ligam ambas as distribuições.	115
Figura 46: Etiquetas semânticas que interligam conjuntos de dados publicados na instância de portal	116

LISTA DE QUADROS

Quadro 1: Expressões utilizadas no texto desta dissertação.....	21
Quadro 2: Comparativo das plataformas de dados abertos de código aberto	40
Quadro 3: Exemplos de vocabulários e classes para anotação semântica	47
Quadro 4: Estratégias das abordagens de interligação que não consideram os metadados associados aos conjuntos de dados	64
Quadro 5: Estratégias das abordagens de interligação que consideram os metadados associados aos conjuntos de dados	65
Quadro 6: Elementos de metadados extraídos para a camada de catálogo de dados	89
Quadro 7: Elementos de metadados extraídos para a camada de conjunto de dados	89
Quadro 8: Elementos de metadados extraídos para a camada de distribuições de dados	90
Quadro 9: Elementos de metadados extraídos para a camada de registro de dados.....	91
Quadro 10: Comparativo para os metadados da camada de catálogo para um FDP	130
Quadro 11: Comparativo para os metadados da camada de conjuntos de dados para um FDP	131
Quadro 12: Comparativo para os metadados da camada de distribuição para um FDP	132
Quadro 13: Distribuições de conjuntos de dados extraídos do Portal Brasileiro de Dados Abertos para execução dos experimentos.....	135
Quadro 14: Distribuições de conjuntos de dados extraídos do Portal de Dados Abertos sobre Agrotóxicos para execução dos experimentos	136

LISTA DE SIGLAS

API	Application Programming Interface
ChEBI	Chemical Entities of Biological Interest
CKAN	Comprehensive Knowledge Archive Network
CSV	Comma-separated values
DCAT	Data Catalog Vocabulary
DCMI	Dublin Core™ Metadata Initiative
DOI	Digital Object Identifier
FAIR	Findable, Accessible, Interoperable e Reusable
FDP	FAIR Data Point
FOAF	Friend of a Friend
GRECO	Grupo de Pesquisa em Engenharia do Conhecimento
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
JSON	JavaScript Object Notation
LOD	Linked Open Data
LDP	Linked Data Platform
OE	Objetivo Específico
ODP	Open Data Portal
OGDP	Open Government Data Portal
PDF	Portable Document Format
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
RDF/XML	Resource Description Framework/Extensible Markup Language
SKOS	Simple Knowledge Organization System
SPARQL	Simple Protocol and RDF Query Language
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
VANN	Vocabulary for Annotating Vocabulary
VOAF	Vocabulary of a Friend
VOID	Vocabulary of Interlinked Datasets
WEB	World Wide Web
W3C	World Wide Web Consortium
XLSX	Excel Microsoft Office Open XML Format Spreadsheet file

SUMÁRIO

1.	INTRODUÇÃO.....	15
1.1.	PROBLEMA.....	21
1.2.	OBJETIVOS.....	23
1.3.	ABORDAGEM DE SOLUÇÃO	24
1.4.	JUSTIFICATIVA	25
1.5.	METODOLOGIA DE PESQUISA	26
1.6.	ORGANIZAÇÃO DO TRABALHO	27
2.	WEB SEMÂNTICA, PLATAFORMAS DE DADOS ABERTOS E METADADOS	29
2.1.	METADADOS	29
2.2.	PLATAFORMAS DE DADOS ABERTOS	31
2.2.1.	Comprehensive Knowledge Archive Network (CKAN)	31
2.2.2.	Dataverse Project	35
2.2.3.	DSpace	37
2.2.4.	Comparativo entre plataformas de dados abertos	39
2.3.	WEB SEMÂNTICA E SUAS TECNOLOGIAS.....	41
2.4.	PRINCÍPIOS DE DADOS FAIR	49
2.4.1.	Processo de FAIRificação	50
2.4.2.	FAIR Data Point	51
3.	INTERLIGAÇÃO DE DADOS ABERTOS	54
3.1.	INTERLIGAÇÃO SEM CONSIDERAR METADADOS	56
3.2.	INTERLIGAÇÃO CONSIDERANDO OS METADADOS	60
3.3.	SUMARIZAÇÃO DAS ESTRATÉGIAS DE INTERLIGAÇÃO DE DADOS	64
4.	ABORDAGEM LigADOS PARA INTERLIGAÇÃO DE CONJUNTOS DE DADOS.....	67
4.1.	PROCESSO DE INTERLIGAÇÃO DE CONJUNTOS DE DADOS EM PLATAFORMAS DE DADOS ABERTOS	68
4.1.1.	Extração de metadados de um portal	69
4.1.2.	Geração das interligações semânticas entre conjuntos de dados de um portal	70
4.1.3.	Geração das etiquetas semânticas	78

4.1.4. Publicação das interligações semânticas entre conjuntos de dados	79
4.1.5. Geração de metadados semânticos como suporte a interligação de portais de dados.....	83
4.2. PROCESSO DE SUPORTE ONTOLÓGICO PARA ANOTAÇÕES SEMÂNTICAS DE CONJUNTOS DE DADOS.....	92
4.2.1. Geração da biblioteca de recursos semânticos	93
4.2.2. Publicação da biblioteca de recursos semânticos	94
4.3. ARQUITETURA DE SUPORTE.....	94
4.3.1. Componentes da Arquitetura	94
4.3.2. Implementação da Arquitetura	97
4.4. CONSIDERAÇÕES FINAIS.....	98
5. APLICAÇÃO DE LigADOS.....	101
5.1. CONFIGURAÇÃO DO AMBIENTE.....	101
5.2. PORTAL BRASILEIRO DE DADOS ABERTOS	103
5.3. PORTAL DE DADOS ABERTOS SOBRE AGROTÓXICOS.....	111
5.4. DISCUSSÃO	116
6. CONCLUSÃO	119
6.1. CONTRIBUIÇÕES.....	119
6.2. LIMITAÇÕES.....	121
6.3. TRABALHOS FUTUROS.....	121
REFERÊNCIAS.....	123
APÊNDICES	129
APÊNDICE A – COMPARATIVO DE <i>PROFILES</i> : FAIR DATA POINT (FDP) E LigADOS	129
APÊNDICE B – RECURSOS COMPUTACIONAIS DA APLICAÇÃO LigADOS	133
APÊNDICE C – TRIPLIFICAÇÃO E ANOTAÇÃO SEMÂNTICA DE CONJUNTOS DE DADOS DA APLICAÇÃO DE LigADOS.....	135
APÊNDICE D – EXTRATO DO GRAFO DE INTERLIGAÇÕES SEMÂNTICAS PARA O EXTRATO PORTAL DE DADOS ABERTOS SOBRE AGROTÓXICOS	137

1. INTRODUÇÃO

O acesso à informação tem sido tema cada dia mais recorrente em diferentes contextos e localidades. Desde o século passado, diversos países já estabeleceram leis de acesso à informação de nível nacional. No Brasil, a Lei de Acesso à Informação (LAI) foi sancionada em 2011¹, especificando um direito já previsto desde a Constituição de 1988. Embora a lei abrace o termo "informação", a parte central para edificação deste conceito inicia-se nos dados utilizados e produzidos em diferentes situações.

Em tecnologia da informação, dados são representações de objetos e eventos, que têm significado e importância para um grupo de usuários, e são armazenados no formato digital (HENDERSON; EARLEY, 2017). Dados abertos são dados que qualquer pessoa pode acessar, usar, reutilizar e compartilhar (OPEN KNOWLEDGE FOUNDATION, 2020). Dados governamentais são quaisquer dados e informações produzidos ou encomendados por órgãos públicos (UBALDI, 2013). Por conseguinte, dados abertos governamentais são dados produzidos por processos internos de órgãos públicos que são disponibilizados por esses órgãos para uso, reuso e compartilhamento de quaisquer interessados e sobre os quais não há restrições de privacidade aplicadas.

Plataformas de dados abertos funcionam como um catálogo de dados, que fornece uma interface entre quem disponibiliza os dados (publicador) e aqueles interessados em utilizá-los (consumidores) (OLIVEIRA et al., 2016). Na literatura com foco em dados de pesquisa científica, as plataformas para disponibilização de dados abertos de pesquisa, com funções de depósito de dados, têm sido referenciadas como "repositórios digitais de dados", "repositórios digitais de dados de pesquisa" ou, simplesmente, "repositórios de dados" (SAYÃO; SALES, 2016). As expressões "plataforma de dados abertos", "catálogo de dados abertos" e "repositório de dados abertos" são frequentemente usadas como sinônimos.

Uma plataforma de dados abertos é um *software* que implementa os principais recursos para gerenciar os dados abertos. Esses recursos incluem, mas não estão limitados a, gerenciamento de usuários, publicação de dados, gerenciamento de metadados, armazenamento de conjunto de dados, controle de acesso, pesquisa e visualização (BERENDS et al., 2020). O vocabulário DCAT define catálogo como uma coleção de metadados sobre conjuntos de dados e serviços de dados (DCAT, 2020). Um repositório de dados é uma das partes de uma infraestrutura de informação sustentável que fornece acesso e armazenamento em longo prazo (SAYÃO; SALES, 2016).

¹ http://www.presidencia.gov.br/ccivil_03/_Ato2011-2014/2011/Lei/L12527.htm

Nesta dissertação será utilizada a expressão **plataforma de dados abertos** para se referir às soluções tecnológicas disponíveis no mercado, na forma de *software*, para gerenciamento e publicação de dados abertos, utilizadas para construção de **portais de dados abertos** de uma determinada entidade, organização ou instituição, isto é, uma instância de portal.

Sendo os portais de dados abertos o produto do uso do *software* de plataforma de dados abertos, nesta dissertação será utilizada a expressão **instância de(o) portal**, ou a palavra **instância** seguida do nome ou sigla de uma solução tecnológica, para se referir a uma instalação específica da solução tecnológica de forma a atender às demandas de uma determinada entidade, organização ou instituição.

É importante ressaltar que, nesta dissertação, será utilizada a expressão **instância(s) de dados** para se referir aos dados disponibilizados, isto é, as representações de objetos e eventos, que têm significado e importância para um grupo de usuários, e são armazenados no formato digital (HENDERSON; EARLEY, 2017).

Nos últimos anos houve um rápido crescimento do número de portais disponibilizados na Web, construídos sobre plataformas de dados abertos. Na época de escrita desta dissertação, o registro do catálogo de dados datacatalogs.org, por exemplo, lista 589 catálogos de dados em todo o mundo. Exemplos para a crescente popularidade dos catálogos de dados são portais de dados abertos governamentais, de níveis nacionais e locais, portais de organizações internacionais governamentais ou não, bem como portais de dados científicos. Portais como o Portal Brasileiro de Dados Abertos², o Portal de Dados Abertos Sobre Agrotóxicos³ e o Metabuscador de Dados de Pesquisa da FAPESP⁴ são exemplos de pontos centrais para acesso da sociedade a conjuntos de dados sem restrições de uso. Muitos dos portais são construídos utilizando *softwares* de plataformas de dados abertos comerciais, como o Socrata⁵, ou de código-aberto, como o CKAN⁶ e o Dataverse⁷, para catalogar e publicar dados. De um modo geral, as plataformas de dados abertos fornecem suporte à descrição, armazenamento e descoberta de conjuntos de dados na Web. No que diz respeito ao armazenamento, algumas plataformas permitem que os conjuntos de dados sejam carregados diretamente na infraestrutura pertencente à instância do portal ou a criação de um apontamento (*hyperlink*) para obtenção do recurso no *website* do fornecedor de dados,

² <http://www.dados.gov.br/>

³ <https://dados.contraosagrototoxicos.org/>

⁴ <https://metabuscador.uspdigital.usp.br/>

⁵ <https://dev.socrata.com/>

⁶ <https://ckan.org/>

⁷ <https://dataverse.org/>

sendo esta última opção o uso mais comum. Nesta dissertação será utilizado o verbo **publicar**, ao invés de catalogar, de forma a abranger ambas as formas de disponibilização de conjuntos de dados nas plataformas.

Nos portais construídos sobre plataformas de dados abertos não há restrições de formatos para publicação dos conjuntos de dados. Contudo, um dos princípios de abertura de dados governamentais — políticas para dar transparência a quaisquer dados criados, coletados e mantidos por órgãos e entidades da administração pública — por exemplo, concita ao uso de dados em formatos não proprietários, isto é, “formatos sobre os quais nenhuma entidade possua controle exclusivo” (UBALDI, 2013; ATTARD et al., 2015). No Portal Brasileiro de Dados Abertos, por exemplo, formatos não proprietários costumam ser utilizados, tais como *Comma-Separated Values* (CSV) e *JavaScript Object Notation* (JSON). Contudo, ainda é possível encontrar conjuntos de dados disponibilizados em formatos proprietários tais como *eXcel Spreadsheet* (XLS). No Metabusador de Dados de Pesquisa da FAPESP, por exemplo, dados de pesquisa têm sido disponibilizados em formatos não proprietários, pelos próprios pesquisadores que geraram e/ou extraíram dados de/ou para suas pesquisas. Porém, também é possível encontrar alguns dados gerados por *software* específicos que ainda precisam ser disponibilizados em formatos proprietários, como dados gerados por *software* de geração de *workflows*.

Há alguns formatos não proprietários que seguem uma especificação padrão do *World Wide Web Consortium* (W3C) para representação e interligação de dados na Web, chamada de RDF (*Resource Description Framework*) (BIZER; HEATH; BERNERS-LEE, 2011). RDF é a base para publicação de *linked data*, ou dados conectados, na tradução do W3C Brasil⁸. Nesta dissertação será utilizada a expressão traduzida **dados conectados**.

A representação RDF permite que os dados sejam associados a elementos e/ou termos cujos significados são definidos em vocabulários, ontologias ou tesauros, isto é, recursos cujas terminologias têm significados comumente compreendidos. A associação de dados a vocabulários, ontologias ou tesauros é uma das formas de anexar dado a outro dado (OREN et al., 2006), isto é, uma forma de se realizar anotação semântica. Anotações semânticas inibem possíveis ambiguidades entre conceituações. Sendo assim, conjuntos de dados modelados em RDF que utilizem os mesmos elementos e/ou termos de vocabulários, ontologias ou tesauros para anotação semântica têm, minimamente, uma relação estabelecida: tratam de assuntos correlatos. Essa correlação pode, inclusive, ser estabelecida ao se identificar algum tipo de classe e/ou conceito mais geral que classifique o elemento e/ou termo efetivamente utilizado para anotar uma tripla.

⁸ <https://www.w3c.br/Padroes/WebSemantica>

A correlação de dados de diferentes fontes fornece uma integração parcial centrada na entidade (CORDEIRO et al., 2011). Em função disso, no contexto desta dissertação, as correlações entre dois conjuntos de dados abertos, modelados em RDF, estabelecidas por meio do uso de classes e/ou conceitos de vocabulários, ontologias e tesauros, que categorizam os elementos e/ou termos que tenham sido utilizados para anotações semânticas de recursos pertencentes aos conjuntos de dados serão chamadas de **interligações semânticas indiretas**. “Semânticas” porque dependem das classes e/ou conceitos definidos em vocabulários, ontologias ou tesauros. “Indiretas” porque as classes e/ou conceitos atuam como pontos de mediação das interligações. Ademais, nesta dissertação, quaisquer metadados associados a elementos e/ou termos de vocabulários, ontologias e tesauros serão chamados de **metadados semânticos**. Também é importante mencionar que as correlações entre dois conjuntos de dados abertos, modelados em RDF, estabelecidas no nível de instâncias pela associação de dados anotados semanticamente, significando que tratam do mesmo recurso, serão chamadas de **interligações semânticas diretas**.

Apesar das possibilidades de interligações proporcionadas pela representação de dados em RDF, a publicação de conjuntos de dados em RDF em portais de dados abertos é somente mais uma opção dentre as tecnologias nos quais os dados podem ser disponibilizados. Assim, geralmente está disponível a opção de *download* de conjuntos de dados em RDF, assim como para outras formas de modelagem de conjuntos de dados (estruturados ou não). Em função disso, nem o consumidor de dados humano, nem aplicações executadas por máquina têm acesso às interligações existentes entre os dados publicados a partir da interface do próprio portal. A falta de acesso às interligações de dados ocorre para dados modelados em quaisquer tecnologias, incluindo aqueles modelados em RDF. Sendo assim, a forma atual de publicação de conjuntos de dados em RDF, em portais de dados abertos, não permite que usuários do portal busquem, consultem e naveguem pelas interligações entre esses conjuntos de dados.

Nas plataformas de dados abertos, os metadados usados para descrever conjuntos de dados compreendem normalmente título, descrição, última atualização, dados do publicador, formato, licença e palavras ou conjuntos de palavras usadas como rótulos – as chamadas etiquetas (*tags*). Segundo Miller (2011), esses metadados são chamados de descritivos. Como o nome sugere, metadados descritivos são compostos por elementos para catalogar o recurso digital, dentre os quais somente as etiquetas objetivam descrever o conteúdo dos dados em si. No entanto, algumas palavras e/ou expressões utilizadas como etiquetas podem ser ambíguas. Uma etiqueta que se refira à palavra “capital” em português, por exemplo, segundo Ferreira (2004) pode significar “a cidade que aloja a alta administração

de um país ou estado, província, etc”; “bens ou valores acumulados”; e “dinheiro aplicado numa empresa por seus proprietários.” As múltiplas definições representam o tipo de ambiguidade semântica que aparece quando nem as instâncias, nem as etiquetas estão associadas a vocabulários que as definam. O uso de vocabulários, ontologias e tesauros pode gerar insumos para tornar estes metadados mais expressivos para os consumidores de dados, assim como processáveis por máquina.

De qualquer forma, os metadados descritivos podem criar um tipo de relação entre os conjuntos de dados publicados em um mesmo portal. Quando conjuntos de dados são descritos por itens similares, tais como, o mesmo publicador, as mesmas etiquetas, ou, ainda, títulos com uma ou mais expressões em comum, os motores de busca das plataformas de dados abertos retornam os conjuntos de dados no mesmo resultado, sugerindo alguma relação.

Além disso, em algumas plataformas de dados abertos é possível a criação de grupos temáticos de conjuntos de dados. Eles representam mais um tipo de metadado. No entanto, nenhum tipo de regra de negócio ou de regra semântica controla a inclusão de conjuntos nos grupos. Mesmo assim, os grupos temáticos adicionam um tipo de relação entre os conjuntos de dados incluídos nos mesmos grupos. Contudo, da mesma forma que os metadados descritivos, os grupos temáticos não consideram o significado das instâncias de dados. Assim, nesta dissertação, as relações geradas a partir de elementos de metadados coincidentes, que descrevem conjuntos de dados e que não estejam associados a vocabulários, ontologias e tesauros, serão chamadas de **relações não semânticas**. Assim sendo, as relações não semânticas são formadas por **metadados não semânticos**. Inversamente, cabe ressaltar que as relações entre conjuntos de dados geradas a partir de metadados semânticos serão consideradas **relações semânticas**. As relações semânticas são, de fato, interligações semânticas indiretas, conforme previamente definido para esta dissertação.

Conforme já mencionado, as características nativas de dados modelados em RDF possibilitam o estabelecimento de interligações de dados. De um modo geral, o uso de RDF pode ser explorado quando utilizado em conjuntos de dados abertos: na publicação e no consumo.

A publicação e o consumo de dados representados em RDF requerem atenção à semântica embutida nos dados, aos metadados disponibilizados aos consumidores de dados do portal e aos recursos de apoio disponíveis para busca, consulta e navegação tanto por humanos, quanto por máquinas. Em tese, a forma como os dados são disponibilizados em portais de dados abertos é beneficiada pela utilização de boas práticas para publicar dados na Web, que contribuem para melhorar a comunicação entre os publicadores e consumidores

de dados. Os benefícios alcançados pela aplicação das boas práticas incluem melhor compreensibilidade; facilidade de processamento; facilidade de descoberta; reuso; confiança; capacidade de interligação de dados; facilidade de acesso; e interoperabilidade (LÓSCIO; GUIMARÃES; CALEGARI, 2016). Um conjunto de boas práticas para dados na Web pode ser consultado em Lóscio et al. (2017).

Ainda no contexto da publicação e consumo de dados abertos, dados em RDF viabilizam a interligação segura entre dados, gerando publicações no mais alto nível do esquema *Five Star of Open Data* (Cinco Estrelas dos Dados Abertos), aumentando o valor dos dados e sua capacidade de descoberta (OPEN KNOWLEDGE FOUNDATION, 2020).

Complementarmente, a publicação e o consumo de dados modelados em RDF também podem ser beneficiados pela implementação de diretrizes para gestão de dados e metadados. Neste contexto, FAIR é uma iniciativa, inicialmente, focada no contexto de dados científicos, mas que vem sendo adequada a outros tipos de fontes de dados. FAIR é o acrônimo, originalmente em língua inglesa, para *Findable, Accessible, Interoperable e Reusable*. A iniciativa FAIR define uma série de princípios orientadores. O objetivo é que recursos digitais, conjunto de dados, códigos e objetos gerados ou utilizados por pesquisas científicas sejam descobertos, estejam acessíveis, aptos para interoperabilidade global, e reutilizáveis, tanto por humanos, quanto por máquinas (WILKINSON et al., 2016). Por isso, é natural que a maioria dos princípios FAIR esteja relacionada também a metadados ricamente descritos, a fim de aumentar a possibilidade de descoberta, que é a raiz da acessibilidade, da interoperabilidade e da reutilização. Dentre os princípios FAIR, o enriquecimento de metadados, com uso de vocabulários, ontologias e tesouros para descrevê-los, bem como a interligação entre metadados são aspectos abordados. Os metadados detalhados são essenciais para as tarefas de gerenciamento de dados e dão suporte às tarefas de interligação de dados (para interoperabilidade) e de pesquisa (para descoberta). Além disso, o uso de formatos de dados não proprietários elimina restrições sobre quem os utiliza, o que também melhora a interoperabilidade (FIGUEIREDO et al., 2019).

Por fim, de forma a guiar o leitor, o Quadro 1 reúne algumas das expressões que serão utilizadas ao longo desta dissertação.

Quadro 1: Expressões utilizadas no texto desta dissertação

Plataformas de Dados Abertos	Soluções tecnológicas disponíveis no mercado, na forma de <i>software</i> , para gerenciamento e publicação de dados abertos e utilizadas para construção de portais de dados abertos de uma determinada entidade, organização ou instituição.
Portal de dados abertos	Uma instância de portal.
Instância de(o) portal	Uma instalação específica de uma solução tecnológica de plataforma de dados abertos para atender às demandas de uma determinada entidade, organização ou instituição.
Instância(s) de dados	Os dados disponibilizados, isto é, as representações de objetos e eventos, que têm significado e importância para um grupo de usuários, e são armazenados no formato digital (HENDERSON; EARLEY, 2017).
Publicar ou Publicação	Ato ou efeito de disponibilizar conjuntos de dados em portais de dados abertos construídos sobre plataformas de dados abertos.
Dados conectados	A expressão <i>linked data</i> traduzida pela W3C Brasil.
Interligação Semântica Indireta	Correlação entre dois conjuntos de dados abertos, modelados em RDF, estabelecidas por meio do uso de classes e/ou conceitos de vocabulários, ontologias e tesauros, que categorizam os elementos e/ou termos que tenham sido utilizados para anotações semânticas de recursos pertencentes aos conjuntos de dados.
Interligação Semântica Direta	Correlação entre dois conjuntos de dados abertos, modelados em RDF, estabelecida no nível de instâncias pela associação de dados anotados semanticamente, significando que tratam do mesmo recurso.
Metadados semânticos	Quaisquer metadados associados a elementos e/ou termos de vocabulários, ontologias e tesauros.
Metadados não semânticos	Metadados dos quais são geradas as relações não semânticas.
Relação não semântica	Relação entre conjunto de dados gerada a partir de elementos de metadados coincidentes, que descrevem os conjuntos de dados, e que não estejam associados a vocabulários, ontologias e tesauros.
Relação semântica	Relação entre conjuntos de dados geradas a partir de elementos metadados semânticos, que descrevem os conjuntos de dados. Nesta dissertação, significa o mesmo que interligação semântica indireta.

1.1. Problema

Em sua concepção inicial, plataformas de dados abertos não têm como foco ou não priorizam interligação e interoperabilidade dos dados publicados. Em função disso, nos

portais de dados abertos construídos sobre plataformas de dados abertos, os conjuntos de dados são disponibilizados de forma isolada. Isso significa que, mesmo que haja dados ou metadados relacionados dentre os conjuntos de dados publicados, mecanismos de referenciamento mútuo não são normalmente utilizados. A situação de isolamento ocorre para conjuntos de dados representados em quaisquer formatos, incluindo os conjuntos de dados representados em RDF.

Considerando especialmente dados modelados em RDF, as interligações que carregam os fundamentos tecnológicos do *framework* na maior parte das vezes não são exploradas em todo o seu potencial. Após a obtenção do conjunto de dados via *download*, o consumidor de dados precisa ser um especialista em algumas tecnologias de Web Semântica para obter proveito do conteúdo e compreender as interligações de dados em RDF. Ademais, a capacidade de processamento por máquina dos conjuntos de dados publicados em RDF é subutilizada, pois usualmente não há suporte, integrado ou não, para acesso às estruturas de organização de dados utilizadas. Para ser processável por máquina, o conjunto de dados precisa ser obtido em sua fonte e extraído do ambiente da plataforma.

Por outro lado, o responsável pela publicação dos dados pode também não ser um especialista em tecnologias de Web Semântica. Por isso, ele pode disponibilizar um conjunto de dados em RDF sem maiores interligações a outros recursos. Isto ocorre pelo fato dele não ter à sua disposição alguma interface simples que o apoie no processo de anotação da semântica de seus dados, utilizando vocabulários, ontologias ou tesouros e fornecendo suporte à ligação de seus dados a dados relacionados de outras proveniências ou até mesmo de sua própria organização.

Ainda no que diz respeito à forma como os conjuntos de dados são publicados, observa-se que os metadados registrados nos portais de dados referem-se ao conjunto de dados como um todo, incluindo muitas vezes níveis de abstração mais altos, tal como a área temática. Como nenhum esquema de organização e classificação de metadados é utilizado, mecanismos de validação deles não podem ser estabelecidos.

A forma como os dados são publicados influencia diretamente na maneira como os consumidores de dados buscam dados. Muitas das vezes, as relações existentes entre dados são mais importantes que os dados tratados isoladamente. Em função disso, é comum que consumidores de dados (os interessados) busquem dados em mais de um local para obter a completude de uma informação.

Face ao exposto, configura-se como problema o fato de que **o consumo de conjuntos de dados modelados em RDF, em plataformas de dados abertos, não considera o potencial oferecido pela tecnologia de dados conectados. Nessas plataformas, a falta de suporte para consumo de dados conectados prejudica, principalmente, a exposição das interligações**

semânticas com outros conjuntos de dados modelados em RDF. Esse fato, diminui as possibilidades de descoberta e interoperabilidade de dados, conforme proposto pela iniciativa FAIR e distancia-se de boas práticas para publicação de dados na Web que geram melhorias na compreensibilidade dos dados. A não exploração dessas interligações subutiliza os benefícios do uso de RDF, assim como subutiliza sua capacidade de ser consumido por máquina.

1.2. Objetivos

O objetivo geral desta pesquisa é desenvolver uma abordagem que permita explicitar interligações semânticas indiretas entre conjuntos de dados modelados em RDF, publicados em instâncias de portais de dados abertos.

Sendo assim, partindo da premissa de que os dados estejam publicados modelados em RDF e anotados fazendo uso de vocabulários senso comum, o foco desta pesquisa é utilizar os benefícios da abordagem de dados conectados, de forma a identificar interligações semânticas entre conjuntos de dados que possam não estar explícitas, por estarem interligados de forma indireta (interligações semânticas indiretas) e, conseqüentemente, fomentar o processamento por máquina.

Com base no objetivo geral é possível identificar dois grupos de objetivos específicos: o primeiro centrado na interligação de conjuntos dados; o segundo centrado no suporte à anotação semântica de conjuntos de dados a serem publicados em portais de dados. Esses grupos se coadunam para elevar o entendimento dos dados publicados nos portais de dados abertos, por parte dos envolvidos no consumo e na publicação.

Assim sendo, os objetivos específicos (OE), ou intermediários, foram divididos e agrupados da seguinte forma para atender ao objetivo geral:

- OE-Interligação-de-dados
- OE-Suporte-semântico

O OE-Interligação-de-dados considera proporcionar aos consumidores de dados de portais de dados abertos, ampliação da compreensão das interligações dos conjuntos de dados em RDF publicados nos portais, sem que para isto, necessite explorar cada conjunto de dados individualmente. São três os OE nesse sentido:

- OE-Interligação-de-dados-1: Estabelecer interligações semânticas indiretas entre os conjuntos de dados em RDF publicados em um portal de dados abertos, a partir das ligações das classes/conceitos dos vocabulários utilizados (interligações semânticas indiretas);

- OE-Interligação-de-dados-2: Gerar, associar e expor metadados semânticos para os conjuntos de dados em RDF publicados em um portal de dados abertos;
- OE-Interligação-de-dados-3: Empregar soluções para navegação e exploração das interligações semânticas entre os conjuntos de dados em RDF publicados em um portal de dados abertos; e
- OE-Interligação-de-dados-4: Estabelecer interligações semânticas indiretas entre portais de dados abertos para os quais as interligações semânticas entre seus conjuntos de dados já tenham sido estabelecidas.

O OE-Suporte-semântico considera proporcionar aos responsáveis pelas criações e publicações de novos conjuntos de dados em RDF, uma forma pela qual possam acessar as soluções ontológicas que já foram utilizadas por outrem. Nesse sentido, um OE faz-se necessário:

- OE-Suporte-semântico-1: Proporcionar uma interface de exposição de soluções ontológicas, alinhadas aos conceitos da Web Semântica, que tenham sido utilizadas por conjuntos de dados, publicados em RDF, em um portal de dados abertos.

1.3. Abordagem de Solução

A forma como o problema da pesquisa foi trabalhado está alinhada à divisão dos OE descritos. Por isso, a abordagem foi dividida em dois processos:

- Interligação de conjuntos de dados em plataformas de dados abertos;
- Suporte ontológico para anotações semânticas de conjuntos de dados.

Para o processo “Interligação de conjuntos de dados em plataformas de dados abertos”, foram definidos cinco sub-processos:

- Extração de metadados de um portal;
- Geração das interligações semânticas entre conjuntos de dados de um portal;
- Geração das etiquetas semânticas;
- Publicação das interligações semânticas entre conjuntos de dados; e
- Geração de metadados semânticos como suporte a interligação de portais de dados.

Para o processo “Suporte ontológico para anotações semânticas de conjuntos de dados”, foram definidos dois sub-processos:

- Geração da biblioteca de recursos semânticos utilizados; e
- Publicação da biblioteca de recursos semânticos utilizados pelos conjuntos de dados.

LigADOS é a sigla para **InterLIgação de conjuntos de dADOS em plataformas de dados abertos na Web Semântica**, a abordagem de solução aqui proposta. O Capítulo 4 apresenta o detalhamento da abordagem.

1.4. Justificativa

Atualmente, portais de dados abertos já disponibilizam conjuntos de dados modelados em RDF, uma tecnologia habilitadora da Web Semântica. Contudo, as plataformas de dados abertos ainda não oferecem funcionalidades de busca, consulta e navegação que explorem as características de dados modelados em RDF, principalmente, a exploração das interligações de dados. É possível encontrar conjuntos de dados modelados em RDF, contendo triplas com pouco ou nenhum uso de vocabulários, ontologias ou tesouros, optando por terminologias proprietárias ou descrições textuais de recursos, o que dificulta a exploração de dados.

Destarte, a abordagem de solução proposta entende que há diversas interligações entre os conjuntos de dados publicados em portais de dados abertos que não estão explícitas aos consumidores de dados dos portais (usuários). Ademais, as diversas interligações não estão adequadamente acessíveis para processamento por máquina. Outrossim, as interligações devem estar acessíveis para que o público em geral possa usufruir dos benefícios da Web de Dados, democratizando o acesso aos diferentes perfis de consumidores.

Em função disso, a abordagem proposta considera conjuntos de dados já publicados, elevando a compreensão de seus dados, ao fornecer uma solução para estabelecimento e exploração das interligações semânticas existentes. A elevação do entendimento dos conjuntos de dados é também apoiada pela geração de metadados semânticos.

Adicionalmente, a abordagem proposta considera que a forma de consumo depende diretamente de como conjuntos de dados foram publicados. Por isso, a abordagem também fornece suporte ao publicador de dados interessado em disponibilizar conjuntos de dados anotados com vocabulários, ontologias e tesouros já utilizados por outrem. Para isso, a abordagem proposta expõe conceitos gerais de um domínio, que foram utilizados pelos

conjuntos de dados publicados em uma instância de portal. A exposição de conceitos permite que um publicador de dados verifique o que já foi utilizado para anotação semântica de outros conjuntos de dados disponíveis no portal. Assim, ele pode utilizar esta informação como sugestão para possíveis anotações de futuros conjuntos de dados.

Para portais temáticos, como o Portal de Dados Abertos sobre Agrotóxicos, é bastante comum que conjuntos de dados, mesmo que publicados por diferentes organizações, tratem de assuntos correlatos. Neste sentido, obter informação de quais vocabulários, ontologias e tesouros estão sendo usados, funciona como uma espécie de mecanismo de sugestão de conceituações de dados utilizadas em um determinado domínio e que podem fomentar e sugerir possíveis interligações.

Para portais de dados não focados num só tema, como o Portal Brasileiro de Dados Abertos, dada a diversidade de temas, a exposição de conceitos de um domínio auxiliam nas buscas de dados. Diante da variedade de temas, as expressões de buscas precisam ser mais elaboradas, principalmente quando se considera, os múltiplos significados que os termos podem ter em uma língua. Neste contexto, o suporte aqui proposto pode indicar as soluções ontológicas utilizadas, agrupando por temas. A utilização de termos e/ou elementos já utilizados para anotar novos dados permite a geração automática de interligações indiretas entre os conjuntos de dados de um portal.

A proposta de reutilização de termos e/ou elementos alinha-se aos princípios FAIR que tratam de interoperabilidade (*Interoperability*), pois se as mesmas soluções ontológicas são utilizadas, mais simples é o processo de interoperar; e de reusabilidade (*Reusability*), pois as sugestões são baseadas em soluções ontológicas efetivamente já utilizadas. Se conjuntos de dados estão mais interligados, em um ou vários portais, a capacidade de descoberta (*Findability*) deles pela navegação nas interligações aumenta. Soma-se a isso, o fato da navegação ser atividade intuitiva até mesmo para usuários mais leigos.

1.5. Metodologia de Pesquisa

Para atingir os objetivos descritos, esta pesquisa seguiu as seguintes fases: (i) revisão da literatura, objetivando identificar as características da Web de Dados e como ela tem sido tratada pelas plataformas de dados abertos de código aberto; (ii) exploração das abordagens, publicadas na literatura, utilizadas para interligação de dados catalogados em plataformas de dados abertos; (iii) exploração das abordagens que envolvem enriquecimento de metadados; (iv) especificação da abordagem de solução proposta por meio da modelagem de seus fluxos de processos; (v) projeto e desenvolvimento de uma estrutura para dar suporte à aplicação da abordagem proposta; e (iv) definição e operacionalização da avaliação da abordagem.

Para revisão da literatura foi executada uma revisão sistemática. Três expressões de buscas foram utilizadas: “*dataset interlinking semantic RDF*”, “*semantic metadata interlinking RDF*” e “*‘data catalog’ semantic RDF*”. As três expressões foram submetidas ao: Portal de Periódicos da CAPES, IEEE Xplorer, Springer e Journal of Web Semantics da Science Direct. Em todas as buscas foi aplicado o filtro para publicações entre 2013 a 2020, além de terem sido aplicadas algumas variações na expressão de busca, tais como aplicação de conectores “AND” e “OR” entre as palavras da expressão. Com as referências selecionadas passou-se para a segunda fase da metodologia, onde a ação de exploração das abordagens permite caracterizar este trabalho como uma pesquisa exploratória, já que pretende aprofundar o entendimento das lacunas nas abordagens existentes (ALVES, 2003, p.52).

Cabe ressaltar que esta pesquisa faz parte de um conjunto de investigações na área de apoio semântico à interoperabilidade de dados que vêm sendo conduzidas pelo Grupo de Pesquisa em Engenharia do Conhecimento (GRECO⁹) do Programa de Pós-Graduação em Informática (PPGI) da Universidade Federal do Rio de Janeiro (UFRJ).

1.6. Organização do trabalho

O Capítulo 2 inicia com uma revisão sobre conceitos dos tipos de metadados, focando naqueles que consideram algum nível semântico no contexto de descoberta de dados. Já que o contexto de trabalho inclui plataformas de dados abertos, suas características são analisadas. Esta análise gerou subsídios para uma comparação entre três plataformas de código aberto existentes na época da escrita desta dissertação. Na sequência, os conceitos e tecnologias habilitadoras da Web Semântica, incluindo reificação com RDF* são revisados. Como a pesquisa trata de dados abertos, não se pode furtar a comentar e considerar a iniciativa FAIR, focando no processo de FAIRificação e na estrutura de um FAIR *Data Point*.

O Capítulo 3 apresenta algumas abordagens para interligação de dados abertos. Assim, no referido capítulo são apresentadas abordagens de interligação de dados que propõem apoiar a descoberta de interligações para enriquecer novos conjuntos de dados em RDF a serem publicados, isto é, abordagens que não consideram os dados que desempenham o papel de metadados. Também são apresentadas abordagens que propõem a interligação de dados considerando os metadados associados aos conjuntos de dados.

O Capítulo 4 detalha o coração da abordagem proposta – LigADOS – discorrendo sobre a estratégia de interligação, seus processos, sua arquitetura, além de geração de metadados semânticos e de suporte a estruturas que orquestram metadados de diversos

⁹ <https://sites.google.com/view/greco-ppgi-ufrrj>

portais. O capítulo finaliza com algumas considerações, conduzindo o leitor por aspectos sumários ao mesmo tempo que amplia sua visão de análise da abordagem.

O Capítulo 5 apresenta a aplicação de LigADOS em instâncias de portais distintas. Demonstra-se suas facetas de utilização e o surgimento de novas perspectivas de análise de dados, como consequência de seu uso. Por fim, os resultados são discutidos.

No Capítulo 6 discute as conclusões do trabalho, suas contribuições, limitações e trabalhos futuros.

2. WEB SEMÂNTICA, PLATAFORMAS DE DADOS ABERTOS E METADADOS

Muitas tecnologias emergiram desde o surgimento da *World Wide Web* (The Web) no final dos anos 80. A busca constante por melhores capacidades de conectar documentos, de mecanismos de busca e de recuperação da informação são somente alguns dos inúmeros aspectos motivacionais para o desenvolvimento dessas tecnologias. Um aspecto motivacional que mudou a forma de “ver” esta Web foi a ideia de conectar dados, serviços, programas, sensores, dispositivos pessoais e até mesmo eletrodomésticos, não só para uso humano, mas também para processamento por máquina. Assim, no final da década de 90 surgiu a Web Semântica (HENDLER, BERNERS-LEE, MILLER, 2002). Acompanhando este fato, caminhavam os dados em formato digital publicados na Web e a necessidade de mais dados para descrever estes dados, despondo novos desafios. Dentre estes desafios, a necessidade de entender e estabelecer as relações entre esses dados.

A seção 2.1 apresenta os tipos de metadados existentes. Em função da necessidade de publicar dados na Web, a seção 2.2 apresenta as plataformas para publicação de dados, comentando as principais características de três plataformas de código aberto. A seção 2.3 apresenta os conceitos básicos da Web Semântica e suas tecnologias habilitadoras, bem como conceitos da Web de Dados. Como na Web de Dados os dados precisam ter significado, isto é, definições semânticas, são abordados mecanismos para estas definições. E, por fim, na seção 2.4 a iniciativa FAIR é apresentada, destacando-se questões do processo de tornar dados FAIR e seu armazenamento.

2.1. Metadados

Diferentes tipos de metadados são necessários para descrever, descobrir, recuperar, usar, apresentar e preservar recursos digitais, tal como um conjunto de dados (XIE, MATUSIAK, 2016). A classificação dos tipos de metadados varia de autor para autor.

No contexto desta dissertação, a classificação dada por Miller (2011) adequa-se bem. Miller (2011) classifica metadados em três categorias: descritivos, administrativos e estruturais, seguindo uma abordagem comum utilizada pela Ciência da Informação. Metadados descritivos provêm elementos para catalogar o recurso digital, de forma a identificar o conteúdo, incluindo termos para recuperar informações. Metadados administrativos são elementos usados para gerenciar o recurso digital, tal como o ciclo de vida dos dados, incluindo informações de proveniência, sobre a criação e distribuição do recurso. Metadados estruturais fornecem elementos para especificação da estrutura dos

dados, tal como múltiplos arquivos para definição do recurso digital ou as múltiplas visões de um mesmo recurso.

Dentre os metadados descritivos, etiquetas comumente fazem parte dos metadados associados aos conjuntos de dados de um portal de dados abertos. De um modo geral, uma etiqueta é um elemento de metadados, definida por uma cadeia de caracteres que representa seu nome (TYGEL, 2016). Segundo Kalender (2010), etiquetas são subjetivas e inconsistentes, pois são palavras-chave não hierarquizadas e que falham na captura exata do significado e do contexto de uso, visto que as línguas possuem elementos ambíguos. Em Tygel (2016) são discutidos alguns problemas frequentes na geração de etiquetas para uso em portais de dados abertos. O responsável pela publicação de conjuntos de dados nos portais de dados abertos tem a liberdade de atribuição de etiquetas. Naturalmente, as etiquetas podem remeter à semântica do conteúdo, mas não necessariamente isso ocorre. Observa-se com frequência etiquetas que representam a organização que publicou o conjunto de dados ou referências aos nomes dos sistemas que as geraram, os quais não necessariamente representam o conteúdo dos dados.

Apesar dos problemas, é frequente o uso de etiquetas. Um conjunto de dados publicado em portal pode ter várias etiquetas associadas ou até mesmo nenhuma etiqueta. Uma etiqueta pode estar associada a mais de um conjunto de dados. Isso beneficia a descoberta de conjuntos de dados que tratem de assuntos correlatos.

Kalender (2010) menciona a expressão “palavras-chave”. Palavras-chave representam o conteúdo do documento ou de um conjunto de documentos em uma ou algumas palavras (CASTRO, 2019). Sendo assim, há um viés semântico ao considerar o conteúdo do documento.

A extração de palavras-chave do conteúdo do conjunto de dados para uso como etiquetas enriquecem a qualidade dessas etiquetas. Na extração, as palavras mais relevantes são analisadas para selecionar aquelas mais representativas, formando um resumo do conjunto de dados (CASTRO, 2019). As abordagens de associação automática de etiquetas fornecem precisão e consistência com o conteúdo dos dados (KALENDER, 2010).

De um modo geral, os metadados são livremente atribuídos pelos responsáveis pela publicação de dados. Alguns elementos podem ser gerados e associados por *software* de apoio à publicação de dados, tais como datas de criação e atualização. Além disso, elementos de metadados podem ser associados a elementos e/ou termos de vocabulários, ontologias e tesauros, formando **metadados semânticos**, sejam eles: (i) descritivos, que se refiram à temática dos conjuntos de dados e as etiquetas associadas; (ii) administrativos, que se refiram à proveniência dos dados; ou (iii) estruturais, que se refiram ao formato no qual os dados foram disponibilizados.

2.2. Plataformas de Dados Abertos

Diversas organizações governamentais e não governamentais – de níveis local, regional, nacional e internacional – têm tornado público dados produzidos por seus processos internos. Naturalmente, os dados a serem abertos não incluem aqueles sobre os quais algum requisito de privacidade faz-se necessário, tal como alguns dados pessoais ou estratégicos. Como exemplos de iniciativas de nível nacional, vários países lançaram seus próprios portais de dados abertos (em inglês: *Open Data Portal* - ODP) de forma que a sociedade possa participar e entender decisões governamentais.

Ainda no sentido de tornar dados públicos, os dados gerados ou utilizados por pesquisas científicas se tornaram foco de grande interesse para o mundo científico. Para que dados de pesquisa revelem e transmitam conhecimento no tempo, ações específicas precisam ser implementadas (SAYÃO; SALES, 2016). A falta de gerenciamento desses dados dificulta a validação e, às vezes, a continuidade da pesquisa a partir do ponto onde ela contribuiu. Com o intuito de apoiar o gerenciamento e o compartilhamento de dados, algumas plataformas de dados abertos estão em uso e continuam se desenvolvendo ao longo dos anos.

Nas subseções 2.2.1, 2.2.2 e 2.2.3 serão apresentadas as principais características de três plataformas de dados abertos e de código aberto. A subseção 2.2.4 apresenta um comparativo entre as plataformas considerando suas características.

2.2.1. Comprehensive Knowledge Archive Network (CKAN)

A maioria das iniciativas de abertura de dados gerados por processo governamentais tem sido disponibilizados em ODPs estruturados usando a plataforma *Comprehensive Knowledge Archive Network* (CKAN): uma ferramenta de código aberto e gratuita, com foco na criação de sites para disponibilização de dados abertos, desenvolvida e mantida pela *Open Knowledge Foundation*¹⁰ (COSTA et al., 2017; NEUMAIER et al., 2017, CORREA; ZANDER; SILVA, 2018).

A proposta do CKAN é ser uma plataforma de gerenciamento de dados com o intuito de torná-los acessíveis, localizáveis e apresentáveis, onde os dados são organizados e disponibilizados ao público em geral em unidades chamadas conjuntos de dados. Conjunto de dados é uma coleção de dados que foi catalogada numa instância de portal. Os conjuntos de dados podem atender às especificações de diferentes padrões de organização de dados,

¹⁰ <https://okfn.org/>

caracterizando os formatos de arquivos de dados. Significa que um conjunto de dados pode ser disponibilizado ao público em diferentes formatos.

Um aspecto a esclarecer está no fato de que, no CKAN, os conjuntos de dados também podem ser organizados em partes distintas para formar um todo em comum. Essas partes podem ser do mesmo formato ou de formatos diferentes. As partes que efetivamente contêm os dados do conjunto de dados são chamadas de distribuições (ou ainda, distribuições de dados ou distribuições de conjuntos de dados). As distribuições podem:

- referir-se aos mesmos dados, porém em diferentes formatos (Ex.: XLS, CSV, RDF);
- possuir as mesmas estruturas de dados, porém contendo dados de diferentes escopo de tempo e/ou espaço (Ex.: uma contendo dados referentes ano de 2019; outra dados referentes ao ano de 2020, etc.); e
- possuir diferentes níveis de detalhamento dos dados (Ex.: uma contém os totais por localizações geográficas; outra contém dados brutos, não totalizados, organizados de acordo com uma escala temporal).

O CKAN não impõe regras sobre como as distribuições de dados devem ser organizados. O único requisito do CKAN é que a publicação do conjunto de dados tenha pelo menos uma distribuição associada. A Figura 1 apresenta um esquema de como os dados são organizados em uma instância CKAN.

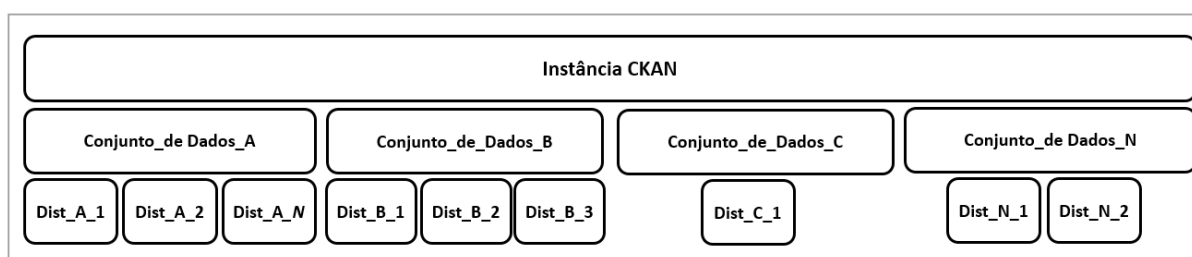


Figura 1: Esquema de organização dos conjuntos de dados e suas distribuições em uma instância CKAN

A Figura 2 apresenta um extrato da publicação de um conjunto de dados publicado em uma instância CKAN com oito distribuições associadas em diferentes formatos.

Conjunto de dados Grupos Fluxo de Atividades Relacionado

PARA - Programa de Análise de Resíduos de Agrotóxicos em Alimentos

Tabelas extraídas do relatório do PARA, com diversos dados relativos aos resíduos de agrotóxicos encontrados em alimentos.

Dados e recursos

	Dados - PARA 2008 Tabelas extraídas do relatório final do Programa de Análise de Resíduos de...	Explorar
	Dados - PARA 2010 Tabelas extraídas do relatório final do Programa de Análise de Resíduos de...	Explorar
	PARA 2008 Relatório final do Programa de Análise de Resíduos de Agrotóxicos em...	Explorar
	PARA 2010 Relatório final do Programa de Análise de Resíduos de Agrotóxicos em...	Explorar
	PARA 2011/2012 Relatório final do Programa de Análise de Resíduos de Agrotóxicos em...	Explorar
	PARA 2012 - 2a Etapa Relatório do Programa de Análise de Resíduos em Alimentos - PARA 2012, 2a...	Explorar
	Amostras com e sem resíduos, por ano A partir dos relatórios do PARA, a tabela mostra o número de amostras...	Explorar
	PARA 2013-2015	Explorar

Figura 2: Conjunto de dados com oito distribuições associadas: dois em XLS, cinco em PDF e um em CSV

No processo de publicação dos conjuntos de dados no CKAN, a organização responsável pela disponibilização dos dados pode ser registrada. O registro da organização pode ser obrigatório ou não, dependendo de como o administrador da instância configurou o ambiente. Cada conjunto de dados de uma organização recebe sua própria página com uma coleção de metadados. Para visualização de conjuntos de dados em formatos tabulares, imagens e até mesmo informações de localização geográficas é possível, ainda, visualizá-los usando gráficos, tabelas e mapas na própria interface do CKAN, se assim o publicador configurar (CKAN, 2020). Todavia, para aqueles modelados em RDF, nenhum suporte está disponível, sendo o *download* de dados a única opção.

O CKAN permite ainda que conjuntos de dados sejam agrupados de forma a estabelecer alguma relação temática entre eles, embora não haja obrigatoriedade de utilizar

nenhum aspecto físico ou semântico no agrupamento, conforme proposto nas abordagens de dados conectados. Os agrupamentos são feitos pelo responsável pela publicação.

Nativamente, o CKAN fornece uma série de opções de descritores (metadados) para cada conjunto de dados. Os metadados são definidos pelo publicador. Exemplos incluem: título, identificador único por meio de um URL, descrição, licença, etiquetas (*tags*) entre outros. Os metadados podem ser extraídos e explorados externamente ao CKAN, como dados conectados, por meio da instalação da extensão (não nativa) “*ckanext-dcat*”, detalhada no Apêndice B. A extensão “*ckanext-dcat*” triplicifica e anota os metadados. O intercâmbio de metadados com outras instâncias CKAN é apoiado pela instalação da extensão (não nativa) “*ckanext-harvest*”¹¹, que viabiliza o acesso a metadados associados a conjuntos de dados de outras instâncias CKAN.

Apesar dos conjuntos de dados em RDF possuírem metadados em grãos menores que o conjunto como um todo, isto é, metadados no nível de triplas, quando anotadas semanticamente, o CKAN não explora essa vantagem nos mecanismos disponíveis para atribuição de metadados, nem em seus mecanismos de acesso aos dados.

De fato, a maior parte dos elementos que compõem o conjunto de metadados disponíveis referem-se a características descritivas (título), administrativas (licenças de uso) e estruturais (identificador único), com pouca semântica. O elemento que objetiva descrever o conteúdo dos dados em si resume-se às etiquetas. Etiquetas, por sua vez, têm sido atribuídas pelos próprios publicadores de dados, não garantindo a melhor representatividade dos dados, visto que geralmente não há regras ou restrições na atribuição de etiquetas (TYGEL, 2016). Ademais, no CKAN as etiquetas são atribuídas para cada conjunto de dados, muitas vezes referindo-se a uma área temática, não necessariamente representando as instâncias dos conjuntos de dados. Diversas pesquisas têm sido desenvolvidas para apoiar e automatizar a geração de etiquetas (SOUZA, 2013; LIRA, 2014; TYGEL, 2016; CASTRO, 2019).

No CKAN, os consumidores de dados podem iniciar uma pesquisa submetendo uma ou mais expressões textuais, e o mecanismo de busca da plataforma utilizará todos os metadados associados aos conjuntos de dados. No CKAN também é possível executar consultas facetadas de modo a filtrar os resultados. Por exemplo, pode-se iniciar uma pesquisa utilizando a expressão “vegetação” e facetar a pesquisa pela seleção do tipo de formato de arquivo, pela organização que publicou, dentre outras combinações (CKAN, 2020).

Embora a informação sobre o tipo de licença de uso para cada conjunto de dados seja necessária, o CKAN não oferece nenhum tipo de controle de acesso para os recursos em si. O único controle de acesso existente é para o responsável pela publicação pertencente a uma

¹¹ <https://github.com/ckan/ckanext-harvest>

organização (USER GUIDE CKAN - 2.9.0, 2020). A ausência de controle de acesso pelos consumidores em geral justifica sua ampla utilização pelas organizações governamentais no intuito de tornar seus dados abertos alinhados às previsões legais de leis de acesso à informação. Ainda com relação ao controle de acesso, alguns dos conjuntos de dados devem ser periodicamente atualizados pelos seus responsáveis que precisam ser autenticados. Como alguns consumidores de dados podem ter o interesse de acompanhar continuamente a evolução de determinados dados, existe a possibilidade de solicitar o recebimento de uma notificação a cada atualização de dados. Naturalmente, esta opção requer que uma conta de acesso ao CKAN seja criada, mas essa conta não limita em nada buscas e acessos a todos os conjuntos de dados disponibilizados em um portal.

O CKAN é bastante flexível no tocante à personalização. Para sua interface Web, a CKAN Action-API¹² está disponível e bem documentada. Mais que isso, a flexibilidade atende a capacidade de se estender características e funcionalidades existentes ou de inserção de novas características, por meio da adoção de extensões já desenvolvidas por terceiros ou pela opção de desenvolvimento de uma nova extensão¹³. Essa capacidade corrobora para a escolha do CKAN para criação de portais para dados abertos governamentais. Cada instância do CKAN pode desenvolver, alterar ou utilizar as extensões que lhe forem necessárias, sem nenhum tipo de hierarquia ou vinculação com outras instâncias (portais).

Além do CKAN, Socrata¹⁴ e Opendatasoft¹⁵ também são plataformas para desenvolvimento de portais com foco na publicação de dados abertos governamentais. Conforme mencionado anteriormente, o CKAN tem sido bastante utilizado em portais de governos de países da Europa, América do Sul, América do Norte, África e Austrália. O Socrata tem sido utilizado principalmente por governos locais dos Estados Unidos da América; e o Opendatasoft tem maior popularidade entre organizações europeias. Todavia, Socrata e Opendatasoft não serão considerados na experimentação da abordagem de solução proposta nesta pesquisa por serem soluções pagas.

2.2.2. Dataverse Project

No que concerne aos dados de pesquisa, a maior parte deles têm sido disponibilizados em repositórios, muitas das vezes separados das publicações científicas que lhes deram origem ou que os utilizaram. Neste contexto, o *Dataverse Project*¹⁶ é um aplicativo de código

¹² <https://ckan.org/portfolio/api/>

¹³ <https://docs.ckan.org/en/2.8/extensions/>

¹⁴ <https://www.tylertech.com/products/socrata>

¹⁵ <https://www.opendatasoft.com/>

¹⁶ <https://dataverse.org/>

aberto, desenvolvido e mantido pelo *Institute for Quantitative Social Science* (IQSS) em parceria com a *Harvard University Library* e *Harvard University Information Technology*, que desde 2006 tem sido utilizado por várias universidades ao redor do mundo como uma plataforma de compartilhamento e preservação de dados para essas instituições.

Um conjunto de dados no Dataverse é um contêiner para seus dados, documentação, código e os metadados que descrevem esse conjunto de dados. Cumprindo seu perfil científico, o Dataverse estrutura a citação de dados gerando um Identificador de Objeto Digital (DOI) para os conjuntos de dados e fornece algum suporte para vincular a submissão do artigo científico a uma revista científica juntamente com a submissão do conjunto de dados na plataforma (a revista precisa cumprir algumas premissas para isso). Embora seja possível armazenar quaisquer formatos de arquivos, incluindo ligações estruturais para outras instâncias Dataverse, a plataforma não fornece suporte para dados conectados. Para dados conectados, apenas a opção de *download* é disponibilizada. Para alguns formatos de arquivos, como PDF e imagens, é possível a visualização de uma prévia na interface do próprio Dataverse. O Dataverse disponibiliza funcionalidades para busca, citação e reutilização de conjuntos de dados utilizados ou gerados por pesquisas científicas.

Há três níveis de metadados para cada conjunto de dados: de citação; de domínio específico (um suporte específico para dados de ciências da vida, sociais, geoespaciais e astronomia); e no nível de arquivo. Após a publicação de um conjunto de dados, seus metadados podem ser exportados em vários formatos e padrões de metadados, incluindo Dublin Core¹⁷ e Schema.org JSON-LD¹⁸ (FEATURES | THE DATAVERSE PROJECT - DATAVERSE.ORG, 2020). O intercâmbio de metadados entre repositórios pode ser apoiada por um *Harvesting Client* que utiliza *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH)¹⁹, uma interface para exposição de metadados armazenados muito conhecida na área de repositórios e pouco conhecida fora dela.

Pesquisas facetadas por outros repositórios, conjuntos de dados e arquivos contidos neles também são permitidas. É possível iniciar uma pesquisa pelo termo “desmatamento” e facetá-la selecionando-se um ano e, dentro do ano selecionado, selecionar uma outra palavra-chave como “bioma”, por exemplo.

Diferente do CKAN, o Dataverse fornece mecanismo de controle para vários níveis de acesso: dados abertos, dados com termos de uso e dados com restrições que exigem que o usuário seja autenticado e autorizado (CROSAS, 2015).

¹⁷ <https://www.dublincore.org/specifications/dublin-core/dces/>

¹⁸ <https://schema.org/docs/datamodel.html>

¹⁹ <http://www.openarchives.org/pmh/>

Por padrão, a todos os conjuntos de dados criados pela interface Web do Dataverse é atribuída uma licença de domínio público. Além disso, caso restrições sejam necessárias, termos de uso podem ser adicionados, indicando as restrições, permissões e requisitos de citações. Se algum arquivo pertencente a um conjunto de dados for restrito, a plataforma fornece suporte para estabelecimento de um termo de acesso.

No que diz respeito à personalização, é possível alterar alguns elementos na interface gráfica como cores de fundo, imagens de rodapé, imagens de logotipo, mas nada estrutural ou funcional.

2.2.3. DSpace

Outra iniciativa de código aberto desenvolvida com base na comunidade científica é o DSpace²⁰. Essa iniciativa surgiu no *Massachusetts Institute of Technology* (MIT), em parceria com a *Hewlett-Packard* (HP), tendo como foco inicial a disseminação de literatura científica, principalmente em formato de artigos que tenham sido publicados anteriormente em periódicos (SHINTAKU; MEIRELLES, 2010). Recentemente o DSpace tem sido amplamente utilizado por diversos tipos de organizações. Além disso, as discussões sobre capacidade de citação, disseminação e reúso de todos os recursos gerados e/ou utilizados por pesquisas, estimulou a ampliação do escopo das publicações, incluindo e, às vezes, enfatizando seu uso como repositório de dados, tal como o Metabuscador de Dados de Pesquisa da FAPESP. Atualmente o DSpace é mantido pelo Duraspace²¹ com apoio de uma comunidade mundial, num projeto que direta e indiretamente envolve muitos profissionais de vários países.

Como repositório, o DSpace organiza os recursos armazenados em estruturas alinhadas às necessidades organizacionais acadêmicas, de forma hierárquica, composto por comunidades, coleções, itens, *bundle* e arquivo binário. Um arquivo binário é um objeto digital (texto, vídeo, áudio, conjunto de dados, entre outros) que é organizado em uma estrutura denominada *bundle* pertencente a um item. Item é a nomenclatura utilizada pelo DSpace para uma unidade informacional, composta por um conjunto de metadados e, minimamente, dois objetos digitais representados pelo arquivo binário de conteúdo útil em si e sua licença, isto é, uma estrutura de conjuntos de dados. Coleções e comunidades (e subcomunidades) são estruturas organizacionais (SHINTAKU; MEIRELLES, 2010). As coleções agrupam documentos com alguma característica em comum e pertencem a uma comunidade. Os arquivos binários podem ser de quaisquer formatos. Para arquivos que representem o texto da publicação científica, os possíveis conjuntos de dados gerados ou utilizados pela pesquisa

²⁰ <https://duraspace.org/dspace/>

²¹ <https://duraspace.org/>

exposta na publicação podem ser armazenados no mesmo item, como um arquivo binário em separado. Porém, nenhuma distinção para formatos dos conjuntos de dados é feita. Assim, mesmo que haja um conjunto de dados em RDF, ele será tratado como um arquivo binário cuja opção de uso disponível é o *download* de dados. Para alguns formatos de arquivos, é possível a visualização de uma prévia na interface do próprio DSpace, quando configurado. Uma característica de destaque é a identificação de alguns formatos de arquivo, para determinar se é uma instância bem formada/válida utilizando JHOVE²² integrado à plataforma.

Para armazenar um objeto digital, um processo de avaliação dos registros antes da publicação, com bastante foco na etapa de catalogação, deve ser seguido. O processo tem como base práticas de catalogação, organização de assuntos e coleções muito utilizadas pela comunidade científica (SHINTAKU; MEIRELLES, 2010).

O DSpace utiliza o *Dublin Core* como esquema de metadados e a cada item é atribuído uma URI. O conteúdo e os metadados podem ser lidos e exportados por aplicações externas ao DSpace, de modo a viabilizar mecanismos para interoperabilidade entre repositórios. O suporte original fornecido para intercâmbio de metadados é a exportação em *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH), assim como o Dataverse. Porém, de forma a ampliar as possibilidades de interoperabilidade entre repositórios, desde a versão 5 o DSpace fornece suporte para publicar conteúdo e metadados como dados conectados (DSpace 6.X DOCUMENTATION, 2020).

Em termos de pesquisas, em uma operação padrão, o usuário submeterá uma expressão a consultar e poderá aplicar alguns filtros. Adicionalmente, algumas combinações para pesquisas facetadas são exequíveis, porém, dependendo da combinação das facetas, a pesquisa desconsidera algumas delas e aplica um filtro mais amplo. As facetas são autor, assunto e ano. Ao selecionar um ano, por exemplo, a busca combinada por assunto não é disponibilizada, restando somente a combinação com autor. Um outra maneira de adicionar facetas às buscas em um repositório é a nuvem de etiquetas ("*tag cloud*"), cujo tamanho da fonte utilizada para a etiqueta denota a popularidade de uso dela no repositório. Quanto maior a fonte, mais popular.

O DSpace é capaz de aplicar políticas de acesso aos repositórios e de direitos de disseminação de conteúdos armazenados em repositórios. Destarte, o fato de um documento estar armazenado em um repositório não significa, necessariamente, que o mesmo esteja disponível ao público em geral (SHINTAKU; MEIRELLES, 2010). Esse fato se adequa à

²² <https://wiki.lyrasis.org/display/DSpace/JhoveIntegration>,
<https://jhove.openpreservation.org/>

necessidade de sigilo de alguns dados utilizados por pesquisas científicas, tais como dados pessoais que não passaram por um processo de anonimização, por exemplo.

Uma licença de permissão de uso pode ser especificada por ocasião da publicação de um item. Assim, os termos de uso são registrados em um campo do conjunto de metadados. O DSpace não estabelece diretamente restrições ao uso em função da licença. Mas, como foi descrito anteriormente, políticas de acesso ao repositório, podem ser configuradas.

Com relação aos aspectos de personalização, é possível configurar a interface de busca para os usuários e ajustar o DSpace às diversas necessidades de organização de comunidades de uma instituição. Entretanto, essas opções baseiam-se em conceitos que orientaram o desenvolvimento de um repositório, isto é, aspectos organizacionais.

2.2.4. Comparativo entre plataformas de dados abertos

Há literatura científica que compara as diferentes plataformas de dados abertos (ROCHA et al., 2018; GABRIEL JUNIOR et al., 2019). Todavia, as comparações muitas das vezes não incluem uma ou outra plataforma, assim como as comparam considerando características específicas e contextualizadas às necessidades de um determinado trabalho ou iniciativa. Em função disso, as características das três plataformas de dados abertos, e de código aberto, contextualizadas às necessidades desta pesquisa, foram apresentadas nas subseções 2.2.1, 2.2.2 e 2.2.3. As características foram organizadas em categorias para estabelecimento de um comparativo entre elas. O Quadro 2 apresenta um comparativo das características observadas nas plataformas de dados abertos e de código aberto.

Dentre as semelhanças, observa-se que nas três plataformas os dados organizados recebem identificadores únicos globais e persistentes, assim como as distribuições associadas a esses dados. Além disso, nas três soluções arquivos de quaisquer naturezas são aceitos para publicação, buscas facetadas são suportadas, mas nenhuma delas fornece suporte para dados conectados. Apesar disso, todas fornecem suporte para intercâmbio de dados com outras plataformas de dados abertos.

Como o Dataverse e o DSpace estruturam a organização dos dados considerando alguma literatura científica que os referencia, até mesmo estruturando a citação, entende-se que ambos estão bastante voltados ao suporte da comunidade científica, distinguindo-os do CKAN, onde uma referência para os dados inclui a organização que os publicou.

Quadro 2: Comparativo das plataformas de dados abertos de código aberto

	CKAN	Dataverse	DSpace
Unidade mínima de organização dos dados	Conjunto de dados	Conjunto de dados	Conjunto de dados
Identificadores globais e persistentes	Sim, via HTTP URI	Sim, via DOI	Sim, via HTTP URI
Identificadores globais e persistentes para distribuições relacionadas	Sim, via HTTP URI	Sim, via DOI	Sim, via HTTP URI
Metadados anotados com vocabulários	Não, mas existe uma extensão que permite exportação de metadados anotados para uso fora da plataforma	Não, mas permite nativamente a exportação de metadados anotados para uso fora da plataforma	Sim, nativamente
Natureza da organização dos dados	Arquivo de qualquer formato: Texto, Tabelas, Multimídia, RDF etc.	Arquivo de qualquer formato: Texto, Tabelas, Multimídia, RDF etc.	Arquivo de qualquer formato: Texto, Tabelas, Multimídia, RDF etc.
Formatos de arquivos aceitos (gerenciados pela solução)	Não verifica formatos, apenas os identifica pela terminação do nome do arquivo. Fornece mecanismos para visualização de alguns formatos.	Não verifica formatos, apenas os identifica pela terminação do nome do arquivo. Fornece mecanismos para visualização de alguns formatos.	Usa JHOVE para verificar formatos. Fornece mecanismos para visualização de arquivos.
Fornecer suporte a dados conectados publicados	Não	Não	Não
Permite pesquisas facetadas	Sim	Sim	Sim
Intercâmbio com outras instâncias	Sim, com uso de extensão por meio de dados conectados	Sim, por meio de OAI-PMH	Sim, por meio de OAI-PMH ou dados conectados
Permite restrições de acesso de acordo com a licença de uso	Não	Sim	Sim
Permite personalização	Sim, da interface gráfica, das funcionalidades por meio do desenvolvimento de extensões e da instalação de extensões já desenvolvidas por terceiros, e por meio da seleção de quais funcionalidades de exploração de dados serão disponibilizadas pela instância do portal	Sim, de itens básicos da interface gráfica	Sim, da interface de pesquisa e da estrutura organizacional do repositório

Dentre as diferenças, observa-se que, dentre as três plataformas de dados abertos comparadas, o DSpace é a única que nativamente já disponibiliza um esquema de metadados anotado semanticamente. E quando restrições de uso devem ser consideradas, o CKAN é o que não permite restrições de acesso de acordo com a licença de uso.

A possibilidade de desenvolvimento e uso de extensões já codificadas e testadas, associada ao fato da maioria das iniciativas de ODP publicarem seus dados abertos fazendo uso da plataforma CKAN (COSTA et al., 2017; NEUMAIER et al., 2017), levou a escolha desta plataforma para uso nos experimentos da abordagem proposta.

2.3. Web Semântica e suas tecnologias

A Web Semântica é uma extensão da Web de documentos. Ela fornece a infraestrutura que permite páginas da Web, bancos de dados, serviços, programas, sensores e até dispositivos consumam e produzam dados na Web. Todo esse conteúdo é compartilhado e processável por ferramentas automatizadas e por pessoas. A infraestrutura da Web Semântica fornece as condições para o desenvolvimento da Web de Dados. A Web de Dados estende os benefícios da Web de documentos especificamente aos dados publicados na Web (HENDLER; BERNERS-LEE; MILLER, 2002). Tradicionalmente, dados são publicados na Web em formatos tais como CSV, PDF, XML ou como tabelas embutidas em código HTML, sacrificando a estrutura e a semântica. Na convencional Web de hipertextos, a natureza dos relacionamentos entre dois documentos é implícita, já que HTML não é suficientemente expressiva em termos semânticos, isto é, não é capaz de descrever o tipo de relacionamento entre os documentos.

Por trás dessa evolução para a Web de Dados está um conjunto de práticas recomendadas para publicar e conectar dados estruturados na Web, conhecido como *linked data* ou dados conectados, na sua tradução para português pela W3C. Dados conectados são conjuntos de dados de diferentes fontes, publicados na Web, interligados por meio de relacionamentos tipados, isto é, definidos conceitualmente em vocabulários e capazes de serem inteligíveis por máquina. Desta forma, a semântica dos relacionamentos é explicitamente definida (BIZER; HEATH; BERNERS-LEE, 2011). Sendo assim, dados conectados formam um caminho cujo fim é a Web Semântica.

Dentre as tecnologias de Web Semântica, o *Resource Description Framework* (RDF) é a base para publicação de dados conectados. RDF organiza dados na forma de triplas compostas por sujeito, predicado e objeto, e utiliza *Uniform Resource Identifiers* (URIs) para nomear e identificar recursos individuais na Web. Recursos são quaisquer tipos de objetos ou

conceitos (RDF - SEMANTIC WEB STANDARDS, 2020). Assim, ao predicado de uma tripla também é atribuído um URI. A estrutura base do modelo de dados de qualquer expressão em RDF é representada como um grafo que consiste em nós e arcos direcionados e rotulados que vinculam pares de nós (STELLATO, 2012). Por isso, um conjunto de triplas é chamado de grafo RDF. A Figura 3-a apresenta a estrutura de um grafo RDF; a Figura 3-b apresenta um exemplo de grafo RDF, isto é, uma instância do mundo real.

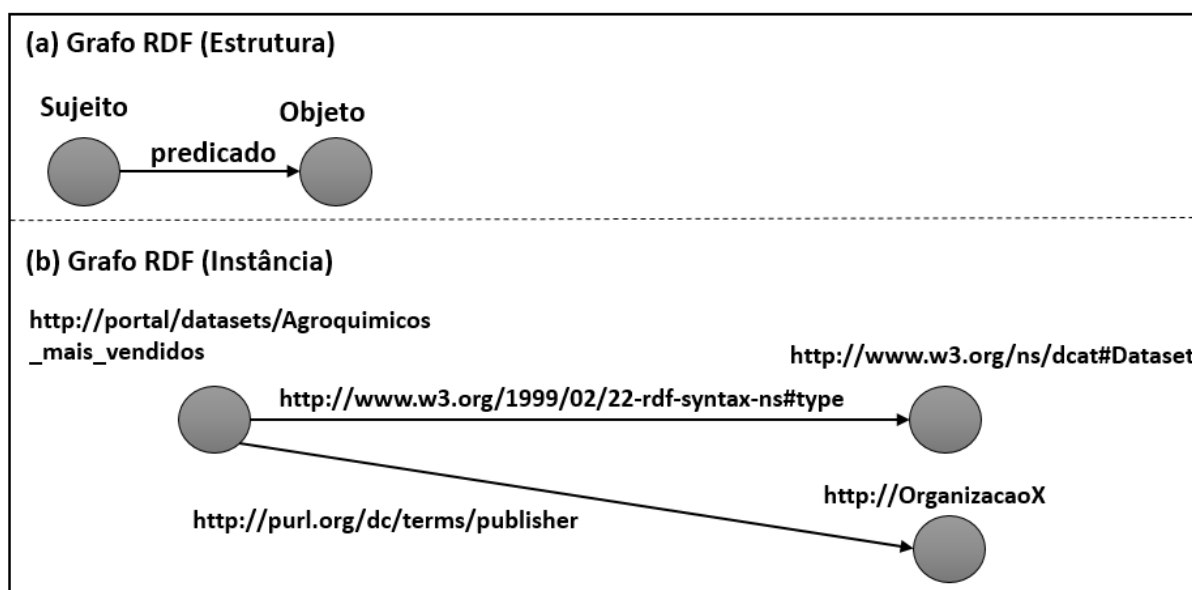


Figura 3: Uma estrutura e uma instância de grafo RDF com um nó sujeito conectado por dois arcos direcionados e rotulados a dois objetos distintos

As triplas RDF podem ser escritas seguindo alguns formatos de serialização, que representam a mesma informação, diferindo no aspecto de facilidade de leitura por humanos e máquinas. Para as plataformas de dados abertos, a serialização não faz diferença, visto que essas plataformas não tratam arquivos em RDF considerando os aspectos semânticos de dados conectados. N-Triples, RDF/XML, N3 e Turtle são tipos de serialização RDF (SPAHIU, 2017). A Figura 4 apresenta uma tripla RDF representada no grafo apresentado na Figura 3-b serializada em N-Triples.

```
<http://dataset/agroquimicos_mais_vendidos>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/ns/dcat#Dataset> .
```

Figura 4: Uma tripla RDF serializada em N-Triples

As triplas RDF também são escritas segundo um modelo de dados. *RDF Schema* (RDFS) é um vocabulário para criação de modelos de dados em RDF. *Web Ontology Language* (OWL) é uma linguagem para construção de ontologias Web. OWL estende RDFS. Enquanto RDFS é utilizada para definição da estrutura dos dados, OWL descreve a semântica embutida

nos relacionamentos entre os recursos. Ambas estão definidas como recomendação da W3C²³.

Para armazenar dados em RDF são utilizados bancos de dados de triplas denominados *Triplestores*. Sendo assim, *Triplestores* armazenam e recuperam dados na forma de triplas, ou seja, dados modelados em RDF cujo modelo de dados é um grafo direcionado e nomeado. Virtuoso²⁴ e GraphDB²⁵ são exemplos de *Triplestores*. Para recuperar os dados armazenados nessas estruturas faz-se necessária uma linguagem de consulta.

SPARQL é a sigla em inglês para *Simple Protocol and RDF Query Language*. Como o próprio nome esclarece, trata-se de um protocolo e de uma linguagem de consulta. Ambos são recomendações da W3C^{26,27}. Um serviço SPARQL é uma implementação do protocolo SPARQL (*SPARQL Protocol*). O protocolo estabelece como transmitir consultas e atualizações em linguagem SPARQL (*SPARQL Query Language*) para processadores de consulta e como os dados devem ser retornados. Um serviço de processamento SPARQL retorna os resultados via HTTP para a entidade que os solicitou. Neste sentido, um SPARQL *Endpoint* é um serviço SPARQL, isto é, um serviço que implementa o protocolo SPARQL, que atende solicitações de clientes que também seguem esse protocolo. O serviço é disponibilizado a clientes como um serviço HTTP, isto é, identificado por uma determinada URL. Há diversos SPARQL *Endpoints* disponibilizados por diferentes iniciativas na Web. A diversidade de serviços gera a oportunidade de identificar algum nível de correspondência entre recursos em RDF distintos. Neste sentido, a linguagem de consulta SPARQL oferece mecanismos para o disparo de consultas a serviços remotos por meio das consultas federadas. Assim, uma consulta federada é executada em um SPARQL *Endpoint* remoto e seu resultado é combinado com o resultado da consulta que a disparou, gerando um único conjunto de triplas como resultado. Consultas federadas também estão definidas como uma especificação da W3C²⁸.

Triplas RDF precisam passar por processos de enriquecimento semântico de forma a se tornarem mais úteis sob o ponto de vista da descoberta de conhecimento promovida por dados conectados. O enriquecimento semântico de conjuntos de dados em RDF pode ser realizado basicamente de duas maneiras: por anotação semântica ou pelo estabelecimento de ligações entre conjuntos de dados por meio de relações existentes entre as instâncias de dados dos conjuntos (SHERIF, NGOMO, LEHMANN, 2015).

²³ <https://www.w3.org/TR/rdf-schema/> e <https://www.w3.org/TR/owl-ref/>

²⁴ <https://virtuoso.openlinksw.com/>

²⁵ <https://www.ontotext.com/products/graphdb/graphdb-free/>

²⁶ <https://www.w3.org/TR/rdf-sparql-protocol/>

²⁷ <https://www.w3.org/TR/rdf-sparql-query/>

²⁸ <https://www.w3.org/TR/sparql11-federated-query/>

Anotação semântica é uma associação entre expressões ou termos relevantes de um recurso da Web Semântica a metadados, conceitos e instâncias descritas em um instrumento terminológico, com uma ontologia (BELLOZE et al., 2012). Isso posto, a associação trata de descrições e conceituações de recursos utilizando instrumentos terminológicos. Outro tipo de instrumento terminológico são os vocabulários. No contexto de Web Semântica, vocabulários são coleções de classes e propriedades que também são expressos em RDF. Os vocabulários também podem estar relacionados a outros vocabulários, definindo-se um mapeamento entre eles utilizando triplas RDF (BIZER; HEATH; BERNERS-LEE, 2011), conforme detalhado mais adiante nesta seção. O uso de vocabulários é uma boa prática para dados conectados (SCHMACHTENBERG; BIZER; PAULHEIM, 2014).

Ontologias são recursos mais expressivos do que vocabulários, utilizadas para representação do modelo conceitual de um determinado domínio, que incluem um conjunto de axiomas com regras de utilização bem definidas para conceituação de indivíduos e dos tipos de relacionamentos entre eles (CIMIANO, 2006). Ontologias computacionais são um meio para formalmente modelar a estrutura de um sistema, com as entidades e relações que emergem da observação do ponto de vista consensual entre várias partes (GUARINO; OBERLE; STAAB; 2009).

No contexto de Web Semântica, as palavras “vocabulário” e “ontologia” são utilizadas sem uma clara distinção de escopo. De um modo geral, a palavra “ontologia” é utilizada para referenciar uma coleção mais formal de termos, enquanto a palavra “vocabulário” é utilizada quando a formalização de termos é menos rigorosa que nas ontologias (SPAHIU, 2017). Além disso, vocabulários não estão necessariamente associados a uma conceitualização mais completa, faltando na maior parte das vezes, relações mais ricas entre os termos.

Neste contexto, tesauros também são uma espécie de vocabulário concebidos, inicialmente, com uma abordagem centrada no usuário (MOREIRA, 2003). Embora inicialmente as pessoas considerassem tesauros como dicionários de sinônimos, eles têm um campo de aplicação muito mais amplo e carregam muito mais informações do que apenas sinônimos. Um tesauro é um vocabulário controlado e estruturado composto por palavras ou frases que representam uma ideia, denominadas termos. Termos representam conceitos (unidades de pensamento) e as relações estabelecidas entre eles (MARTINEZ-GONZÁLES; ALVITE-DÍEZ, 2019). Tesauros não incluem axiomas e os termos são organizados em estruturas que estabelecem relações com outros termos. Essas relações podem ser de equivalência, hierarquia e associação (SALES; CAFÉ, 2009). Essas estruturas de sistemas de organização do conhecimento existentes podem ser transportadas para a Web Semântica

utilizando modelos de dados como o *Simple Knowledge Organization System* (SKOS)²⁹ (SKOS, 2020). SKOS é um esquema RDF, sendo considerado um modelo de dados e um vocabulário. Está organizado em classes e propriedades para construção de sistemas de organização do conhecimento, tal como tesouros. SKOS é um padrão para compartilhamento de dados definidos ou conceituados em diferentes sistemas de conhecimento, criando uma espécie de ponte entre diferentes conceituações. Em função disso, a maior parte de suas propriedades está relacionada ao estabelecimento de relações entre conceitos, visto que *Concept* é sua principal classe. O modelo de dados do SKOS está definido como triplas RDF e, por consequência, suas classes e propriedades também são empregadas em triplas para definições de recursos da web semântica, isto é, são empregadas como vocabulário. Ou seja, SKOS é uma aplicação de RDF.

Em suma, vocabulários, ontologias e tesouros são os habilitadores das atividades de anotação semântica e são fundamentais para compartilhamento do conhecimento e interoperabilidade. Embora possam conceituar elementos que tenham algum tipo de relação, ontologias são construídas em diferentes níveis de abstração (SCHMIDT; TROJAHN; VIEIRA, 2016). Comparando-se tesouros e ontologias evidencia-se que as ontologias transcendem as possibilidades de relacionamento conceitual dos tesouros (SALES; CAFÉ, 2009). Por isso, mecanismos de alinhamento de ontologias são essenciais para usufruir de todos os benefícios que ontologias oferecem. As técnicas de alinhamento de ontologias dependem de dois fatores: do que será alinhado, como esquemas, suas instâncias ou ambos; e de quais outras fontes existem, como outros vocabulários e ontologias, e se já estão alinhadas (CROSS, 2018). O alinhamento de ontologias funciona como um acordo mútuo, onde duas ontologias se tornam consistentes e coerentes, gerando como resultado final duas ontologias em separado, mas com ligações adicionais entre suas entidades equivalentes (LACHTIM; MOURA; CAVALCANTI, 2009). Em função disso, alinhamento de ontologias é atividade importante para ampliar as possibilidades de interligação de conjuntos de dados.

Dados conectados devem ser os mais descritivos possíveis e, portanto, devem incluir metadados (SCHMACHTENBERG; BIZER; PAULHEIM, 2014). Conjuntos de dados em RDF naturalmente fornecem metadados, ao mesmo tempo que fornecem organização para eles. Significa dizer que as descrições dos dados fazem parte da estrutura do RDF.

Uma maneira ostensiva para se descrever e observar metadados em dados em RDF é por meio de declarações RDF sobre declarações RDF. Esse tipo de estrutura chama-se reificação — “RDF Reification”³⁰. A proposta de reificação foi apresentada desde os primórdios do *framework* RDF. Porém, mais recentemente, uma abordagem alternativa, mas ainda

²⁹ <https://www.w3.org/TR/skos-reference/> e <https://www.w3.org/TR/swbp-skos-core-spec/>

³⁰ <https://www.w3.org/TR/rdf11-mt/#whatnot>

compatível com a anterior, foi apresentada no intuito de eliminar alguns inconvenientes da proposta original, tal como a necessidade de inclusão de quatro triplas adicionais para cada tripla sendo descrita. Essa nova abordagem de reificação RDF é chamada de RDF*. RDF* baseia-se na ideia de usar uma tripla diretamente na posição de sujeito ou de objeto de (outras) triplas, ou seja, aninhar (embutir) uma tripla, que represente metadados, como sujeito ou objeto de outra tripla (HARTIG; THOMPSON, 2019). A linguagem de consulta SPARQL adaptada para acomodar a estrutura reificada é chamada SPARQL*. A Figura 5 apresenta um conjunto de triplas RDF* serializadas em Turtle*, uma extensão da gramática de Turtle para acomodar as triplas embutidas. Em Turtle* as triplas embutidas são delimitadas pelos símbolos "<<" e ">>".

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix ex: <http://ex.org/#> .

<<ex:Jose ex:temEsposa ex:Maria>> ex:dataCasorio "2020-02-11"^^xsd:date .
```

Figura 5: Representação de tripla embutida serializada em Turtle*

A extração de metadados pode ser favorecida quando os conjuntos de dados em RDF são enriquecidos por meio de anotação semântica. Para plataformas de dados abertos, boas práticas recomendam a anotação de metadados identificando as partes que os compõem. Vocabulários como Schema.org³¹, Dublin Core Metadata Initiative³², DCAT (Data Catalog Vocabulary)³³ e VOID (Vocabulary of Interlinked Datasets)³⁴ são exemplos de alternativas para anotação de metadados. De um modo geral, os exemplos de vocabulários citados, estendem partes entre si. Especialmente o DCAT utiliza alguns termos definidos por Dublin Core, além de outros vocabulários. Aplicando essas reutilizações, DCAT tornou-se robusto e tem sido referenciado e aplicado (PESCE et al., 2015; NEUMAIER, UMBRICH, POLLERES, 2017; MOREIRA et al., 2019). O Quadro 3 apresenta alguns vocabulários, bem como algumas classes e propriedades desses vocabulários, alinhados às tecnologias da Web Semântica, que podem ser úteis para anotação semântica de conjuntos de metadados.

³¹ <https://schema.org/docs/data-and-datasets.html>

³² <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>

³³ <https://www.w3.org/TR/vocab-dcat-2/>

³⁴ <https://www.w3.org/TR/void/>

Quadro 3: Exemplos de vocabulários e classes para anotação semântica

Vocabulário	Descrição e exemplos de classes definidas no vocabulário
Friend of A Friend (FOAF)	Vocabulário definido como parte do FOAF <i>Project</i> que define classes e propriedades para anotar relações sociais entre pessoas e redes de informações. http://xmlns.com/foaf/0.1/Organization → classe utilizada para representar instituições sociais, tais como companhias, empresas, sociedades. http://xmlns.com/foaf/0.1/name → propriedade que define um nome para alguma coisa.
Data Catalog Vocabulary (DCAT)	Vocabulário definido como uma recomendação da W3C com o intuito de facilitar a interoperabilidade de catálogos na Web. Considerando que para interoperar faz-se necessário um conjunto de metadados bem definido, DCAT fornece classes e propriedades para definição das características de conjuntos de dados e serviços de dados disponibilizados em catálogos de dados. http://www.w3.org/ns/dcat#Dataset → classe que representa uma coleção de dados, ou seja, um conjunto de dados conceitual que pode ser disponibilizado (distribuído) sob um ou mais formatos pelo publicador. http://www.w3.org/ns/dcat#Relationship → classe que caracteriza uma relação entre recursos definidos com DCAT, e sobre a qual podem ser adicionadas informações. http://www.w3.org/ns/dcat#Distribution → classe para indicar uma representação específica de um conjunto de dados. DCAT define que um conjunto de dados pode estar disponível em várias distribuições que podem diferir de várias maneiras, incluindo esquema de organização, resolução temporal e espacial, nível de detalhe, etc. Notas na documentação do vocabulário indicam que a decisão de como organizar os conjuntos de dados e suas distribuições são do publicador de dados, caso o responsável pelo portal de dados não tenha definido um padrão (DCAT, 2020). Assim, um conjunto de dados pode ter uma distribuição para cada ano ao qual os dados se referem ou os mesmos dados organizados em diferentes formatos de serialização em RDF, por exemplo. http://www.w3.org/ns/dcat#distribution → propriedade que conecta uma classe do tipo conjunto de dados (dcat:Dataset) a uma classe do tipo distribuição (dcat:Distribution). Observa-se que a propriedade tem o mesmo nome da classe. Contudo, seguindo as boas práticas para nomenclatura de classes e propriedades, a diferença sintática está na escrita, onde a classe inicia-se com letra maiúscula e a propriedade inicia-se com letra minúscula. A diferença semântica está na definição formal do vocabulário que especifica um domínio (classes que podem estar como sujeito da tripla) e um <i>range</i> (classes que podem estar como objeto de uma tripla). http://www.w3.org/ns/dcat#theme → propriedade utilizada para especificar temas/assuntos de um recurso.
VoID	Vocabulário que provê termos e padrões para descrever conjuntos de dados em RDF como uma ponte entre os publicadores e o consumidores de dados. O vocabulário pode ser usado para expressar metadados, incluindo formas de acesso e interligações entre conjuntos de dados.

Vocabulário	Descrição e exemplos de classes definidas no vocabulário
	http://rdfs.org/ns/void#sparqlEndpoint → propriedade que especifica uma forma de acesso a um conjunto de dados via protocolo SPARQL.
DCMI Metadata Terms	Vocabulário organizado pela <i>Dublin Core™ Metadata Initiative</i> (DCMI) que define um conjunto de classes e propriedades que podem ser combinadas com termos de outros vocabulários para definições de perfis para aplicações específicas (AP - <i>Application Profiles</i>). O vocabulário DCAT é um exemplo para o qual podem ser criados AP em combinação com DCMI. Há quatro <i>namespaces</i> definidos: http://purl.org/dc/elements/1.1/, http://purl.org/dc/terms/, http://purl.org/dc/dcmitype/ e http://purl.org/dc/dcam/. http://purl.org/dc/terms/publisher → propriedade utilizada para ligar o recurso publicado e o recurso que represente a entidade que disponibilizou o recurso. http://purl.org/dc/terms/references → propriedade utilizada para indicar o recurso objeto que é referenciado, citado ou apontado pelo recurso sujeito. http://purl.org/dc/dcam/memberOf → propriedade utilizada para definir um relacionamento entre um recurso e um esquema de codificação de vocabulário, que indica que o recurso é membro de um conjunto.
Vocabulary of A Friend (VOAF)	Vocabulário que provê classes e propriedades para descrição, extensão, especificação e vinculação de vocabulários. http://purl.org/vocommons/voaf#Vocabulary → classe que tipifica um vocabulário utilizado para anotação de uma distribuição de um conjunto de dados ou referenciado em uma nuvem de dados conectados.
VANN	Vocabulário para anotação de descrições de vocabulários. http://purl.org/vocab/vann/preferredNamespacePrefix → propriedade que indica o prefixo preferencial para identificação de um vocabulário em uma distribuição de um conjunto de dados.

Dados conectados são disponibilizados na Web seguindo alguns princípios relacionados à interligação e anotação de dados, sob algum tipo de licença aberta, fazendo reuso de vocabulários e ontologias de acordo com seus objetivos. Dados em RDF que descrevem vocabulários e ontologias e que sigam todos os princípios de *Linked Open Data* (LOD) definidos pelo inventor da Web, Sir Tim Berners-Lee, formam um ambiente de nuvem de dados conectados organizadas por domínio que, interligados, formam uma nuvem geral – a nuvem de dados conectados (LOD Cloud)³⁵. No centro das iniciativas de LOD, reside a ideia de interligar recursos de informação não associados originalmente, aproveitando a infraestrutura da Web usando HTTP, URIs e RDF como uma estrutura de representação de dados (CORDEIRO, 2011). Além disso, muitos conjuntos de dados em RDF que, não

³⁵ <https://lod-cloud.net/>

necessariamente sejam vocabulários, mas estejam anotados com vocabulários, ontologias ou tesouros de domínios específicos, também são disponibilizados na Web em plataformas de dados abertos, viabilizando o acesso do público em geral e organizando os dados considerando suas temáticas.

2.4. Princípios de Dados FAIR

A necessidade de organização de dados abertos, demandou o surgimento de iniciativas de gerenciamento de dados como o FAIR. FAIR é o acrônimo em inglês para *"Findable"*, *"Accessible"*, *"Interoperable"* e *"Reusable"*, que representam facetas de organização de um conjunto de princípios orientadores que buscam tornar dados gerados e/ou utilizados, inicialmente, por pesquisas científicas mais fáceis de serem descobertos, acessíveis, interoperáveis e reutilizáveis, tanto por pessoas quanto por máquina. Os princípios representam requisitos baseados nas melhores práticas de gerenciamento e arquivamento de dados e são aplicáveis tanto para dados quanto para metadados (WILKINSON et al., 2016). Embora o foco inicial tenha sido dados científicos, as diretrizes surgiram no contexto de dados abertos. Sendo assim, quaisquer dados abertos podem ser gerenciados e/ou administrados alinhados aos princípios FAIR. O detalhamento dos princípios FAIR pode ser consultado em Wilkinson, et al. (2016). A Figura 6 apresenta um resumo dos princípios FAIR.

Findable (F) – Localizáveis F1. Identificadores únicos e persistentes F2. Metadados enriquecidos F3. Interligação de dados e metadados F4. Metadados indexados	Accessible (A) – Acessíveis A1. Uso de protocolos de comunicação padronizados A1.1. Protocolos abertos e livres A1.2. Autorização e autenticação (quando necessário) A2. Metadados sempre disponíveis
Interoperable (I) – Interpreráveis I1. Representação do conhecimento por uma linguagem formal, acessível e compartilhada I2. Vocabulários que utilizem identificadores únicos e persistentes I3. Ligações entre dados deve ser descrita, rotulada	Reusable (R) – Reutilizáveis R1. Metadados descritos por atributos precisos e relevantes R1.1. Metadados associados a licenças de uso R1.2. Proveniência de metadados R1.3. Adoção de padrões de arquivamento e compartilhamento em comunidades

Figura 6: Resumo dos princípios FAIR (Adaptado de Wilkinson et al. (2016))

Os princípios FAIR são orientadores, no sentido de não especificar aspectos de implementação. Todavia, há grupos que têm se dedicado à definição de especificações para implementação dos princípios, conforme descrito nas seções 2.4.1 e 2.4.2.

2.4.1. Processo de FAIRificação

O processo de FAIRificação (*"FAIRification process"*) tem o intuito de colocar em prática os princípios FAIR. Neste sentido, um fluxo de atividades genérico foi especificado por Jacobsen et al. (2020) com a intenção de servir como guia para adequação de fontes de dados e metadados, independente do domínio, aos princípios FAIR.

O processo de FAIRificação (*"FAIRification"*) é dividido em três fases: Pré-FAIRificação, FAIRificação e Pós-FAIRificação. Na primeira fase são identificados os objetivos do processo e análises de dados e metadados. A análise dos dados foca na preparação para a interoperabilidade; a análise dos metadados quanto à disponibilidade de serem localizáveis, acessíveis e reutilizáveis. Na segunda fase ocorrem as transformações dos dados e dos metadados de forma a alinhá-los aos princípios FAIR. Para isso, modelos semânticos são definidos e as transformações de dados e metadados para formatos legíveis por máquina são executadas. Em seguida, dados e metadados são disponibilizados para humanos e máquinas. Na terceira fase são avaliados os resultados em relação aos objetivos definidos na primeira fase. A Figura 7 apresenta as atividades do processo para adequar dados e metadados aos princípios FAIR.

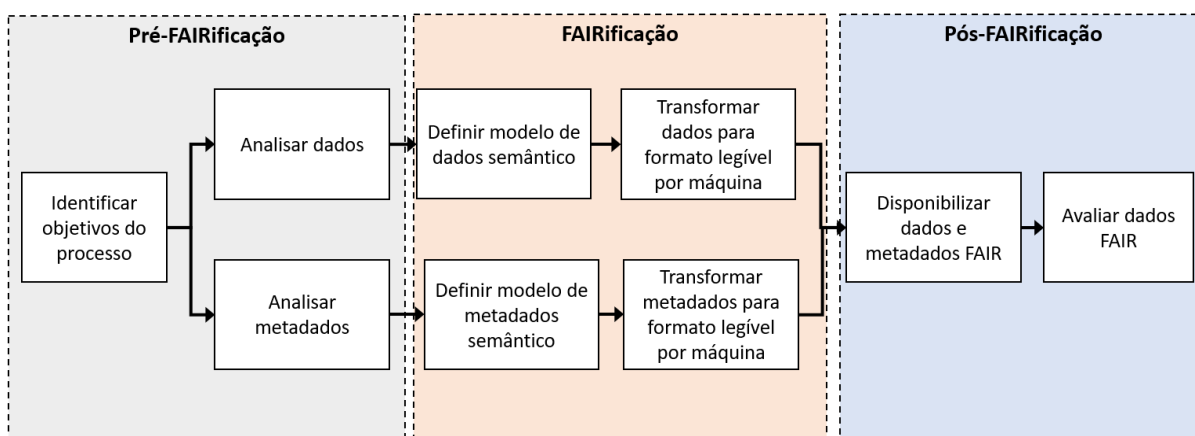


Figura 7: Fluxo de atividades do processo para tornar dados e metadados FAIR (Adaptado de Jacobsen et al. (2016))

Após a execução do processo de FAIRificação, surge a necessidade de especificações sobre o local onde dados e metadados FAIR serão armazenados e disponibilizados. Neste sentido, um FAIR *Data Point* (FDP), detalhado na próxima seção, pode ser utilizado.

2.4.2. FAIR Data Point

Considerando a disseminação de metadados em vários repositórios, há uma tendência ao estabelecimento de pontos centrais de metadados para melhorar a descoberta de dados por seres humanos e por máquina. Como parte da iniciativa FAIR, esses pontos centrais são chamados de FAIR *Data Points* (FDP) e são modelados como repositórios de metadados para acesso a dados de forma federada.

A arquitetura de um FDP inclui interface gráfica para o usuário baseada na Web e uma interface de programação de aplicativos (API - *Application Provider Interface*) para expor suas funcionalidades aos usuários. O *Metadata Provider* é um componente central, sendo o responsável pelo provisionamento do conteúdo de metadados disponível no FDP. Nesse sentido, foram definidos esquemas de metadados em cinco camadas complementares de organização de dados, alinhados aos princípios FAIR. As cinco camadas de metadados são: para o próprio repositório (nível 1); para os catálogos (nível 2); para os conjuntos de dados (nível 3); para as distribuições dos conjuntos de dados (nível 4); e para os registros de dados (nível 5). Na época da escrita desta dissertação, a especificação do esquema de metadados para um FDP fazia menção à camada de metadados para registros de dados, todavia ela ainda não estava completamente especificada, indicando somente que os metadados deveriam descrever os conteúdos dos dados por meio dos tipos envolvidos, dos domínios, das relações com outros tipos, etc. (MOREIRA et al., 2019; FAIRDATAPOINT-SPEC, 2020).

A divisão em camadas do *Metadata Provider* estabelece uma hierarquia entre os metadados e facilita que mecanismos automatizados dos repositórios verifiquem e validem metadados definidos, incluindo regras de quais itens são obrigatórios (MOREIRA et al., 2019). A definição de uma hierarquia entre os metadados, onde no topo os metadados são menos granulares, até uma base mais granular, move-se em direção a uma melhor capacidade de descoberta de conjunto de dados (*Findability*). O aumento de granularidade dos metadados ocorre à medida em que se detalha metadados a cada nível de forma a prover mais clareza sobre o conteúdo dos dados, sem a necessidade de acessar as instâncias de dados. Os metadados também incluem informações sobre licenças de uso de conjuntos de dados e protocolos de acesso, facilitando a interoperabilidade e a reutilização de dados, conforme proposto pela iniciativa FAIR. Ademais, fornece meios para a interligação de dados. Cabe ressaltar que a especificação de um FDP não configura um novo padrão, mas sim a combinação de padrões existentes para atingir os objetivos dos princípios FAIR (MOREIRA et al., 2019). A Figura 8 apresenta o esquema de metadados em níveis para um FDP.

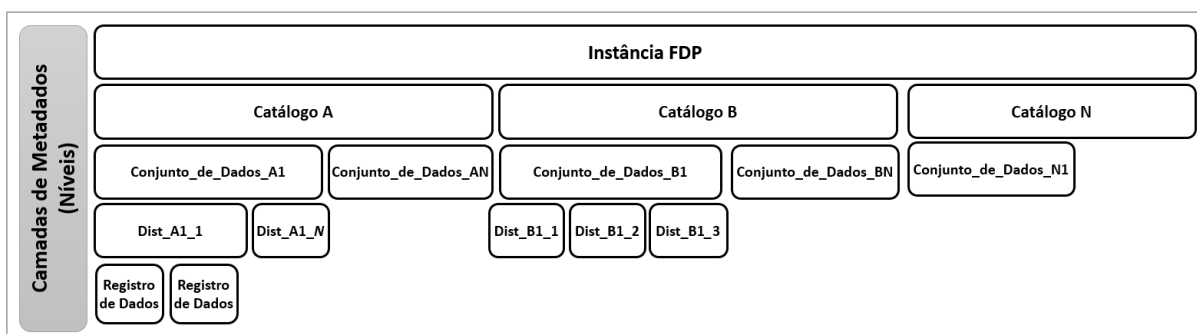


Figura 8: Esquema de metadados em camadas para um FDP

Para cada esquema de metadados é especificado um identificador do conjunto de metadados e suas datas de criação e de modificação no FDP, por meio do uso de uma ontologia própria para o FDP ("*FDP Ontology*"). A Figura 9 apresenta um extrato das propriedades utilizadas.

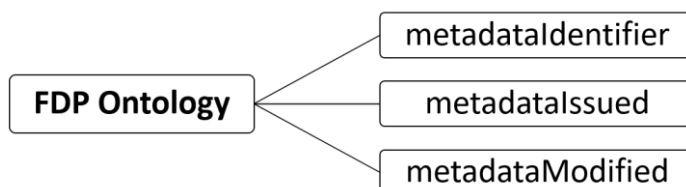


Figura 9: Propriedades da *FDP Ontology* (Fonte: FAIRDATAPoint-SPEC, 2020)

Ainda de acordo com a especificação, cada camada de metadados deve ser tipificada. Isto é, cada camada deve ter a explicitação do tipo de recurso para o qual os metadados estão sendo organizados. Para isso, seguindo o *profile* do FDP é utilizado o predicado "*rdf:type*" que aponta para um tipo definido no vocabulário DCAT. Assim, tem-se a tipificação para o catálogo (*dcat:Catalog*), para o conjunto de dados (*dcat:Dataset*) e para a distribuição (*dcat:Distribution*).

No entanto, a tipificação da camada que contém os metadados de registros de dados requer algumas considerações. O vocabulário DCAT não tem especificamente uma classe que tipifica registros de dados, assim como tem para outros níveis de organização de dados. Pode-se dizer que metadados para registros de dados referem-se às descrições dos conteúdos dos dados individualmente, independentemente de como estejam organizados: seja em linhas de um ou mais tabelas, seja em triplas em um ou mais conjuntos de dados modelados em RDF. Isso significa que é preciso identificar à qual estrutura os registros (os dados) pertencem para que seja possível identificar a camada de metadados de registros de dados.

É importante esclarecer que o nível de metadados de registro de dados deve ser recuperado a partir do nível de conjuntos de dados (FAIRDATAPoint-SPEC, 2020). Entretanto, quando se considera os conjuntos de dados em portais construídos sobre plataformas de

dados abertos, as distribuições são os elementos que contêm os dados em si. As distribuições podem variar em escopo, formato, estrutura e nível de detalhamento dos dados, assim como foi explicado na seção 2.3. Embora a variação no nível de detalhamento possa parecer não adequada ao conceito de “conjunto de dados”, ela é frequente ao se consultar portais de dados abertos governamentais. Isso ocorre porque o publicador de dados organiza sua publicação em uma área temática de alto nível, delegando às distribuições a definição da organização e escopo dos dados. De fato, a própria definição de distribuição no vocabulário DCAT considera estas possibilidades. Em função disso, faz-se necessária o uso de uma abordagem para tipificação da camada de metadados de registros de dados.

A especificação, que tem sido utilizada como referência para desenvolvimento de um FDP³⁶, define os elementos de metadados obrigatórios e opcionais para cada um dos níveis. Os esquemas de metadados definido para um FDP podem ser consultados no Apêndice A.

³⁶ <https://github.com/FAIRDataTeam/FAIRDataPoint-Spec/blob/master/spec.md>

3. INTERLIGAÇÃO DE DADOS ABERTOS

Interligação de dados é um princípio fundamental da Web de Dados. As interligações têm como objetivo principal agregar valor aos dados. As interligações entre dados podem ser utilizadas para descrever, associar e descobrir novas informações. A descoberta de novas informações é possível graças à capacidade de exploração e navegação por dados conectados (SCHMACHTENBERG; BIZER; PAULHEIM, 2014).

As interligações são utilizadas para descrever dados por meio da associação com outros dados que descrevem o recurso, enriquecendo-os, pela associação a vocabulários (SCHMACHTENBERG; BIZER; PAULHEIM, 2014). As descrições podem contextualizar os dados, melhorar a semântica, adicionar novas informações, revelar a fonte dos dados e até mesmo adicionar valores simples como seu tipo (CORDEIRO, 2015). A Figura 10 apresenta a interligação do recurso “http://dbpedia.org/page/Vinicius_de_Moraes” da ontologia DBPedia ao recurso “<http://dbpedia.org/page/Poet>” que representa uma de suas descrições na mesma ontologia.



Figura 10: Exemplo de descrição de recurso

As interligações são utilizadas para associar os dados por meio da associação a outros dados nos níveis de triplas, integrando-os (CORDEIRO, 2015). A Figura 11 apresenta o esquema de interligação de dois conjuntos de dados, por meio da utilização do mesmo recurso, utilizando a propriedade “`owl:sameAs`” definida em um vocabulário.

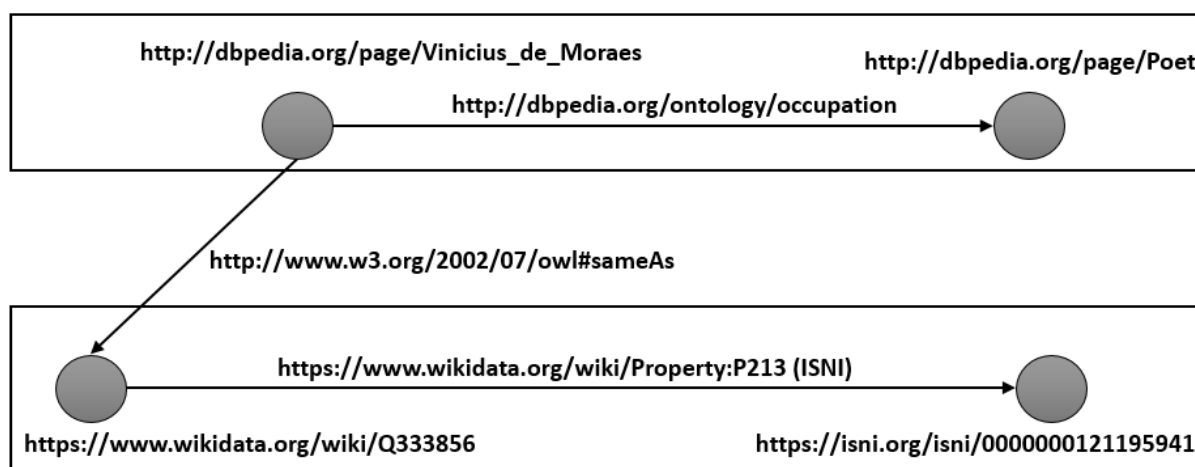


Figura 11: Exemplo de interligação por meio de uma propriedade

As interligações são utilizadas para descobrir novas informações por meio da navegação pelas ligações de um recurso a outro, permitindo a inferência de conhecimentos (CORDEIRO, 2015). A Figura 12 apresenta a navegação a partir do recurso “http://dbpedia.org/page/Vinicius_de_Moraes” e a descoberta de seu ISNI definido em outra ontologia (Wikidata), assim como a navegação a partir do recurso “<https://www.wikidata.org/wiki/Q333856>” e a descoberta de sua ocupação definida em outra ontologia (DBpedia).

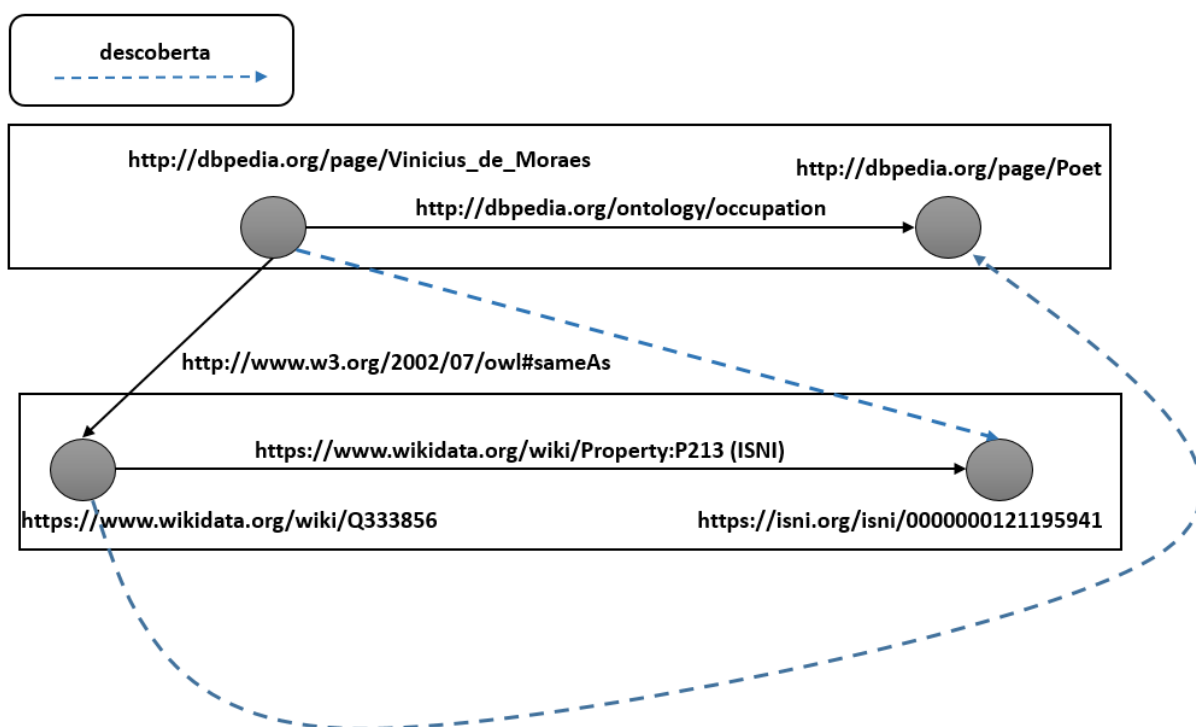


Figura 12: Exemplo de descoberta de novas informações por meio da navegação

Para apoiar a interligação de dados abertos, diversas abordagens foram desenvolvidas, as quais podem ser divididas em dois grupos. Na seção 3.1 são apresentados as abordagens que propõem apoiar a descoberta de interligações para enriquecer novos conjuntos de dados em RDF a serem publicados. Isso significa que eles não consideram os dados que desempenham o papel de metadados. Na seção 3.2 são apresentados as abordagens que propõem a interligação de dados considerando os metadados a eles associados. Por fim, na seção 3.3 é feita uma comparação entre as abordagens apresentadas.

Abordagens que focam na interligação de dados sem considerar os metadados podem: (i) utilizar cálculos estatísticos para identificação de potenciais conjuntos de dados a interligar; (ii) aplicar clusterização de conjuntos de dados; (iii) lançar sementes de buscas para alcançar potenciais interligações a outros conjuntos de dados; (iv) analisar similaridade de

expressões textuais; (v) analisar frequência de expressões textuais; e (vi) analisar relações entre elementos definidos em instrumentos terminológicos, como vocabulários, ontologias e tesouros. Essas estratégias podem ser utilizadas isolada ou simultaneamente.

Abordagens focadas na interligação de dados considerando os metadados publicados podem: (i) anotá-los com vocabulários, ontologias e tesouros; (ii) utilizar ou não as instâncias de dados; e (iii) interligar dados de uma mesma instância de portal ou de portais distintos. Quando vocabulários são considerados, eles podem se restringir àqueles que objetivam a melhor descrição de metadados, ou àqueles que podem ampliar a semântica, ao utilizar vocabulários de domínios específicos. Todas as estratégias também podem ser aplicadas simultaneamente ou em separado.

3.1. Interligação sem considerar metadados

Este grupo de abordagens está voltado para a identificação e seleção de possíveis interligações de conjuntos de dados afins, publicados ou não em portais de dados abertos. Algumas pesquisas consideram a dificuldade de interligação de novos conjuntos de dados àqueles publicados. Outras pesquisas consideram a identificação de interligações, dado dois conjuntos de dados previamente definidos. Assim, a etapa de interligação pode fazer parte do processo de publicação.

O uso de cálculos estatísticos para análise de similaridades de recursos e de clusterização de conjuntos de dados pode gerar uma lista de opções de possíveis interligações que, usualmente, são ranqueados de acordo com a probabilidade de se definir ligações entre um novo conjunto de dados àqueles já publicados (LEME et al. 2013; CARABALLO et al., 2016).

Leme et al. (2013) utilizaram conjuntos de dados em RDF já publicados e a publicar. Os autores propuseram uma técnica de identificação de interligações utilizando uma classificação probabilística baseada na teoria Bayesiana. A técnica identifica padrões de interligações existentes em conjuntos de dados publicados para montar sugestões para o novo conjunto de dados. Feito isso, os conjuntos de dados que já estão publicados, e com interligações possivelmente correlatas ao novo conjunto de dados, são ranqueados de acordo com a probabilidade de se definir ligações entre suas URIs à URIs do novo conjunto. Assim, inspecionando-se os conjuntos de dados mais relevantes na classificação, encontra-se as interligações ao conjunto de dados a ser publicado.

Caraballo et al. (2016) também consideraram a dificuldade de interligação de novos conjuntos de dados àqueles publicados. Os autores propuseram uma abordagem que gera um grafo para descrever uma nuvem de dados conectados, utilizando metadados extraídos

de catálogos de dados. A abordagem define um processo automático de geração de comunidades (*clustering*) de conjuntos de dados do grafo gerado. Além disso, a abordagem gera descrições para essas comunidades aplicando técnicas de criação de *profiles* de conjunto de dados (*dataset profiling*). Cada *profile* é expresso como um vetor cujas dimensões correspondem à pontuação de relevância para as principais categorias da Wikipédia. Portanto, embora a abordagem utilize inicialmente metadados na geração do grafo de conjuntos de dados publicados, a definição da relevância da possível interligação entre o novo conjunto de dados àqueles já publicados é realizada com base nas categorias da Wikipédia, associadas a medidas de similaridade entre o perfil do novo conjunto de dados e o perfil dos conjuntos de dados já publicados, sem aplicar enriquecimento de metadados.

Diferentes tipos de sementes de buscas também são utilizadas para alcançar potenciais interligações de um novo conjunto de dados a outros já publicados. Elas são lançadas para iniciação de buscas e definições de caminhos a seguir em busca de possíveis interligações. Uma semente pode ser um conjunto de URLs de catálogos de dados de domínios semelhantes às necessidades da nova publicação (MARTINS; MOTA; CAVALCANTI, 2016); assim como um conceito de um vocabulário, ontologia ou tesauro (GOMES et al., 2014).

Martins, Mota e Cavalcanti (2016) propuseram uma abordagem para classificar e ranquear conjuntos de dados publicados, e modelados em RDF, de forma a identificar possíveis interligações para um novo conjunto de dados a ser publicado. A abordagem, denominada DSCrank, usa um conjunto de URLs de catálogos de dados como semente para coletar conjuntos de dados em RDF, publicados nestes catálogos, cujo domínio seja semelhante às necessidades da nova publicação.

Gomes et al. (2014) propuseram uma abordagem para dar suporte aos responsáveis pela publicação de dados em RDF nos processos de triplificação e ligação. Na proposta de Gomes et al. (2014) o publicador de dados indica um conceito de um vocabulário, ontologia ou tesauro como semente para que um mecanismo rastreador busque os termos relacionados no próprio vocabulário, ontologia ou tesauro e em outros aos quais ele tenha ligação já estabelecida.

O uso de sementes de buscas geralmente são combinadas com outras estratégias para efetivamente propor e interligar dados. Assim, as estratégias podem fazer uso de análises de frequência e/ou similaridade de expressões textuais em conjuntos de dados, assim como análises de relações entre elementos definidos em instrumentos terminológicos. Neste contexto, expressões textuais extraídas de rótulos de sujeitos, predicados e objetos de triplas em RDF podem ser utilizados como insumos para aplicação de métodos para indicar a importância de uma palavra em um documento com base na sua frequência de uso

(MARTINS; MOTA; CAVALCANTI, 2016; ACHICHI et al., 2019). Da mesma forma, rótulos de predicados de triplas em RDF também podem ser utilizados como insumos para aplicação de técnicas de cálculo de similaridade de textos e, assim, identificar as semelhanças entre dados (KETTOUCH; LUCA; HOBBS, 2018). Ademais, análises de relações entre elementos definidos em instrumentos terminológicos podem complementar estratégias de interligação após uma semente atingir um determinado ponto (GOMES et al., 2014).

A abordagem DSCrank de Martins, Mota e Cavalcanti (2016) realiza uma análise da relevância dos conjuntos de dados coletados a partir da semente de busca. Para isso, os conteúdos do novo conjunto de dados e daqueles selecionados como candidatos são comparados, fazendo uso de consultas SPARQL ou acessando diretamente as instâncias do arquivo RDF. Assim, a frequência e cobertura dos termos presentes nos conjuntos de dados é verificada por um algoritmo, para que seja estabelecido o ranqueamento.

Achichi et al. (2019) apresentaram um *framework* para interligar conjuntos de dados modelados em RDF previamente selecionados pelo interessado. Desses conjuntos de dados, pares de classes de instâncias são extraídas automaticamente, pré-processadas e preparadas para comparação. No pré-processamento dos conjuntos de dados é executada uma seleção das propriedades (predicados de triplas) que sejam relevantes para interligação. Após esse pré-processamento, é criado um perfil para cada instância de dados (*instance profile*) permitindo a representação de cada recurso como um subgrafo considerado relevante para comparação. A abordagem utiliza a técnica de CBD (*Concise Bounded Description*), onde um subgrafo é gerado para cada recurso de um grafo. De um modo geral, para gerar o subgrafo são extraídas todas as triplas onde o recurso é utilizado como sujeito, de forma descrever todas as características de um determinado recurso. Após a geração do *instance profile* para cada recurso de ambos os conjuntos de dados, elas são transformadas em documentos texto. Técnicas de processamento de linguagem natural são executadas para filtragem. Feito isso, os documentos são inseridos em um vetor de dimensão limitada para comparar os recursos. Nesse vetor, é aplicado um método para indicar a importância de uma palavra em um documento com base na sua frequência de uso (TF-IDF - *Term Frequency-Inverse Document Frequency*). Em seguida, um novo conjunto de dados de *links* é gerado, como resultado de correspondências de instâncias, desambiguação de instâncias e procedimentos de seleção de *link* (ou fusão de *link*). Assim, o resultado é materializado em um grafo de ligações do tipo “*owl:sameAs*” com referências para ambos os conjuntos de dados efetivamente ligados.

Kettouch, Luca e Hobbs (2018) desenvolveram um *framework* para apoiar motores de buscas de dados modelados em RDF ou semi-estruturados, tais como XML e JSON. Para isso, foi necessário o desenvolvimento de um subsistema para interligação de dados como

ferramenta complementar. A abordagem de interligação de dados proposta pelos autores tem como ponto central as propriedades dos recursos, isto é, os predicados de triplas descritos por meio de URIs. Os rótulos desses predicados são extraídos para que seja utilizada uma técnica de cálculo de similaridade semântica de textos e, assim, identificar as semelhanças entre dados. Para cálculo da similaridade é utilizada uma ferramenta que utiliza similaridade de palavras associadas ao WordNet, uma base de dados que contém relações semânticas entre palavras, incluindo sinônimos, hipônimos e merônimos. Alinhando-se com abordagens que utilizam sementes de busca, pode-se considerar que neste trabalho, os rótulos extraídos das instâncias de dados, a partir do qual as buscas por similaridades são efetuadas atuam como sementes.

Na abordagem de Gomes et al. (2014) é efetuada uma análise de elementos definidos em instrumentos terminológicos de forma a seguir relações previamente definidas nesses vocabulários, ontologias e tesouros, tais como verificação de subclasses, equivalências e termos relacionados. O rastreador proposto na abordagem gera uma lista dos vocabulários, ontologias e tesouros e dos conceitos/termos de vocabulários classificados como mais adequados para a triplificação, bem como uma lista de conjuntos de dados em RDF aos quais o novo conjunto de dados provavelmente pode estar vinculado. As listas geradas pelo rastreador são formadas por propriedades específicas, tais como subclasses, equivalências e termos relacionados. Sendo assim, as sugestões de novas ligações são obtidas por meio de metadados. Ademais, a semântica dos dados é considerada, visto que as interligações são identificadas com base nas classificações e conceituações das instâncias de dados. Portanto, a abordagem utiliza uma estratégia com dupla função: interligar o novo conjunto de dados pela sugestão de suportes ontológicos para descrever os novos dados, assim como pela sugestão de interligações com outros conjuntos de dados.

Embora algumas abordagens utilizem instâncias de dados RDF em suas estratégias, elas não consideram a princípio a perspectiva ontológica dos dados (MARTINS; MOTA; CAVALCANTI, 2016; KETTOUCH; LUCA; HOBBS, 2018; ACHICHI et al., 2019). Isso ocorre quando as estratégias extraem partes de URI para transformar em expressões textuais como insumo para aplicação de algum método de Processamento de Linguagem Natural ou algum método de cálculo da importância ou frequência de um termo em um documento. Por outro lado, quando a semântica dos dados é considerada, além de sugerir conjuntos de dados potenciais para interligação, é possível sugerir opções para anotar os recursos do conjunto de dados (GOMES et al., 2014).

O grupo de abordagens descrito nesta seção pode ser combinado com outras abordagens que considerem o uso de metadados, por exemplo. No entanto, excetuando-se a

abordagem de Gomes et al. (2014), quando utilizadas isoladamente, o nível semântico de interligação ainda é baixo por não considerar o significado dos dados.

3.2. Interligação considerando os metadados

O segundo grupo de abordagens de interligação visa interligar metadados associados a conjuntos de dados publicados em portais de dados abertos e, por consequência, interligar conjuntos de dados. Para serem utilizados como meio de interligações, os metadados precisam ser anotados de forma a estabelecer um protocolo comum para indicar semelhanças.

As estratégias que utilizam vocabulários, ontologias e tesouros para anotação semântica de metadados variam entre aquelas que utilizam vocabulários especializados na descrição mais expressiva da organização dos metadados, sem conotação com o significado dos dados, tais como DCAT, VoID e Schema.org, e aquelas que utilizam vocabulários de domínio específico para descrição mais expressiva dos valores atribuídos a elementos de metadados organizados.

Quando vocabulários especializados na descrição mais expressiva da organização dos metadados são utilizados, o resultado é um conjunto de metadados com algum nível semântico (MILIĆ; VELJKOVIĆ; STOIMENOV, 2015; PESCE et al., 2015; NEUMAIER; UMBRICH; POLLERES, 2017; MOREIRA et al., 2019; WILKINSON et al., 2017; KIRSTEIN et al., 2019). No entanto, o nível semântico é limitado a identificar como os metadados dos conjuntos de dados publicados foram organizados, permitindo processos de interligação ainda com pouca semântica, quando considerada a perspectiva dos significados dos dados.

Milić, Veljković e Stoimenov (2015) propuseram uma abordagem denominada *Linked Relations* (LIRE). A abordagem cria relações entre conjuntos de dados abertos por meio da utilização de um modelo RDF que utiliza o vocabulário VoID na definição das relações. O modelo RDF funciona como o elemento moderador das interligações. As interligações são estabelecidas entre dois conjuntos de dados, classificando as relações como "filho_de" (CHILD_OF) ou "conexões_a_partir_de" (LINKS_FROM). Para isso, a abordagem analisa os elementos de metadados associados aos conjuntos de dados na instância de portal. Como exemplo, é analisada a relação do tipo "filho_de", que verifica se o conjunto de dados candidato a filho possui o número de etiquetas similares (ou exatamente as mesmas) menor do que o número de etiquetas similares (ou exatamente as mesmas) do outro conjunto de dados analisado. A abordagem possui um módulo que apresenta cada conjunto de dados como um elemento gráfico e suas ligações na própria plataforma de portal utilizada como

exemplo, o CKAN. Embora haja uma representação gráfica, não se trata de uma estrutura grafo direcionado, com nós e arcos nomeados.

Neumaier, Umbrich e Polleres (2017) desenvolveram uma abordagem na qual, dentre outras coisas, extraem metadados de vários portais de dados abertos construídos sobre o CKAN, Socrata e OpenDataSoft, expõem e conectam estes metadados, mapeando-os e publicando-os utilizando vocabulários para descrições de metadados, como DCAT e Schema.org. A ontologia PROV³⁷ também é utilizada para registrar e anotar a proveniência dos dados gerados/publicados. Os resultados, com todos os metadados integrados e anotados, foram disponibilizados publicamente como dados conectados em portal que monitora os demais portais, que é o elemento moderador. Adicionalmente, um SPARQL *Endpoint* é disponibilizado para consultar o RDF gerado.

Kirstein et al. (2019) apresentaram a conceituação, a implementação e as lições aprendidas na implementação do *European Data Portal*³⁸ (Portal de Dados Europeu), lançado em 2015 pela Comissão Europeia. O portal funciona como um ponto central único para obtenção de informações do setor público de toda União Europeia. A abordagem efetua a coleta de metadados disponibilizados em portais de dados abertos do setor público de países da União Europeia, os transforma, os anota e os armazena como dados conectados. Os metadados são anotados com a especificação DCAT-AP (*DCAT Application Profile for data portals in Europe*), uma especificação baseada no vocabulário DCAT para atender às necessidades específicas de portais de dados na Europa, enquanto fornece interoperabilidade semântica com outros portais com base na reutilização de vocabulários controlados estabelecidos e mapeamentos para vocabulários de metadados existentes (DRAGAN, 2020).

Alguns dos vocabulários especializados na descrição mais expressiva da organização dos metadados disponibilizam elementos para anotação dos assuntos contidos nos conjuntos de dados ou de etiquetas associadas aos conjuntos de dados. No entanto, o nível de granularidade ainda é alto, pois assuntos e etiquetas referem-se ao conjunto de dados como um todo. No caso específico das etiquetas, outra questão emerge: elas representam apenas um texto simples.

Diante do importante papel das etiquetas para os motores de busca das plataformas de dados abertos, algumas abordagens focam no enriquecimento semântico delas, em separado dos outros elementos de metadados (TYGEL, 2016; CASTRO, 2019). Neste sentido, são utilizados vocabulários, ontologias e tesouros de domínios específicos para descrever de forma mais expressiva os valores atribuídos a elementos de metadados organizados. Assim, eles possibilitam interligações com mais semântica, quando considerada a perspectiva dos

³⁷ <https://www.w3.org/TR/prov-o/>

³⁸ <https://www.europeandataportal.eu/en>

significados dos dados. Vocabulários, ontologias e tesouros de domínio específico elevam expressões textuais simples a expressões com significado bem definido.

Tygel (2016) propôs uma abordagem para extrair, limpar e enriquecer semanticamente etiquetas atribuídas pelos publicadores a conjuntos de dados nos portais de dados abertos. O objetivo é que conjuntos de dados de diferentes portais, eventualmente em idiomas diferentes, possam ser conectados se seus conteúdos se referirem aos mesmos assuntos. Usando a abordagem denominada STODaP (*Semantic Tags for Open Data*), os usuários podem encontrar dados sobre tópicos específicos de várias fontes em um único local. A abordagem STODaP gera metadados semânticos na forma de etiquetas, que são organizadas como recursos no formato RDF/XML. As etiquetas são publicadas em uma espécie de portal de metadados (o componente moderador) que viabiliza as interligações entre conjuntos de dados de diferentes portais por meio de etiquetas em comum, mesmo que anotadas com diferentes vocabulários, ontologias ou tesouros.

Castro (2019) também considerou recursos da Web Semântica para enriquecer metadados ao fornecer uma abordagem para criar etiquetas com base na semântica dos dados, de modo a melhorar a recuperação de conjuntos de dados publicados em portais de dados abertos. Diferente de Tygel (2016), Castro (2019) não utilizou as etiquetas definidas pelos publicadores, visto que sua abordagem gera as próprias etiquetas a partir das instâncias dos conjuntos de dados publicados em formatos tabulares. A abordagem de Castro (2019), denominada RTagE (*Relevant Tag Extractor*), busca e extrai novas etiquetas a partir das instâncias de dados, aplicando técnicas de extração de palavras-chave de documentos. Castro (2019) busca, em uma base de conhecimento, conceituações na forma de recursos (com URI) para anotação das palavras-chave extraídas. Com a conceituação formal, são geradas etiquetas semânticas. Ou seja, como o nome já sugere, etiquetas semânticas consideram a semântica embutida nas instâncias de dados. As etiquetas semânticas são associadas aos conjuntos de dados por meio da geração de um grafo em RDF que conecta conjuntos de dados e etiquetas. Sendo assim, o grafo RDF gerado pela abordagem RTagE interliga conjuntos de dados publicados em uma mesma instância de portal e que utilizem o mesmo vocabulário, ontologia ou tesouro. O grafo RDF é o elemento moderador das interligações e pode ser explorado para encontrar conjuntos de dados semanticamente relacionados por meio de suas etiquetas. Contudo, a abordagem RTagE não publica o grafo resultante no portal, nem o conecta a ele.

Pesce et al. (2015) também extraíram e anotaram os metadados associados a conjuntos de dados distribuídos em variados formatos. Parte dos metadados foi anotado com vocabulários especializados para anotação semântica de metadados descritivos (DCAT, VOID, DOAP, schema.org) integrando esses vocabulários quando necessário; outra parte foi

anotada com vocabulários, ontologias e tesouros de domínios específicos (FAO *Geopolitical Ontology*, AGROVOC) para identificar tópicos, escopo geográfico e tipos de dados. Utilizando os metadados enriquecidos, a abordagem cria um catálogo, sem utilizar plataformas de dados abertos, como ponto central para buscas de conjuntos de dados de um domínio específico. A estratégia de busca permite identificar as interligações entre os conjuntos de dados publicados visto que é possível verificar quais conjuntos de dados utilizam os mesmos metadados.

Face ao exposto, a anotação de metadados utilizando vocabulários especializados para anotação de metadados descritivos, associada a vocabulários, ontologias e tesouros de domínios específicos, leva a descrições mais ricas, fornecendo suporte a processos de interligação semântica por meio de metadados, isto é, interligações semânticas indiretas. No entanto, esta combinação mantém os metadados no nível de conjunto de dados ou nível de distribuição de dados.

Ademais, quando as instâncias de dados são consideradas para extrair algum nível de metadados, mesmo que focada em apenas um elemento de metadado, o enriquecimento semântico é maior, como proposto em Castro (2019). Esse tipo de extração de metadados permite conhecer o conteúdo do conjunto de dados sem a necessidade de acessar as instâncias de dados.

As abordagens apresentadas têm gerado resultados processáveis por máquina. Porém, não se observa exploração do potencial de dados conectados pelas abordagens, caso existam conjuntos de dados em RDF publicados nos portais. As abordagens que interligam dados publicados em portais distintos fornecem interfaces externas aos portais para apresentação dos resultados das interligações (TYGEL, 2015; PESCE et al., 2015; NEUMAIER; UMBRICH; POLLERES, 2017, KIRSTEIN et al. 2019). Essas abordagens propiciam aos consumidores uma camada para entendimento das interligações, mas a busca, consulta e navegação dos dados conectados gerados, ou não estão disponíveis, ou destinam-se apenas aos especialistas nas tecnologias da Web Semântica. As abordagens que interligam dados publicados no mesmo portal (MILIĆ; VELJKOVIĆ; STOIMENOV, 2015; CASTRO, 2019) também não fornecem interfaces que explorem a capacidade de navegação de dados conectados. Assim, as interligações só podem ser exploradas em ferramentas externas ao ambiente do portal (CASTRO, 2019).

Ainda com ênfase na interligação de metadados, uma linha de pesquisa que tem se destacado nos últimos anos envolve o estabelecimento de diretrizes e o desenvolvimento de soluções para lidar com dados FAIR. Já existe iniciativa de estabelecimento de uma arquitetura a ser utilizada como uma “referência de implementação” para os princípios FAIR, em particular quando aplicado a dados que normalmente não seriam considerados

interoperáveis, tais como uma planilha Excel ou um arquivo texto. O foco está na modelagem RDF como base da infraestrutura, já que modelos em RDF foram projetados para compartilhamento de dados na Web (WILKINSON et al., 2017). Nesse contexto, Moreira et al. (2019) descrevem a arquitetura, a metodologia e a implementação de um proxy semântico, isto é, um mecanismo capaz de traduzir metadados recuperados de repositórios de dados para o modelo de dados conectados de acordo com as especificações de um FAIR *Data Point*.

3.3. Sumarização das estratégias de interligação de dados

De um modo geral, a evolução e as diferentes estratégias descritas nas seções 3.1 e 3.2 alinham-se aos princípios da Web de Dados. No entanto, as estratégias descritas ainda deixam algumas lacunas face ao potencial de uso de dados conectados.

O Quadro 4 sumariza as estratégias das abordagens de interligação propostas nos trabalhos de pesquisa comentados na seção 3.1, ou seja, que não consideram os metadados associados aos conjuntos de dados na interligação.

Quadro 4: Estratégias das abordagens de interligação que não consideram os metadados associados aos conjuntos de dados

		Trabalhos de Pesquisa					
		Leme et al. (2013)	Gomes et al. (2014)	Caraballo et al. (2016)	Martins, Mota e Cavalcanti (2016)	Kettouch, Luca e Hobbs (2018)	Achichi et al. (2019)
Estratégias de Interligação sem considerar metadados	Cálculos Estatísticos	X	-	-	-	-	X
	Clusterização de dados	-	-	X	-	-	-
	Sementes de Buscas	-	X	-	X	X	-
	Similaridade de expressões textos	-	-	-	-	X	-
	Frequência de expressões textuais	-	-	-	X	-	X
	Relações entre elementos definidos em instrumentos terminológicos	-	X	-	-	-	-

O Quadro 5 apresenta as estratégias das abordagens de interligação propostas nos trabalhos de pesquisa comentados na seção 3.2, ou seja, que consideram os metadados associados aos conjuntos de dados na interligação.

Quadro 5: Estratégias das abordagens de interligação que consideram os metadados associados aos conjuntos de dados

			Trabalhos de Pesquisa							
			Milić, Veljković e Stoimenov (2015)	Tygel (2015)	Pesce et al. (2015)	Neumaier, Umbrich e Polleres (2017)	Wilkinson et al. (2017)	Castro (2019)	Moreira et al. (2019)	Kirstein et al. (2019)
Estratégias de Interligação considerando os metadados	Suporte Ontológico	Especializados na descrição da organização dos metadados	X	-	X	X	X	-	X	X
		De domínio específico para valores de elementos de metadados	-	X	X	-	-	X	-	-
	Instâncias de Dados	Considera a semântica	-	-	-	-	-	X	-	-
		Não considera a semântica	-	-	X	-	X	-	-	X
	Alcance da interligação	Mesma instância de portal	X	-	-	-	-	X	-	-
		Diferentes instâncias de portais	-	X	X	X	X	-	X	X

Abordagens de interligação que não consideram os metadados associados aos conjuntos de dados servem de apoio no processo de preparação de novos conjuntos de dados. Contudo, para aqueles já publicados, cujas responsabilidades são das organizações ou grupos que os publicaram, a interligação semântica indireta, isto é, que considera os metadados associados aos conjuntos de dados, parece ser o caminho natural. Como ela envolve o uso de metadados anotados, que são de uso público e irrestrito, não altera as publicações originais, ao mesmo tempo que fornece uma camada de interligação.

Especialmente quando se trata de conjuntos de dados modelados em RDF de acesso aberto, verifica-se que ainda há lacunas a serem exploradas. As instâncias de dados em RDF, que naturalmente contêm metadados, ainda podem ser mais exploradas complementando as abordagens descritas. Esta lacuna é base para abordagem proposta nesta dissertação e

explorada no próximo capítulo, chamada LigADOS. A proposta considera estabelecer interligações semânticas indiretas de conjuntos de dados em RDF, definindo e expondo ao público em geral as interconexões de metadados no nível de registro de dados, investigando instâncias de dados e anotações semânticas. Essa granularidade de metadados associada a esquemas de metadados propicia o ambiente para interligações entre dados publicados em diferentes portais.

4. ABORDAGEM LigADOS PARA INTERLIGAÇÃO DE CONJUNTOS DE DADOS

A Web de Dados e tecnologias associadas proporcionam uma série de benefícios para compreensão do significado dos dados. A abordagem LigADOS faz uso das tecnologias utilizadas na Web de Dados para interligar conjuntos de dados publicados em instâncias de portais de dados abertos, gerar metadados semânticos, assim como propiciar suporte ao processamento por máquina. O foco de LigADOS está na interligação de conjuntos de dados já alinhados aos conceitos da Web de Dados, isto é, conjuntos de dados modelados em RDF em quaisquer tipos de serializações.

A interligação entre os conjuntos de dados é estabelecida considerando as classificações semânticas das instâncias de dados, isto é, extraíndo categorizações das anotações semânticas embutidas nelas. As classificações semânticas das instâncias de dados formam os metadados semânticos, alinhados com os princípios FAIR, por estarem definidos em vocabulários que utilizam identificadores únicos para cada classe/conceito definido. Assim, eles servem de insumos para compor esquemas de metadados propostos por FDPs, ao mesmo tempo que expandem as possibilidades de interligações de conjuntos e dados publicados em diferentes portais. A Figura 13 apresenta o esquema de uma interligação semântica indireta gerada por LigADOS, onde a Classe_1 do vocabulário vocab_V é o ponto de interligação das distribuições “Dist_A_1” do “Conjunto_de_Dados_A” e “Dist_B_3” do “Conjunto_de_Dados_B”. LigADOS extrai termos utilizados para anotação semântica das triplas das distribuições “Dist_A_1” e “Dist_B_3” e os utiliza para obter a classe do recurso anotado. Assim, a classe proporciona uma forma indireta de interligar as distribuições de dados e, por consequência, os conjuntos de dados de uma mesma instância de portal.

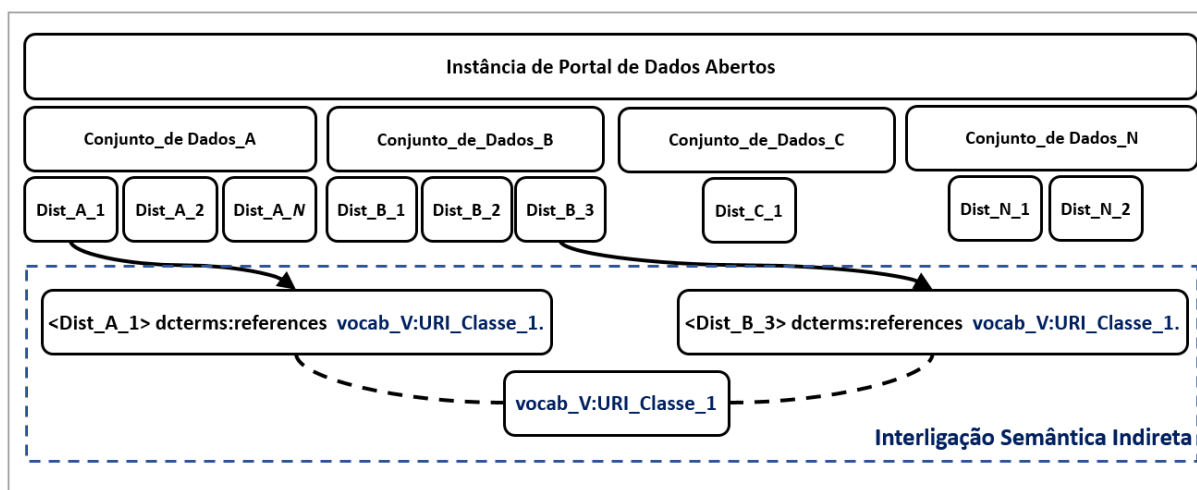


Figura 13: Esquema de interligação semântica indireta gerada por LigADOS para uma instância de portal

LigADOS também propõe que o ambiente do próprio portal de dados deve ser utilizado para expor as interligações entre os conjuntos de dados publicados. Mais que isso, a abordagem considera que estas interligações devem estar disponíveis ao público em geral, mesmo àqueles sem conhecimentos prévios sobre os conceitos da Web Semântica ou da Web de Dados. Um aspecto importante é que os conjuntos de dados envolvidos nas interligações possam ser identificados e acessados a partir da visualização das interligações. Como LigADOS mantém-se alinhada com a necessidade de manter as interligações entre conjuntos de dados processáveis por máquina, são disponibilizados mecanismos tecnológicos para esse suporte. Cabe ressaltar que LigADOS não tem a intenção de alterar as publicações originais que são de responsabilidade das diversas organizações que publicam seus dados de forma aberta.

A abordagem LigADOS é composta de uma série de atividades organizadas sistematicamente em dois processos. O primeiro processo tem como foco a interligação dos conjuntos de dados, com base em metadados semânticos associados aos conjuntos de dados publicados, e na disponibilização das interligações geradas. O segundo processo tem como foco o suporte à anotação semântica de conjuntos de dados modelados em RDF. Ambos os processos são disparados por um agente técnico de forma transparente ao funcionamento da instância de portal. O agente técnico pode ser um recurso tecnológico programado para este fim ou um ser humano. A partir da chamada para execução, os processos são executados sem interação de agentes externos. O resultado da execução do processo, por sua vez, destina-se aos consumidores de dados do portal (usuários), assim como às aplicações aptas a processar dados conectados.

As seções 4.1 e 4.2 detalham os dois processos da abordagem LigADOS. Para representar os processos da abordagem foi utilizado *Business Process Model and Notation* (BPMN), padronizado pela OMG (2011). A arquitetura de suporte para execução dos processos é apresentada na seção 4.3.

4.1. Processo de Interligação de Conjuntos de Dados em Plataformas de Dados Abertos

Este processo engloba o fluxo de atividades principal da abordagem proposta. O cerne desse processo é a identificação das interligações entre os conjuntos de dados em RDF, em quaisquer tipos de serializações, publicados em um portal de dados. A partir das interligações encontradas é construído um grafo em RDF a ser utilizado pelos consumidores de dados do portal. Portanto, cada instância do portal de dados terá somente um grafo associado. O grafo

gerado é denominado *Grafo de Interligações Semânticas*. Conforme apresentado na Figura 14, o processo de interligações foi dividido em cinco macroprocessos. As subseções a seguir apresentam as atividades de cada macroprocesso.

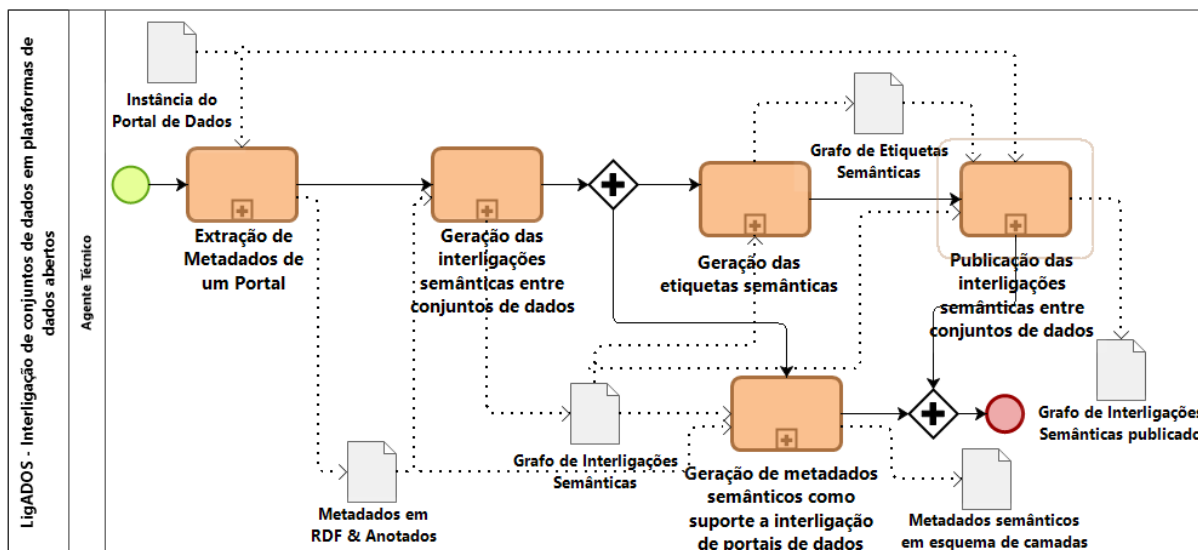


Figura 14: Processo de interligação e disponibilização das interligações de dados de LigADOS

4.1.1. Extração de metadados de um portal

O fluxo de atividades para extração de metadados cadastrados em um determinado portal inicia-se com a varredura de toda a instância do portal com o intuito de extrair os metadados de todos os conjuntos de dados publicados com pelo menos uma distribuição modelada em RDF. A extração gera um novo conjunto de dados modelado em RDF, contendo os metadados dos conjuntos de dados supramencionados. O conjunto de dados de metadados é anotado com o vocabulário DCAT e persistido em um repositório de um *triplestore*. A anotação com o vocabulário DCAT está alinhada com outros trabalhos (PESCE et al., 2015; NEUMAIER, UMBRICH, POLLERES, 2017; MOREIRA et al., 2019), mas que, principalmente, alinha-se com a recomendação de Wilkinson et al. (2017) por permitir a identificação das distribuições de conjuntos de dados. O conjunto de metadados é o insumo para iniciar as interligações. Do conteúdo do conjunto de metadados, são extraídas as seguintes informações:

- a identificação e o título da instância do portal;
- os identificadores e os títulos dos conjuntos de dados;
- os nomes das organizações que publicaram os conjuntos de dados; e
- os identificadores, títulos e URL para *download* das distribuições em RDF que representam os conjuntos de dados.

4.1.2. Geração das interligações semânticas entre conjuntos de dados de um portal

As atividades para geração das interligações semânticas utilizam o conjunto de metadados extraído do portal. Dentre as informações coletadas e armazenadas no conjunto de metadados, constam os URL das distribuições associadas aos conjuntos de dados, isto é, o URL da localização do arquivo de dados em RDF. Utilizando esses URL, são obtidos os dados a serem carregados em um repositório de um *triplestore*. Cada distribuição de conjunto de dados é carregada em um grafo distinto no repositório. Os grafos são lidos individualmente com o intuito de identificar os recursos sujeitos e os objetos de triplas que sejam URI e que estejam anotados com algum vocabulário, ontologia ou tesauro.

Os recursos são utilizados como insumos para o disparo de consultas federadas aos SPARQL *Endpoints* dos vocabulários, ontologias e tesauros de modo a encontrar os URI dos termos e/ou elementos utilizados para anotar as triplas da distribuição de um conjunto de dados. Localizado o URI dos termos e/ou elementos no vocabulário, ontologia ou tesauro, a mesma consulta federada busca pela classe ou conceito que categoriza o termo e/ou elemento utilizado. A identificação da classe ou conceito considera o elemento um nível acima do que foi utilizado para anotação do recurso, buscando por relações do tipo `skos:broader`³⁹, `rdf:type`⁴⁰, `rdfs:subClassOf`⁴¹. Identificada a classe ou conceito, a consulta ainda extrai seu nome buscando por relações do tipo `skos:prefLabel`⁴² ou `rdfs:label`⁴³, que estejam preferencialmente descritas em português. Após a identificação e posterior extração das classes e/ou conceitos atinentes a um grafo, que representa uma distribuição de dados, o grafo pode ser removido do repositório.

A Figura 15 apresenta um esquema de identificação de um conceito do Tesauro AGROVOC⁴⁴ que categoriza dois termos distintos, utilizados por duas distribuições de conjuntos de dados distintas: “Dist_A_1” e “Dist_B_3”. A “Dist_A_1” utiliza o conceito “Cipermetrina”; a “Dist_B_3” utiliza o conceito “Fluvalinato”; e ambas têm “Piretrina Sintética” como conceito mais amplo (`skos:broader`).

³⁹ `skos:broader` refere-se a propriedade “broader” do vocabulário SKOS, cujo *namespace* recomendado pela W3C é <http://www.w3.org/2004/02/skos/core#>

⁴⁰ `rdf:type` refere-se a propriedades “type” definida na sintaxe abstrata do vocabulário RDF, cujo *namespace* recomendado pela W3C é <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.

⁴¹ `rdfs:subClassOf` refere-se a propriedade “subClassOf” do vocabulário RDFS Schema, cujo *namespace* recomendado pela W3C é <http://www.w3.org/2000/01/rdf-schema#>

⁴² `skos:prefLabel` refere-se a propriedade “prefLabel” do vocabulário SKOS, cujo *namespace* recomendado pela W3C é <http://www.w3.org/2004/02/skos/core#>

⁴³ `rdfs:label` refere-se a propriedade “label” do vocabulário RDFS Schema, cujo *namespace* recomendado pela W3C é <http://www.w3.org/2000/01/rdf-schema#>

⁴⁴ <http://www.fao.org/agrovoc/>

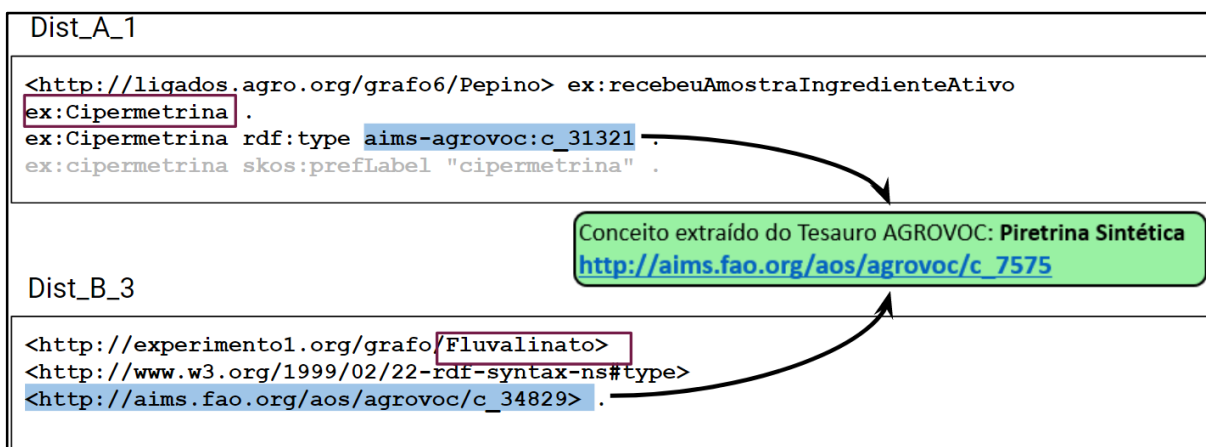


Figura 15: Identificação de conceito que categoriza termos utilizados por duas distribuições de conjuntos de dados

As classes e/ou conceitos são utilizados como insumos para montagem das triplas que farão parte do *Grafo de Interligações Semânticas*, que é a saída deste fluxo de atividades. As triplas montadas representam metadados no nível de registro de dados, propiciando a configuração do nível 5, quando comparado a uma estrutura de FDP (vide seção 2.4.2). A Figura 16 apresenta uma representação dos metadados no nível de registro de dados.



Figura 16: Esquema de metadados no nível de registro de dados

O *Grafo de Interligações Semânticas* é formado por nós que representam:

- a instância do portal de dados abertos;
- as organizações que publicaram o conjunto de dados;
- os conjuntos de dados;
- as distribuições de dados modelados em RDF dos conjuntos de dados;
- os vocabulários, ontologias e tesauros referenciados nas distribuições dos conjuntos de dados; e
- as classes e/ou conceitos dos vocabulários, ontologias e tesauros que classificam os termos e/ou elementos utilizados para anotar as triplas das distribuições dos conjuntos de dados.

Para definir nós, seja como sujeito, seja como objeto, bem como estabelecer as ligações entre os nós é preciso fazer uso de vocabulários especializados. Para o *Grafo de Interligações Semânticas* os seguintes vocabulários, classes e propriedades neles definidas, são utilizados:

- Friend of A Friend (FOAF)
 - <http://xmlns.com/foaf/0.1/Organization> → anota os recursos que representam as organizações que publicaram os conjuntos de dados.
 - <http://xmlns.com/foaf/0.1/name> → associa um literal que representa um nome de uma organização ao nó anotado com foaf:Organization.

A Figura 17 apresenta a utilização do vocabulário pela abordagem de solução que utiliza uma classe e uma propriedade.

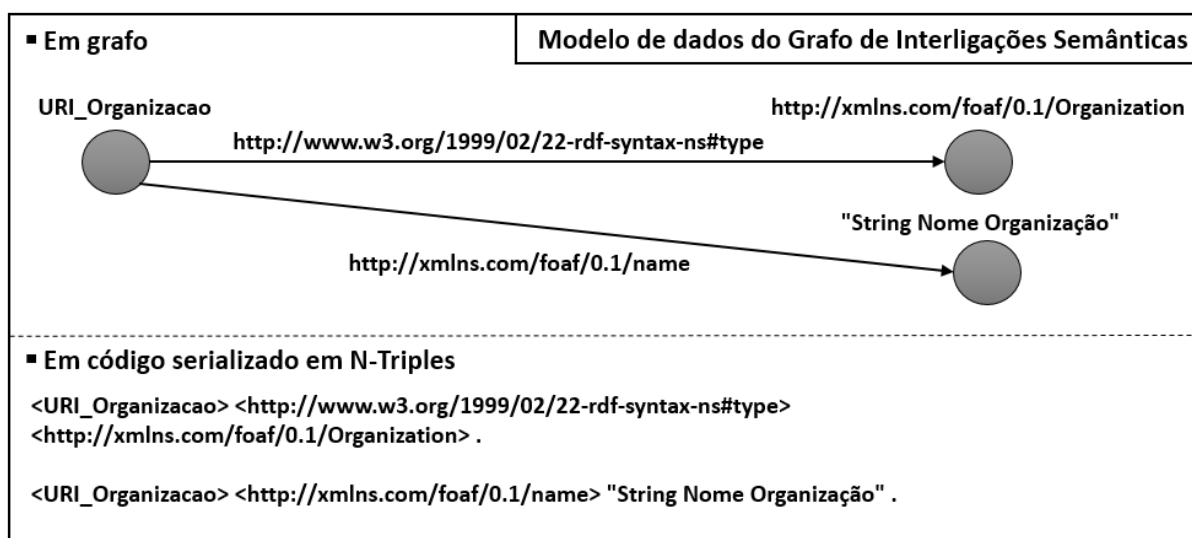


Figura 17: Utilização do vocabulário FOAF

- Data Catalog Vocabulary (DCAT)
 - <http://www.w3.org/ns/dcat#Dataset> → anota os recursos que representam os conjuntos de dados publicados.
 - <http://www.w3.org/ns/dcat#Distribution> → anota os recursos que representam as distribuições modeladas em RDF para cada conjunto de dados.
 - <http://www.w3.org/ns/dcat#distribution> → liga uma ou mais distribuições ao conjunto de dados.

- <http://www.w3.org/ns/dcat#theme> → indica os vocabulários, ontologias ou tesouros utilizados pelas distribuições de conjuntos de dados.

A Figura 18 apresenta o uso do vocabulário na construção do *Grafo de Interligações Semânticas* no qual são utilizadas duas classes e duas propriedades.

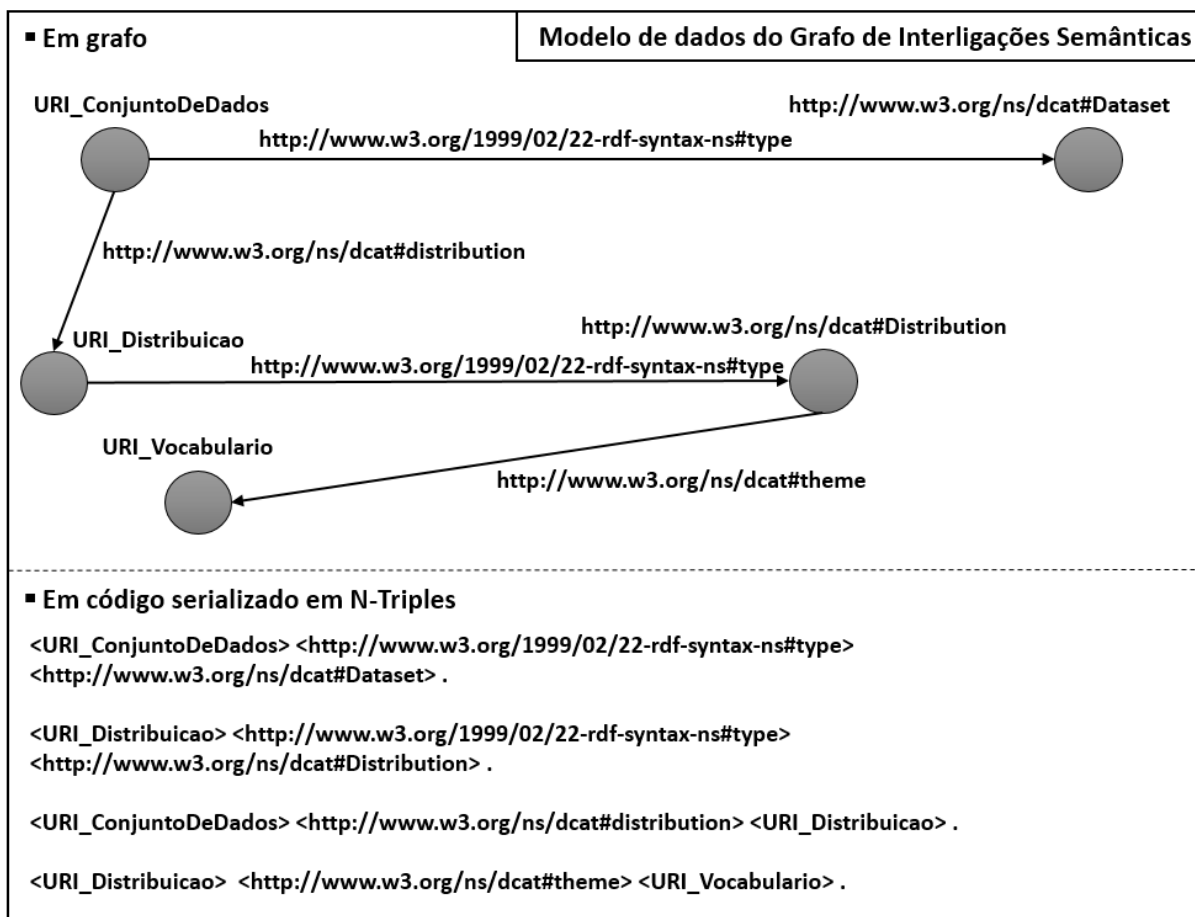


Figura 18: Utilização do vocabulário DCAT

- DCMI Metadata Terms
 - <http://purl.org/dc/terms/publisher> → liga o conjunto de dados à organização que o publicou no portal. Assim, a propriedade está relacionada com a definição do recurso do tipo “Organization” anotado com o vocabulário FOAF.
 - <http://purl.org/dc/terms/references> → indica a classe/conceito do vocabulário, ontologia ou tesouro que é referenciada indiretamente pela distribuição do conjunto de dados, isto é, que categoriza termos/elementos de uma distribuição de dados.

- <http://purl.org/dc/dcam/memberOf> → especifica que uma classe/conceito de um vocabulário utilizado por uma distribuição do conjunto de dados faz parte do esquema de um vocabulário, ontologia ou tesouro.

A Figura 19 apresenta o uso do vocabulário pelo *Grafo de Interligações Semânticas* no qual foram utilizados termos de dois *namespaces* distintos.

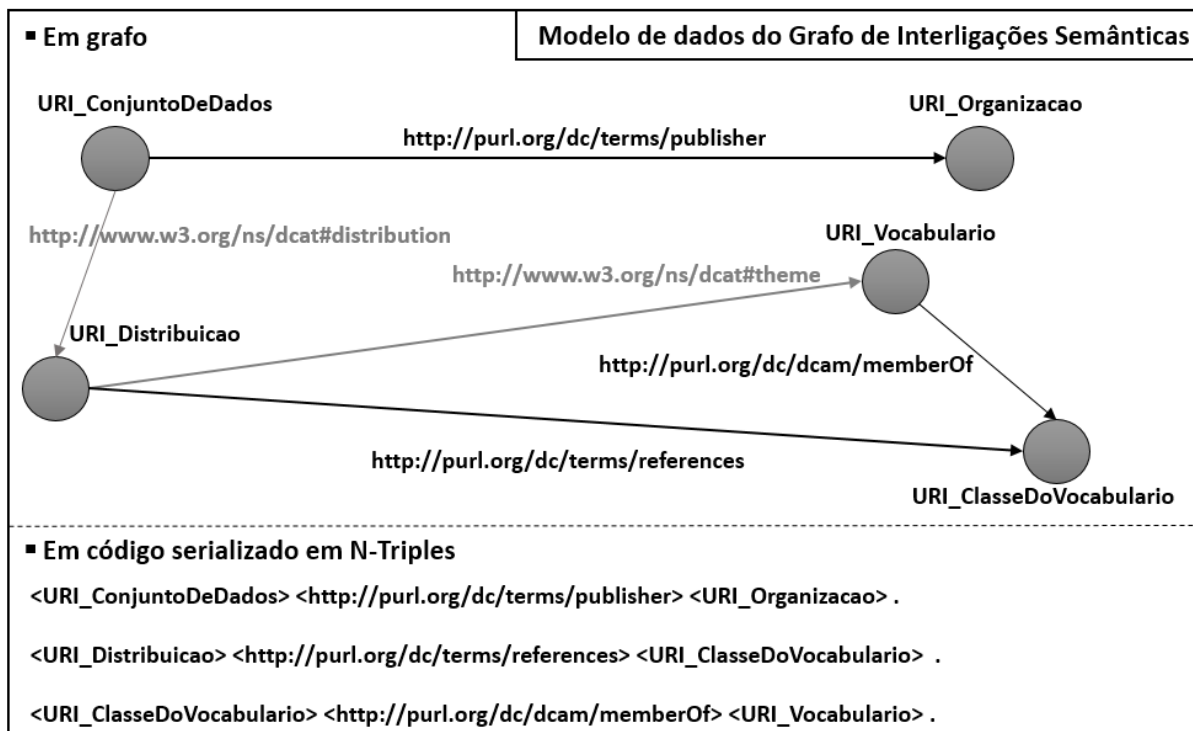


Figura 19: Utilização do vocabulário DCMI Metadata Terms

- Vocabulary of A Friend (VOAF)
 - <http://purl.org/vocommons/voaf#Vocabulary> → anota os recursos que representam vocabulários, ontologias e tesouros.

A Figura 20 apresenta a utilização do vocabulário no *Grafo de Interligações Semânticas* que emprega uma única classe.

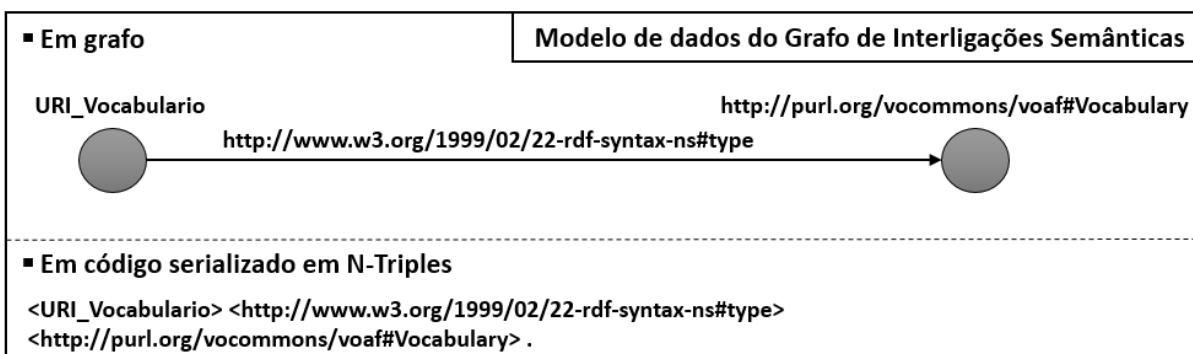


Figura 20: Utilização do vocabulário VOAF

- VANN
 - <http://purl.org/vocab/vann/preferredNamespacePrefix> → indica prefixos textuais e humanamente legíveis para os vocabulários, ontologias e tesouros utilizados por uma distribuição de um conjunto de dados.

A Figura 21 apresenta a propriedade de VANN utilizada pela abordagem de solução.

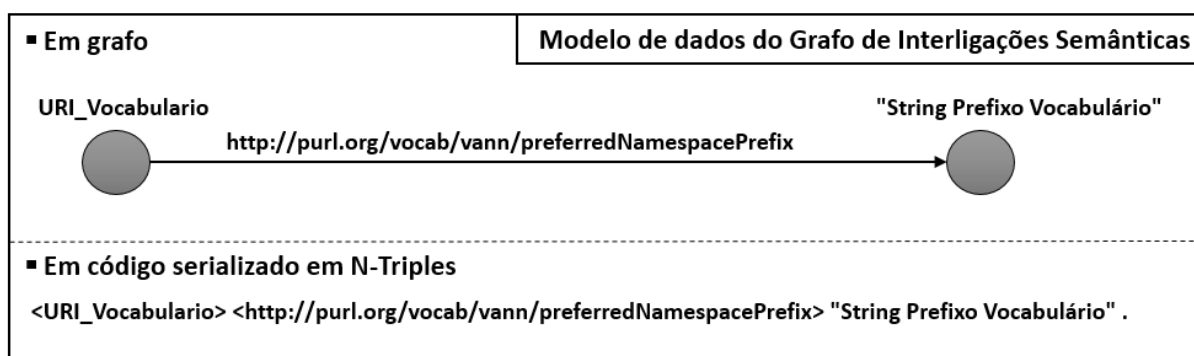


Figura 21: Utilização do vocabulário VANN

Além dos vocabulários específicos descritos nesta seção, foi utilizado RDFS, discutido na seção 2.3. Cabe ressaltar que quaisquer vocabulários podem ser utilizados pelos conjuntos de dados do portal. Fato que não interfere na geração do grafo pela abordagem LigADOS.

Para a geração do *Grafo de Interligações Semânticas* são extraídas somente as classes/conceitos de vocabulários, ontologias e tesouros que classificam os termos e/ou elementos utilizados para anotar sujeitos e objetos de triplas. Predicados (propriedades) não são extraídas pela abordagem pois sozinhos não representam conceitos. De fato, os predicados corroboram com a estruturação de um conceito, caracterizando-o, como na criação de um *Concise Bounded Description*, onde um subgrafo é gerado para cada recurso de um grafo (ACHICHI et al., 2019).

Ao se concentrar na extração dos conceitos e classificações mais gerais dos termos e/ou elementos, LigADOS extrai descritores das instâncias de dados. Isto significa dizer que a abordagem gera metadados no nível de registro de dados, alinhando-se ao nível semântico do esquema de metadados em camadas proposto para um FDP.

Neste contexto, reforça-se o importante papel do uso das ontologias como modelos de referência de domínios específicos nas atividades de anotação semântica. Ontologias assumem a posição de habilitadoras de abordagens como LigADOS, que consideram a

semântica das instâncias de dados como fator-chave para identificação de relações entre dados.

Conjuntos de dados podem utilizar os mesmos ou diferentes vocabulários, ontologias e tesouros. Todavia, pode ser que classes/conceitos de diferentes vocabulários, ontologias e tesouros representem o mesmo conceito. Em função disso, alguns vocabulários, ontologias e tesouros já estabelecem correspondências entre si. Aproveitando-se das correspondências já definidas nos vocabulários, ontologias e tesouros, LigADOS insere no *Grafo de Interligações Semânticas* as correspondências entre as classes/conceitos dos vocabulários, ontologias e tesouros utilizados na instância de portal e as classes da ontologia DBPedia⁴⁵. As correspondências são obtidas recuperando-se as correspondências descritas por meio da propriedade “`skos:closeMatch`”, do vocabulário SKOS, comentado na seção 2.3.

A Figura 22 apresenta a estrutura das triplas que compõem o *Grafo de Interligações Semânticas* serializado em N-Triples. A estrutura é aplicada para cada conjunto de dados representado no grafo publicado no portal.

O *Grafo de Interligações Semânticas* é persistido em um repositório de um *triplestore*. A primeira geração do grafo é uma etapa fundamental da abordagem. Novos conjuntos de dados adicionados ou atualizações dos dados já publicados podem ser monitorados por meio de configurações de soluções tecnológicas.

Nesse ponto é importante ressaltar que, em suma, o *Grafo de Interligações Semânticas* é um de grafo de metadados. Metadados enquadrados em diferentes níveis, quando comparado aos níveis de organização de metadados de um FDP, e aglutinados em um mesmo grafo, cujo diferencial se dá pelos descritores semânticos das instâncias de dados. O *Grafo de Interligações Semânticas* tem uma dupla função: (i) indicar as interligações semânticas indiretas entre conjuntos de dados; e (ii) gerar metadados semânticos mais granulares do que aqueles originalmente associados aos conjuntos de dados.

⁴⁵ <https://wiki.dbpedia.org/services-resources/ontology>

```

#Conjunto de dados e seu publicador
<URI_Organizacao> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://xmlns.com/foaf/0.1/Organization> .

<URI_Organizacao> <http://xmlns.com/foaf/0.1/name> "String Nome Organização"@pt .

<URI_ConjuntoDeDados> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/ns/dcat#Dataset> .

<URI_ConjuntoDeDados> <http://purl.org/dc/terms/publisher> <URI_Organizacao> .

#Distribuições do conjunto de dados
# [ Repete para cada distribuição ]
<URI_Distribuicao> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/ns/dcat#Distribution> .
<URI_Distribuicao> <http://purl.org/dc/terms/title> "String título da
distribuição" .
<URI_ConjuntoDeDados> <http://www.w3.org/ns/dcat#distribution> <URI_Distribuicao>
.

#Vocabulário(s) utilizado(s) pela Distribuição
<URI_Distribuicao> <http://purl.org/dc/terms/references> <URI_Vocabulario> .

<URI_Vocabulario> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://purl.org/vocommons/voaf#Vocabulary> .

<URI_Vocabulario> <http://purl.org/vocab/vann/preferredNamespacePrefix> "String
Prefixo Vocabulário" .

#Classe(s) do(s) vocabulário(s) que foram utilizado(s) pela Distribuição
# [ Uma tripla para cada classe/conceito do vocabulário ]
<URI_Distribuicao> <http://purl.org/dc/terms/references> <URI_ClasseDoVocabulario>
.

#Características de classe/conceito do vocabulário
<URI_ClasseDoVocabulario> <http://purl.org/dc/dcam/memberOf> <URI_Vocabulario> .

<URI_ClasseDoVocabulario> <http://www.w3.org/2000/01/rdf-schema#label> "String
NomeClasse"@pt .

#Correspondências de conceitos de classe(s) de algumas ontologias com DBPedia
<URI_ClasseDoVocabulario> <http://www.w3.org/2004/02/skos/core#closeMatch>
<http://dbpedia.org/ontology/[classe]> .

<http://dbpedia.org/ontology/[classe]> <http://www.w3.org/2000/01/rdf-
schema#label> "String NomeClasseDBPedia"@pt .

<http://dbpedia.org/ontology/[classe]> <http://purl.org/dc/dcam/memberOf>
<http://dbpedia.org/ontology/> .

#Temas dos conjuntos dados
<URI_ConjuntoDeDados> <http://www.w3.org/ns/dcat#theme> <URI_Vocabulario> .

```

Figura 22: Estrutura das triplas do *Grafo de Interligações Semânticas* para cada conjunto de dados

O *Grafo de Interligações Semânticas* alinha-se às diretrizes FAIR para descoberta de dados (*Findability*), que pautam por metadados ricamente descritos (princípio F2), já que gera metadados semânticos e que, explicitamente, referenciam os dados por eles descritos (princípio F3). Além disso, seguindo os princípios FAIR de interoperabilidade (*Interoperability*), os metadados no *Grafo de Interligações Semânticas* fazem uso de vocabulários, ontologias e

tesauros processáveis por máquina (princípio I1). Os vocabulários, ontologias e tesauros possuem identificadores únicos e usam uma linguagem formal e compartilhada para representação dos dados (princípio I2) e utilizam referências qualificadas para outros metadados, pelo uso de propriedades que, por sua vez, também são definidas formalmente em vocabulários (princípio I3). Assim, as atividades para geração do grafo correspondem às atividades de FAIRificação dos metadados, conforme proposto por Jacobsen et al. (2020).

4.1.3. Geração das etiquetas semânticas

O fluxo de atividades descrito nessa seção tem como objetivo gerar metadados semânticos na forma de etiquetas, para os conjuntos de dados em RDF publicados em portal de dados abertos. Baseada na abordagem de Castro (2019) descrita com mais detalhes no Capítulo 3, a abordagem LigADOS também fornece uma estratégia para a extração automática de etiquetas semânticas. Considerando que dados em RDF naturalmente expõem a estrutura dos dados, os dados e os metadados, é possível extrair etiquetas a partir dos recursos que formam as triplas. Recursos anotados semanticamente, isto é, conceituados e associados a vocabulários, ontologias ou tesauros servem de base para extração de etiquetas.

Nesse ponto, cabe uma breve digressão. A versão mais atual do vocabulário DCAT⁴⁶ flexibilizou o domínio da propriedade “keyword” para quaisquer tipos de recursos (`dcat:Dataset`, `dcat:DataService`) mantendo seu *range* como um literal. As etiquetas também são associadas aos conjuntos de dados nas plataformas de dados abertos. Por isso, tanto Castro (2019) quanto a abordagem aqui descrita consideram associar as etiquetas geradas ao conjunto de dados. Entretanto, ambas as abordagens sugerem flexibilizar o *range* da propriedade supracitada, no sentido de apontar para uma classe/conceito de um vocabulário, ontologia ou tesouro, a qual tenha um rótulo associado que, em tempo de exploração do grafo, pode ser plenamente visualizada por caracterizar o nó que representa a classe/conceito.

Voltando à geração de etiquetas, a abordagem LigADOS obtém as classes/conceitos associados às distribuições dos conjuntos de dados persistidas no *Grafo de Interligações Semânticas*. As classes/conceitos são associadas aos conjuntos de dados. Para isso, a abordagem gera triplas usando a propriedade “keyword” do vocabulário DCAT (`dcat:keyword`) de forma a conectar os URI dos conjuntos de dados aos URI das classes/conceitos extraídos. As triplas são incluídas no *Grafo de Etiquetas Semânticas*, que é persistido em um repositório de um *triplestore*. Em função disso, é possível a navegação pelas

⁴⁶ W3C Recommendation 04 February 2020

etiquetas associadas a um conjunto de dados de forma a explorar possíveis outros conjuntos de dados que façam uso das mesmas etiquetas, descobrindo-se, assim, conexões indiretas entre conjuntos de dados. A diferença de LigADOS para a abordagem proposta por Castro (2019) está no fato de que a primeira obtém as etiquetas semânticas em um nível de abstração mais alto que a segunda, pois utiliza classes/conceitos de elementos/termos utilizados na anotação dos recursos das triplas. Outro diferencial de LigADOS está no fato de considerar conjuntos de dados modelados em RDF, enquanto Castro (2019) considera dados em formatos tabulares.

As descrições textuais das etiquetas, ou seja, os rótulos das classes/conceitos também são incluídos no *Grafo de Etiquetas Semânticas*, além de serem armazenadas e organizadas externamente ao grafo. Elas serão utilizadas posteriormente, na publicação do *Grafo de Interligações Semânticas* na instância do portal de dados.

4.1.4. Publicação das interligações semânticas entre conjuntos de dados

Os três fluxos de atividades descritos até aqui, concentram-se na construção das interligações semânticas entre conjuntos de dados, por uma camada de metadados de nível de registro de dados. O fluxo de atividades descrito nesta seção inclui as atividades que, efetivamente, disponibilizam as interligações semânticas em uma instância de portal de dados, para uso pelos consumidores de dados e aplicações externas.

Os consumidores de dados que utilizam portais de dados abertos possuem diferentes perfis de conhecimentos a respeito de tecnologias da Web de Dados. Em virtude da heterogeneidade de perfis, é preponderante que haja algum suporte intuitivo para uso de conjuntos de dados modelados em RDF (grafos). Alguns requisitos para atender essa demanda envolvem o fornecimento de uma interface que faça uso de recursos visuais e interativos.

Inicialmente, a premissa para exploração de qualquer grafo RDF é que ele esteja armazenado em um *triplestore*. O processo que gera o *Grafo de Interligações Semânticas*, o persiste em um *triplestore*, assim como também persiste o conjunto de dados em RDF contendo os metadados da instância do portal. Sendo assim, com o repositório populado, a apresentação visual do *Grafo de Interligações Semânticas* é criada. Para isso, é utilizada a forma CONSTRUCT da linguagem SPARQL. Em um *triplestore*, consultas definidas utilizando a forma CONSTRUCT retornam um grafo RDF simples, em conformidade com um *template* definido, a partir de dados armazenados no repositório (SPARQL Query Language for RDF, 2020). Com o grafo estruturado, a visualização inicial do grafo e os comportamentos das expansões dos nós são estabelecidos.

A apresentação inicial posiciona o tipo `dcat:Dataset` e os nós assim tipificados mais ao centro do grafo, próximo ao nó que representa a instância do portal de dados, isto é, do tipo `dcat:Catalog`. Contudo, uma vez disponibilizado, o consumidor é livre para expandir ou retrain os nós de acordo com sua escolha.

Há quatro tipos de conexões que podem ser expandidas dos nós que representam os conjuntos de dados. A primeira conexão estabelece uma classificação pois aponta para o tipo "Dataset". Os outros dois tipos de conexões representam a organização que publicou o conjunto de dados e as suas distribuições. Sendo assim, a única conexão que aponta para o conjunto de dados é a do portal no qual ele foi publicado.

Há quatro tipos de conexões ao se expandir um nó que representa uma distribuição. O primeiro tipo estabelece uma classificação para o nó ao apontar para o tipo "dcat:Distribution". Os outros dois tipos apontam para os vocabulários, ontologias e tesauros utilizados pela distribuição; e para as classes e/ou conceitos desses recursos ontológicos que representam um nível de abstração mais alto dos termos e/ou elementos utilizados para anotação das triplas. Ou seja, representam os metadados no nível de registro de dados. O último tipo de conexão aponta para o URL de onde pode ser feito o *download* da distribuição.

A Figura 23 apresenta um extrato do *Grafo de Interligações Semânticas* para um portal de dados abertos. A cada tipo de nó é atribuída uma cor distinta. Assim, no exemplo da Figura 23, aos conjuntos de dados foi atribuída a cor vermelha, às suas distribuições a cor roxa, aos vocabulários a cor verde, etc.

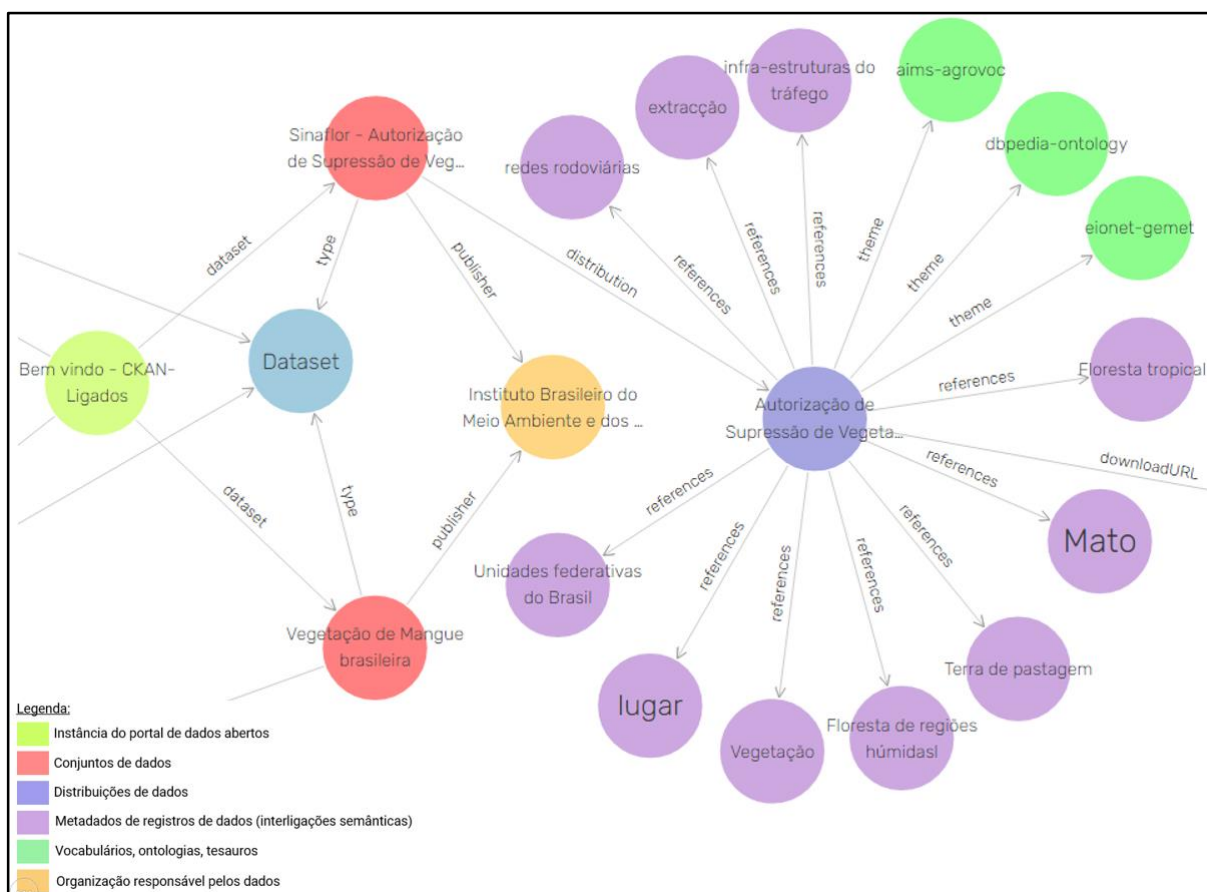


Figura 23: Apresentação do Grafo de Interligações Semânticas

No extrato da Figura 23 fica ostensivo que o conjunto de dados “Sinaflor - Autorização de Supressão de Vegetação” foi publicado pela organização “Instituto Brasileiro de Meio Ambiente” que também publicou o conjunto de dados “Vegetação Mangue Brasileira”. O primeiro conjunto de dados foi distribuído por meio do recurso “Autorização de Supressão de Vegetação”, o qual faz uso da ontologia “dbpedia-ontology” e dos tesouros “aims-agrovoc” e “eionet-gemet” para anotação de suas triplas.

A Figura 24 apresenta o nó do tesouro “eionet-gemet” expandido, de forma a mostrar os conceitos que foram utilizados para anotação semântica da distribuição do conjunto de dados⁴⁷: “redes rodoviárias”, “extracção” e “infra-estrutura do tráfego”. A ligação da distribuição com o tesouro funciona como um facilitador de navegação, pois um nó do grafo pode ser colapsado (ocultando vértices diretamente ligados a ele) para exibir somente os vocabulários, ontologias e tesouros utilizados, facilitando a observação de congruências entre distribuições de dados.

⁴⁷ As grafias dos conceitos utilizados nos rótulos dos nós do grafo foram obtidas de forma automatizada por LigADOS. Sendo assim, grafias de termos não alinhadas às regras da nova ortografia da língua portuguesa não são de responsabilidade de LigADOS.

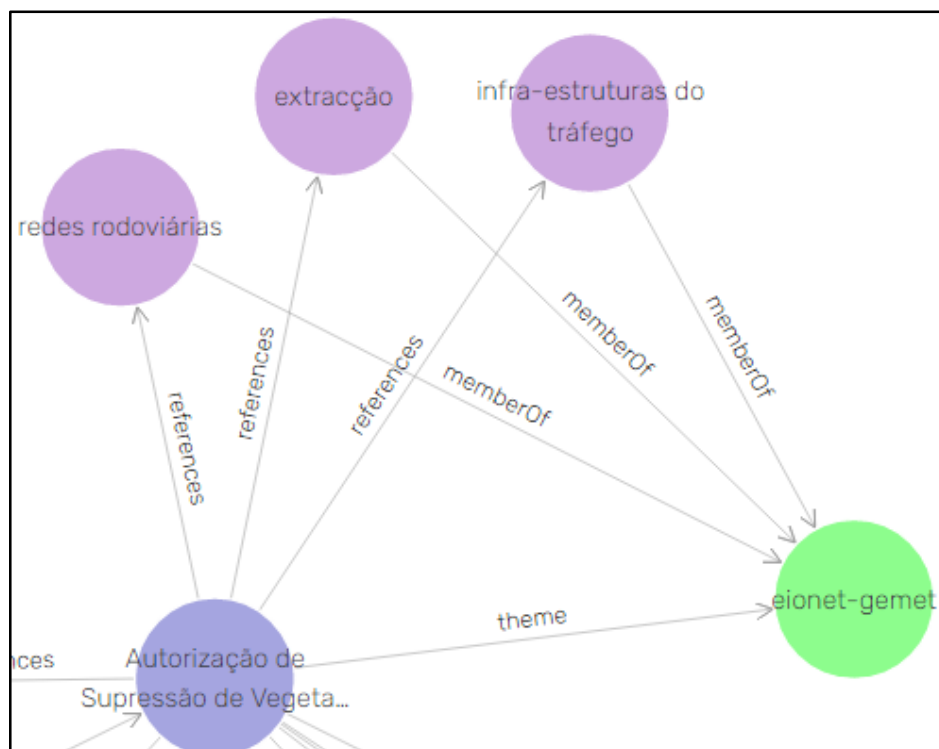


Figura 24: Tesauro “eionet-gemet” e os conceitos que foram utilizados para anotação semântica da distribuição

Em suma, o grafo viabiliza a expansão dos nós que representam:

- os conjuntos de dados de forma a visualizar suas distribuições, isto é, como os conjuntos de dados são disponibilizados aos consumidores;
- as distribuições dos conjuntos de dados de forma a descobrir as classes e/ou conceitos utilizados na anotação semântica de suas triplas;
- as classes e/ou conceitos dos vocabulários, ontologias e tesauros de forma a descobrir outros conjuntos de dados que o utilizam para anotação semântica, isto é, a exploração de interligação indireta;
- as classes e/ou conceitos dos vocabulários, ontologias e tesauros de forma a descobrir as ligações com classes e/ou conceitos de outros vocabulários alinhados, isto é, que tratem dos mesmos conceitos.

As configurações da apresentação são persistidas no próprio *triplestore* e o *Grafo de Interligações Semânticas* pode ser publicado em uma instância de portal. LigADOS faz a publicação, criando o conjunto de dados (o *Grafo de Interligações Semânticas*), suas distribuições e os metadados a serem cadastrados na instância de portal, incluindo as etiquetas semânticas. O grafo é distribuído de três formas:

- *RDF Dump*: disponibiliza a opção para *download* do arquivo;
- *Grafo Visual*: faz a chamada do URL associado a montagem dinâmica da apresentação do *Grafo de Interligações Semânticas* armazenado no *triplestore*; e
- *SPARQL Endpoint*: abre o URL do *SPARQL Endpoint* para submissão de consultas livres ao *Grafo de Interligações Semânticas*. Os objetivos são: a habilitação do processamento por máquina; e o consumidor de dados com conhecimentos nas tecnologias de Web Semântica, tornando a estrutura do grafo bastante transparente. Para facilitar o uso inicial do *Endpoint*, podem ser disponibilizados alguns exemplos de consultas em SPARQL junto à publicação.

O *Grafo de Etiquetas Semânticas* é publicado como uma das distribuições do *Grafo de Interligações Semânticas*. Para cadastramento das etiquetas na plataforma de portal, são recuperadas as descrições textuais das etiquetas semânticas geradas no processo de geração de etiquetas semânticas, conforme descrito na seção 4.1.3.

4.1.5. Geração de metadados semânticos como suporte a interligação de portais de dados

O suporte a interligação de portais de dados ocorre por meio de um fluxo de atividades cujo objetivo é gerar um conjunto de metadados semânticos alinhados ao perfil (*profile*) de metadados definido para um FDP. Ou seja, um perfil definido para um repositório central de metadados que fornece acesso a dados publicados em portais de dados abertos de forma federada.

O conjunto de metadados em RDF é anotado com o vocabulário DCAT, extraído da instância do portal, conforme detalhado na seção 4.1.1, juntamente com o *Gráfico de Interligações Semânticas* contêm a maior parte dos insumos para geração de metadados para uso em um FDP. Em função disso, ambas as fontes de metadados são lidas e, sistematicamente, quatro dos cinco níveis de organização dos metadados de um FDP são extraídos e estendidos com outros metadados definidos no perfil para um FDP. Este processo gera o *Grafo de Metadados FDP* que, por fim, é persistido em um *triplestore*. Embora a maior parte das triplas do *Grafo de Metadados FDP* seja extraída (copiada) das fontes de metadados, nem todos os elementos de metadados da especificação para um FDP são cobertos.

É possível a extração de metadados do nível de catálogo, de conjunto de dados, de distribuição até o nível de metadados de registro de dados, gerados por LigADOS. Ou seja, são extraídos o segundo, o terceiro, o quarto e o quinto nível de metadados previsto para um

FDP. O primeiro nível é representado pelos metadados do próprio FDP. Como toda estrutura de metadados depende da instalação do FDP, é fundamental que LigADOS receba o URL da instalação do FDP como parâmetro de entrada. Assim, entende-se que os metadados do FDP já foram gerados pelo responsável pela sua instalação.

A proposta de LigADOS inclui a geração todos os itens obrigatórios referentes aos segundo, terceiro, quarto e quinto níveis. Somente alguns itens opcionais serão também adicionados. Neste sentido, foram necessárias algumas adaptações e considerações.

Um item obrigatório e fundamental é a geração da estrutura hierárquica de metadados do FDP. Ela é estabelecida pela ligação de um nível ao seu nível imediatamente superior. Assim, os metadados no nível de distribuição (nível 4), por exemplo, contém a identificação dos metadados do nível de conjuntos de dados (nível 3).

LigADOS gera as triplas que definem os identificadores de camadas dos quatro níveis de metadados que estão sob seu alcance (catálogo, conjunto de dados, distribuição e registro de dados). Para isso, utiliza o URL do FDP para compor o URI do identificador dos níveis – “metadataIdentifier”. A Figura 25 apresenta as estruturas das triplas geradas para identificar as camadas de metadados.

```
<URI_Catalogo> <http://fdpontology.metadataIdentifier> <FDP_URI_Catalogo> .

<URI_ConjuntoDeDados> <http://fdpontology.metadataIdentifier>
<FDP_URI_ConjuntoDeDados> .

<URI_Distribuicao> <http://fdpontology.metadataIdentifier> <FDP_URI_Distribuicao> .
<URI_RegistroDados> <http://fdpontology.metadataIdentifier>
<FDP_URI_RegistroDados> .
```

Figura 25: Estruturas das triplas geradas por LigADOS para identificadores de camadas de metadados

Da mesma forma, LigADOS gera as triplas que representam os relacionamentos hierárquicos entre os níveis, isto é, ligando cada nível a seu “nível-pai”. O identificador do FDP, recebido como parâmetro, refere-se aos metadados do próprio FDP, que é a primeira camada. A Figura 26 apresenta as estruturas das triplas geradas para representar a hierarquia entre as camadas.

```
<FDP_URI_Catalogo> <http://purl.org/dc/terms/isPartOf> <FDP_URI_FDP> .

<FDP_URI_ConjuntoDeDados> <http://purl.org/dc/terms/isPartOf> <FDP_URI_Catalogo> .
<FDP_URI_Distribuicao> <http://purl.org/dc/terms/isPartOf>
<FDP_URI_ConjuntoDeDados> .

<FDP_URI_RegistroDados> <http://purl.org/dc/terms/isPartOf> <FDP_URI_Distribuicao>
.
```

Figura 26: Estruturas das triplas geradas por LigADOS para representação dos relacionamentos hierárquicos entre as camadas de metadados

A abordagem LigADOS considera que os metadados para registros de dados precisam referenciar os dados juntamente com a indicação de onde eles estão inseridos. Neste sentido, é estabelecido um relacionamento entre os dados e uma estrutura que os organizem. Nos portais construídos sobre plataformas de dados abertos, os dados estão relacionados às distribuições dos conjuntos de dados. Sendo assim, para descrever a relação entre dados e distribuições dos conjuntos de dados, LigADOS sugere o uso da classe “*Relationship*” (`dcat:Relationship`) do vocabulário DCAT. A Figura 27 apresenta as estruturas das triplas para tipificar as camadas de metadados, cuja fonte é o *Grafo de Interligações Semânticas*.

```
<URI_Catalogo> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/ns/dcat#Catalog> .

<URI_ConjuntoDeDados> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/ns/dcat#Dataset> .

<URI_Distribuicao> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/ns/dcat#Distribution> .

<Identificador_RegistroDeDados> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/ns/dcat#Relationship> .
```

Figura 27: Estruturas das triplas para tipificar as camadas de metadados

Cada item tipificado é identificado por seu URI. Os URI para catálogo, conjuntos de dados e distribuições são obtidos do *Grafo de Interligações Semânticas* sem transformações. Contudo, o identificador do registro de dados é montado neste processo. Para isso, nesta dissertação é proposta a flexibilização da subordinação dos metadados do nível de registro de dados ao nível de distribuição. Como as distribuições contêm as instâncias de dados, elas representam exatamente um nível hierárquico acima dos registros de dados, atuando como um *container*. No processo de geração do *Grafo de Metadados FDP* a flexibilização proposta é utilizada.

Um metadado de registro de dado está diretamente associado a uma classe/conceito extraído de um vocabulário, ontologia ou tesouro. Como uma classe/conceito pode ser utilizada por diferentes distribuições de dados, seja do mesmo conjunto de dados, seja de outros, ela sozinha não é capaz de caracterizar um único registro de dados. Faz-se necessária, então, uma combinação de dados para completa caracterização da classe/conceito.

Assim, a aplicação da abordagem alternativa para reificação de RDF — RDF* — surge como uma opção adequada. Conforme apresentado na seção 2.3, a abordagem RDF* permite aninhar triplas, anotando recursos de triplas com outras triplas de metadados. Triplas

aninhadas (embutidas) podem corresponder naturalmente ao conceito de identificadores de triplas (HARTIG; THOMPSON, 2019).

A caracterização de um registro de dados se dá pela combinação da identificação da distribuição de dados com as classes/conceitos de vocabulários, ontologias ou tesouros que caracterizam termos/elementos utilizados para anotação de suas triplas. A associação representa o cerne do processo de interligação proposto por LigADOS e exposto em detalhes na seção 4.1.2. A Figura 28 apresenta a estrutura da tripla que caracteriza o metadado no nível de registro de dados, conforme extraída do *Grafo de Interligações Semânticas*.

```
#Caracterização do metadado no nível de registro de dados de um recurso de dados

Em N-Triples:
<URI_Distribuicao> <http://purl.org/dc/terms/references> <URI_ClasseDoVocabulario>
.
-----

Em Turtle:
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix vocab: <http://namespacevocabulario.org/ns/> .

<URI_Distribuicao> dcterms:references vocab:URI_Classe .
```

Figura 28: Estrutura da tripla que caracteriza o metadado no nível de registro de dados

Conforme visto na seção 2.3 a serialização em Turtle é obrigatória, dada a extensão de sua gramática para Turtle*, que acomoda a reificação RDF*. A tripla que caracteriza um metadado do nível de registro de dados pode ser aplicada às estruturas apresentadas nas Figuras 23, 24 e 25, porém serializadas em Turtle*. A Figura 29 apresenta a tripla de metadado aninhada à tripla que identifica a camada de metadados, à tripla que define a hierarquia entre as camadas e à tripla que tipifica o recurso de dados.

Note que a tripla que identifica um recurso de dados com insumos extraídos do *Grafo de Interligações Semânticas* é embutida na posição de sujeito na estrutura para identificar a camada de metadados; e adaptada com o URI do FDP para ser embutida como objeto de “metadataIdentifier”. Assim, o *Grafo de Metadados FDP* terá um conjunto dessas triplas para cada classe/conceito utilizada por uma distribuição.


```

@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix vocab: <http://namespacevocabulario.org/ns/> .
@prefix fdp: <http://fdpontology/fdp-o#> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .

#Figura 25: identifica a camada de metadados
<<<URI_Distribuicao> dcterms:references vocab:URI_Classe>>
fdp:metadataIdentifier <<<FDP_URI_Distribuicao> dcterms:references
vocab:URI_Classe>> .

#Figura 26: liga a camada de metadados a sua camada pai
<<<FDP_URI_Distribuicao> dcterms:references vocab:URI_Classe>> dcterms:isPartOf
<FDP_URI_MetadadosDistribuicao> .

#Figura 27: tipifica o recurso ao qual a camada de metadados descreve
<<<URI_Distribuicao> dcterms:references vocab:URI_Classe>> rdf:type
dcat:Relationship .

```

Figura 29: Estrutura da tripla que caracteriza a camada de metadados no nível de registro de dados, hierarquização entre as camadas de registro de dados e distribuição e a tipificação de um recurso que representa a camada de metadados no nível de registro de dados

Outra questão constante da especificação de um FDP está na obrigatoriedade de um elemento de metadados no nível de catálogo que indique seu responsável. Ou seja, que represente a pessoa ou organização que criou e mantém a instância de portal. Porém, algumas plataformas de dados abertos, como o CKAN, não registram o criador/mantenedor como parte dos metadados. Em função disso, LigADOS gera a informação sobre quem é o responsável pelo catálogo. De um modo geral, os Portais de dados abertos governamentais construídos sobre o CKAN têm sido mantidos pela própria administração pública, muitas das vezes por colegiados responsáveis pela gerência central, mantendo cada unidade (organização) pública com algum nível de administração para que possa gerenciar seus próprios conjuntos de dados. Nesses casos, o URL de acesso ao portal contém referências para uma esfera governamental específica: continental, nacional, municipal, etc. Isso posto, LigADOS extrai parte do URL de acesso ao portal e o atribui ao elemento de metadado destinado a indicar o responsável pelo catálogo. A abordagem destaca o protocolo, para extrair e utilizar somente o nome do host (IP+domínio) como responsável pelo portal.

Outros elementos de metadados gerados por LigADOS são as taxonomias que caracterizam a área temática do catálogo e os conceitos que definem o conjunto de dados. Considerando que algumas plataformas de dados abertos não prevêem o registro desse tipo de dado, LigADOS pode buscar no *Grafo de Interligações Semânticas* os URIs dos vocabulários, ontologias e tesouros utilizados pelas distribuições de dados. Os URIs são utilizados para gerar as triplas que definem as taxonomias utilizadas pelos catálogo. Outrossim, LigADOS pode extrair as classes/conceitos associados às distribuições e utilizá-

las para gerar as triplas que definem os conceitos que definem os conjuntos de dados. As estruturas das triplas geradas são apresentadas no Figura 30.

```
<URI_Catalogo> <http://www.w3.org/ns/dcat#themeTaxonomy> "<URI_Vocabulario1>",
"<URI_Vocabulario2>" .

<URI_ConjuntoDeDados> <http://www.w3.org/ns/dcat#theme>
<URI_ClasseDoVocabulario1>, <URI_ClasseDoVocabulario2> .
```

Figura 30: Estrutura das triplas que caracterizam área temática do catálogo e conceitos dos conjuntos de dados

Portanto, LigADOS gera seis tipos de triplas para: identificar, hierarquizar e tipificar a camada de metadados; indicar o responsável pelo portal; atribuir temas ao portal; e vincular aos conjuntos de dados os conceitos que os caracterizam. As demais triplas são replicações das fontes de metadados para o *Grafo de Metadados FDP*.

De qualquer forma, a replicação de dados precisa ser sistematicamente realizada. Sendo assim, o processo de extração inicia-se com foco na segunda camada, que refere-se ao catálogo, isto é, a instância do portal de dados. Para extração das camadas de catálogo, conjunto de dados e distribuição, a fonte de dados é o conjunto de metadados extraído da instância do portal.

O Quadro 6 apresenta os elementos de metadados que são extraídos, a estrutura das triplas serializadas em Turtle e a indicação de obrigatoriedade ou não. A escolha da serialização em Turtle alinha-se com a necessidade de uso de Turtle* para algumas triplas, conforme anteriormente explicado. Os prefixos utilizados são:

```
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix dcam: <http://purl.org/dc/dcam/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
```

Quadro 6: Elementos de metadados extraídos para a camada de catálogo de dados

Elemento de Metadado	Triplas Geradas	Obrigatório
Tipo	<code><URI_Catalogo> rdf:type dcat:Catalog .</code>	Sim
Título	<code><URI_Catalogo> dcterms:title "String título Catálogo" .</code>	Sim
Versão	<code><URI_Catalogo> dcterms:hasVersion "String versão" .</code>	Sim
Organização ou Pessoa responsável pelo catálogo	<code><URI_Catalogo> dcterms:publisher <URI_Publicador_Catalogo> .</code>	Sim
Descrição	<code><URI_Catalogo> dcterms:description "String descrição" .</code>	Não
Conjuntos de Dados catalogados (uma tripla para cada conjunto de dados)	<code><URI_Catalogo> dcat:dataset <URI_ConjuntoDeDados> .</code>	Sim

Na sequência, são extraídos os metadados da terceira camada, que refere-se ao conjunto de dados. O Quadro 7 apresenta os elementos de metadados que são extraídos e a estrutura das triplas serializadas em Turtle.

Quadro 7: Elementos de metadados extraídos para a camada de conjunto de dados

Elemento de Metadado	Triplas Geradas	Obrigatório
Tipo	<code><URI_ConjuntoDeDados> rdf:type dcat:Dataset .</code>	Sim
Título	<code><URI_ConjuntoDeDados> dcterms:title "String título Conjunto de Dados" .</code>	Sim
Versão	<code><URI_ConjuntoDeDados> dcterms:hasVersion "String versão" .</code>	Sim
Organização ou Pessoa responsável pelo catálogo	<code><URI_ConjuntoDeDados> dcterms:publisher <URI_Publicador_Catalogo> .</code>	Sim
Descrição	<code><URI_ConjuntoDeDados> dcterms:description "String descrição".</code>	Não
Data de criação do conjunto de dados	<code><URI_ConjuntoDeDados> dcterms:issued <DataCriacao> .</code>	Não
Data de modificação do conjunto de dados	<code><URI_ConjuntoDeDados> dcterms:modified <DataModificacao> .</code>	Não
Distribuições associadas ao Conjunto de Dados (uma tripla para cada distribuição)	<code><URI_ConjuntoDeDados> dcat:dataset <URI_Distribuicao> .</code>	Sim
Palavras-chave	<code><URI_ConjuntoDeDados> dcat:keyword <URI_Etiqueta>⁴⁸ .</code>	Não

⁴⁸ A seção 4.1.3 discute sobre a atribuição de etiquetas e o uso do predicado `<dcat:keyword>`.

Em seguida, são extraídos os metadados da quarta camada, que referem-se às distribuições dos conjuntos de dados. O Quadro 8 apresenta os elementos de metadados que são extraídos e a estrutura das triplas serializadas em Turtle.

Quadro 8: Elementos de metadados extraídos para a camada de distribuições de dados

Elemento de Metadado	Triplas Geradas	Obrigatório
Tipo	<code><URI_Distribuicao> rdf:type dcat:Dataset .</code>	Sim
Título	<code><URI_Distribuicao> dcterms:title "String título Conjunto de Dados" .</code>	Sim
Versão	<code><URI_Distribuicao> dcterms:hasVersion "String versão" .</code>	Sim
Licença ⁴⁹	<code><URI_Distribuicao> dcat:license <URI_Licença> .</code>	Sim
Descrição	<code><URI_Distribuicao> dcterms:description "String descrição" .</code>	Não
URL de acesso à distribuição	<code><URI_Distribuicao> dcat:accessURL <URL_AcessoRecurso> .</code>	Sim
URL do arquivo que contém a distribuição em um determinado formato de dados	<code><URI_Distribuicao> dcat:downloadURL <URL_LocalArquivo> .</code>	Sim
Tipo do recurso da distribuição	<code><URI_Distribuicao> <dcat:mediaType> "String que define o tipo" .</code>	Sim
Formato do arquivo da distribuição	<code><URI_Distribuicao> dcat:format "String formato do arquivo" .</code>	Não

Para o nível de registro de dados a extração ocorre unicamente a partir do *Grafo de Interligações Semânticas*. O Quadro 9 apresenta os elementos de metadados que são extraídos e a estrutura das triplas serializadas em Turtle. O prefixo utilizado é:

```
@prefix vocab: <http://namespacevocabulario.org/ns/> .
```

⁴⁹ LigADOS faz uso da extensão "*ckanext-dcat*" para anotação dos metadados do portal. No modelo CKAN, a licença está no nível do conjunto de dados, enquanto no modelo DCAT está no nível de distribuição. Por padrão, o analisador RDF da extensão tentará encontrar a licença do conjunto de dados e anexará essa licença à distribuição. Assim, qualquer discrepância entre as licenças de distribuição não poderá ser resolvida sem intervenção de um agente externo à abordagem.

Quadro 9: Elementos de metadados extraídos para a camada de registro de dados

Elemento de Metadado	Triplas Geradas
Classe/conceito do vocabulário que foi utilizada	<URI_Distribuicao> dcterms:references vocab:URI_Classe.
Características de classe/conceito do vocabulário	vocab:URI_Classe dcam:memberOf <URI_Vocabulario> .
	vocab:URI_Classe rdfs:label "String NomeClasse"@pt .

O nível de metadados de registro de dados deve ser fornecido pelo administrador de dados e não pelo responsável pelo FDP. Assim, LigADOS automatiza essa atividade, como suporte aos administradores de dados dos portais. No ambiente FDP, o nível de registro de dados permitirá a identificação das interligações de dados gerenciadas por diferentes portais. O componente “*Metadata Provider*” do FDP pode recuperar o *Grafo de Metadados FDP* por um SPARQL *Endpoint* sempre que necessário e armazená-lo no *triplestore* local FDP.

É importante frisar que o macroprocesso aqui descrito simplifica um conjunto de atividades de inclusão de metadados que seriam realizadas pelo responsável do FDP e pelo administrador de um portal de dados.

A Figura 31 apresenta as estruturas das triplas referentes à camada do nível de registro de dados que compõem o *Grafo de Metadados FDP*, serializadas em Turtle incluindo Turtle*. Essa estrutura é utilizada para cada registro de dados reconhecido por LigADOS.

```
#Estrutura para cada Registro de Dados reconhecido por LigADOS
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix fdp: <http://fdpontology/fdp-o#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix voaf: <http://purl.org/vocommons/voaf#> .
@prefix vann: <http://purl.org/vocab/vann/> .
@prefix vocab: <http://namespacevocabulario.org/ns/> .
@prefix dcam: <http://purl.org/dc/dcam/> .

<<<URI_Distribuicao> dcterms:references vocab:URI_Classe>>
fdp:metadataIdentifier <<<FDP_URI_Distribuicao> dcterms:references
vocab:URI_Classe>> .

<<<FDP_URI_Distribuicao> dcterms:references vocab:URI_Classe>> dcterms:isPartOf
<FDP_URI_Distribuicao> .

<<<FDP_URI_Distribuicao> dcterms:references vocab:URI_Classe>> fdp:metadataIssued
"2020-08-20T10:20:37.08Z"^^xsd:dateTime;
<<<FDP_URI_Distribuicao> dcterms:references vocab:URI_Classe>>
fdp:metadataModified "2020-08-21T13:09:55"^^xsd:dateTime;

vocab:URI_Classe dcam:memberOf <URI_Vocabulario> ;
rdfs:label "String NomeClasse"@pt .
<URI_Vocabulario> rdf:type voaf:Vocabulary ;
vann:preferredNamespacePrefix "String Prefixo Vocabulário" .
```

Figura 31: Estrutura das triplas da camada de Registro de Dados do *Grafo de Metadados FDP*

Visto que as interligações semânticas geradas por LigADOS são embutidas no *Grafo de Metadados FDP*, é possível a observação de interligações semânticas indiretas entre dados armazenados em diferentes instâncias de portais. A exposição e visualização das interligações dependem de implementações no FDP (no *software* FDP), conforme proposto no processo de FAIRificação. Um esquema de interligações de dados entre portais, sem utilizar consultas federadas, é apresentado na Figura 32, onde considera-se que um *Metadata Provider* recuperou metadados de duas instâncias de portais, nas quais LigADOS foi aplicada.

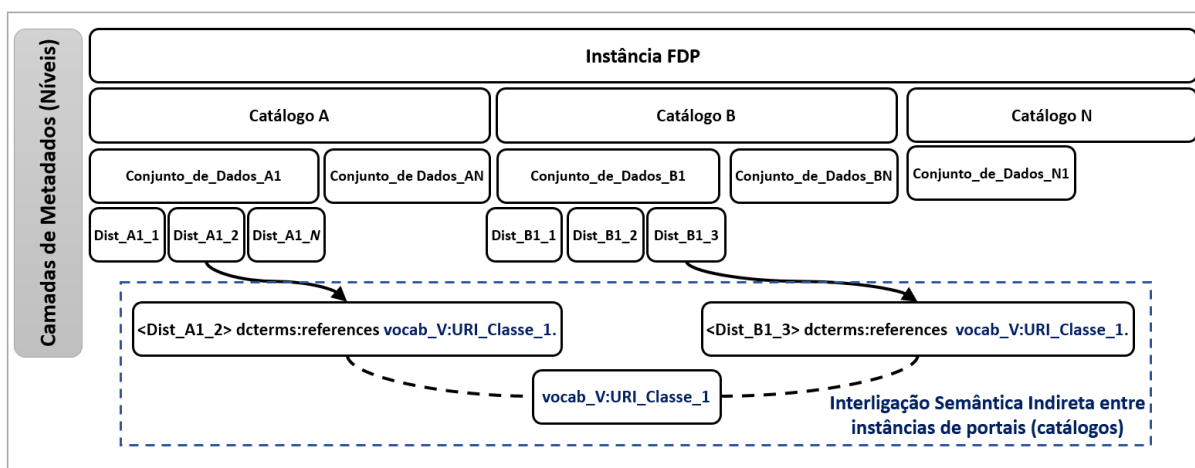


Figura 32: Esquema exemplificativo de interligações de dados de diferentes instâncias de portais em um FDP

O Apêndice A apresenta um comparativo entre os esquemas de metadados definido para um FDP e o definido e utilizado por LigADOS.

O *Grafo de Metadados* FDP alinha-se aos princípios FAIR para capacidade de descoberta (*Findability*) ao gerar metadados: (i) com identificadores únicos e persistentes (princípio F1); (ii) descritos por metadados semânticos (princípio F2); (iii) explicitamente referenciando os dados descritos (princípio F3); e (iv) preparados para publicação em um FDP (princípio F4).

4.2. Processo de Suporte Ontológico para Anotações Semânticas de Conjuntos de Dados

Para aumentar a possibilidade de interligações de dados, a abordagem LigADOS fornece um nível de suporte inicial para atividades de anotação semântica. O intuito é aumentar a probabilidade de que novos conjuntos de dados sejam interligados a outros conjuntos de dados já publicados numa instância de portal. O processo baseia-se na premissa de que, uma parte fundamental no processo de publicação de dados conectados inclui a

definição de um modelo de dados para maximização da interoperabilidade semântica entre conjuntos de dados. Outra questão importante é que o modelo de dados reutilize ao máximo vocabulários, ontologias e tesauros existentes (CORDEIRO et al., 2011).

Destarte, a abordagem LigADOS gera a *Biblioteca de Recursos Semânticos* da instância do portal. A biblioteca atua como um mecanismo de sugestão de soluções ontológicas que já estão em uso por conjuntos de dados publicados. A informação das soluções ontológicas fomenta o reuso. Ademais, ela aumenta a confiança dos envolvidos no processo inicial de anotação semântica, pois eles podem tomar conhecimento de soluções que já foram testadas e utilizadas. Caso alguma classe e/ou conceito tenha relação com os novos dados a anotar, o interessado pode buscar, externamente à solução, URIs específicas para anotar recursos de seu conjunto de dados.

Conforme apresentado na Figura 33, o processo de suporte ontológico foi dividido em dois macroprocessos. As subseções 4.2.1 e 4.2.2 apresentam as atividades de cada macroprocesso.

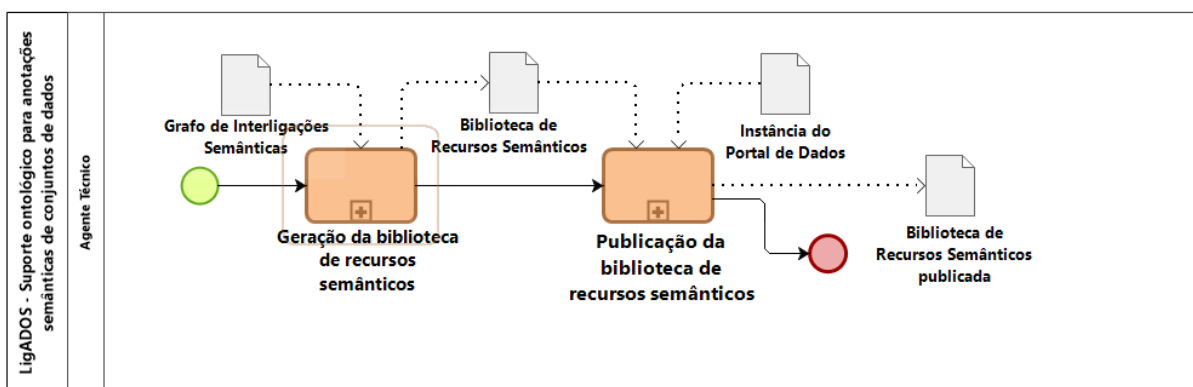


Figura 33: Processo de suporte ontológico fornecido por LigADOS

4.2.1. Geração da biblioteca de recursos semânticos

Com base no *Grafo de Interligações Semânticas* é possível identificar quais classes e/ou conceitos de vocabulários, ontologias ou tesauros estão sendo utilizados para anotação das triplas das distribuições de dados. Desta forma, é possível a geração de uma biblioteca de recursos semânticos que foram utilizados na instância de portal de dados. Esta biblioteca é estruturada na forma de grafo RDF, contendo triplas cujos sujeitos são os URI das classes/conceitos e os objetos são os URI dos vocabulários, ontologias e tesauros. O predicado é anotado com a propriedade “memberOf” do vocabulário DCMI *Metadata Terms* (dcam:memberOf), indicando que uma classe/conceito foi definida em um vocabulário, ontologia ou tesouro. O grafo RDF gerado é denominado *Biblioteca de Recursos Semânticos* e, assim como os demais grafos gerados por LigADOS, é persistido em um repositório de um *triplestore*. Cada instância de portal terá somente uma biblioteca. É importante observar que,

abordagens como a de Gomes et al. (2014) podem utilizar a biblioteca como alvo para lançamento de sementes para obtenção de conceitos de ontologias.

4.2.2. Publicação da biblioteca de recursos semânticos

O fluxo de atividades para publicação da biblioteca semântica tem como objetivo tornar público e de fácil acesso todas as classes e/ou conceitos de vocabulários, ontologias e tesouros utilizados pelas distribuições dos conjuntos de dados publicados em uma instância de portal. Assim como no processo de publicação das interligações semânticas, a forma CONSTRUCT da linguagem SPARQL é utilizada no grafo que contém a *Biblioteca de Recursos Semânticos*. Com base no *template* utilizado na consulta, a apresentação visual da *Biblioteca de Recursos Semânticos* é criada. As configurações da apresentação são persistidas no próprio repositório e a biblioteca pode ser publicada na instância do portal. LigADOS cria o conjuntos de dados que representa a biblioteca, suas distribuições e os metadados a serem cadastrados na instância de portal. O grafo é distribuído de quatro formas:

- *RDF Dump*: disponibiliza a opção para *download* do arquivo;
- *Grafo Visual*: faz a chamada do URL associado a montagem dinâmica da apresentação da *Biblioteca de Recursos Semânticos* armazenado no *triplestore*;
- *SPARQL Endpoint*: abre o URL do *SPARQL Endpoint* para submissão de consultas livres ao grafo; e
- *Tabular*: apresenta os rótulos das classes e/ou conceitos ordenadas pelas classes e/ou conceitos mais utilizados.

Portanto, a *Biblioteca de Recursos Semânticos* contribui para uma estratégia auto-sustentável, pois quanto mais classes de vocabulários forem reutilizadas, mais interligações serão estabelecidas no *Grafo de Interligações Semânticas*.

4.3. Arquitetura de Suporte

Todas as atividades descritas nas seções 4.1 e 4.2 são executadas pelos componentes de uma arquitetura. A arquitetura de suporte à abordagem e sua implementação são detalhados nas seções 4.3.1 e 4.3.2.

4.3.1. Componentes da Arquitetura

A arquitetura de suporte à LigADOS é composta de quatro camadas. A Figura 34 apresenta a arquitetura de suporte e as relações entre as partes que a compõem.

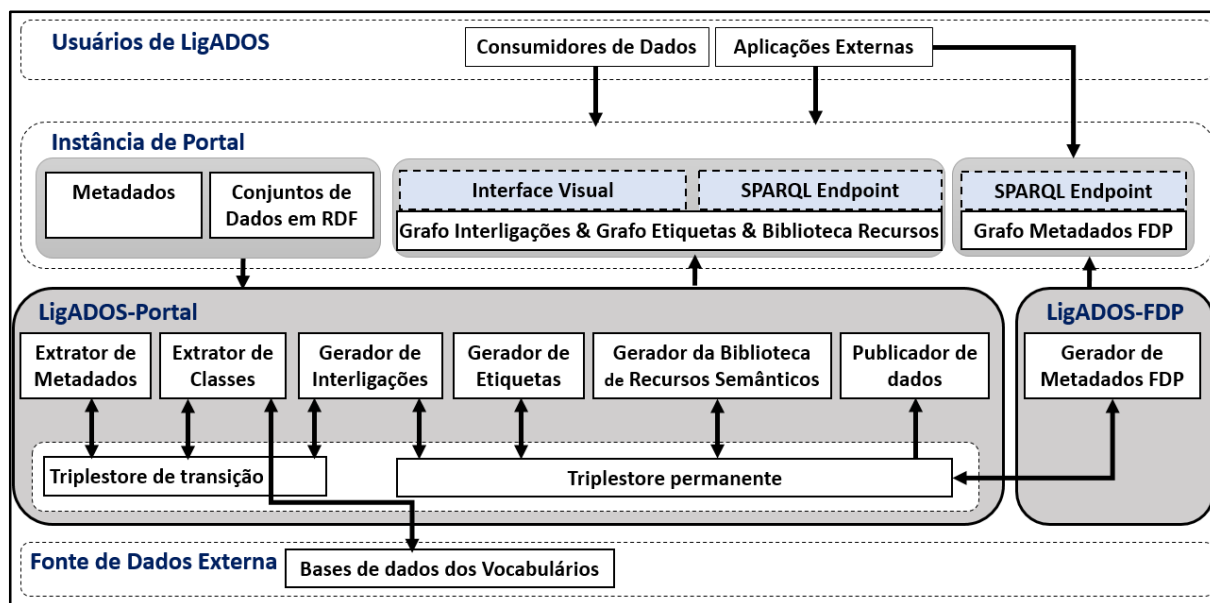


Figura 34: Arquitetura de suporte à abordagem LigADOS

A **primeira camada** da arquitetura faz referência aos usuários de LigADOS que têm interação direta com uma instância de portal de dados. Os usuários de LigADOS são de dois tipos: Consumidores de Dados e Aplicações Externas.

Os consumidores de dados acessam a instância de portal e utilizam a interface visual e interativa de LigADOS. A interface visual permite que os consumidores de dados naveguem e explorem as interligações entre distribuições de dados e conjuntos de dados por meio do *Grafo de Interligações Semânticas* e também por meio do *Grafo de Etiquetas Semânticas*. Consumidores interessados em saber quais vocabulários, ontologias e tesouros são utilizados pelos conjuntos de dados publicados no portal navegam e exploram visualmente a *Biblioteca de Recursos Semânticos*. A biblioteca fornece um panorama das classes/conceitos de vocabulários, ontologias e tesouros que categorizam os dados disponibilizados na instância de portal. Dessa forma, a biblioteca fornece entendimento da composição dos recursos ontológicos organizados em um só local.

As aplicações externas representam o consumo e o processamento por máquina. LigADOS suporta buscas e consultas SPARQL no conjunto de interligações gerado por LigADOS, fornecendo um URL para um *SPARQL Endpoint*. A abordagem também suporta buscas e consultas aos recursos da *Biblioteca de Recursos Semânticos*.

A partir da primeira camada, aplicações externas também fazem a extração de conjuntos de metadados para compor um FDP. As aplicações externas podem recuperar metadados organizados em camadas utilizando buscas e consultas SPARQL.

A instância de portal é representada na **segunda camada**, com duas funções: (i) fornecer os insumos para que LigADOS seja aplicado; e (ii) receber os resultados produzidos por LigADOS. Na função de fornecer insumos à LigADOS, a instância do portal disponibiliza

os metadados cadastrados pelos responsáveis pelas publicações de cada conjunto de dados. Na função de receber os resultados, a instância do portal é utilizada para disponibilização da interface que faz a mediação entre os usuários de LigADOS e as interfaces de acesso ao *Grafo de Interligações Semânticas* e à *Biblioteca de Recursos Semânticos*. A Figura 35 apresenta um extrato da publicação dos resultados gerados por LigADOS aplicado a uma instância CKAN, ilustrando como o resultado é disponibilizado na **segunda camada**, a instância do portal. “LigADOS” é publicado como conjunto de dados, contendo seis distribuições: (i) opção de *download* do *Grafo de Interligações Semânticas*; (ii) apresentação visual do *Grafo de Interligações Semânticas* para exploração com interatividade; (iii) URL do *SPARQL Endpoint* para acesso às interligações entre distribuições de dados em RDF no portal; (iv) opção de *download* do *Grafo de Etiquetas Semânticas*; (v) apresentação visual do *Grafo de Etiquetas Semânticas* para exploração com interatividade; e (vi) URL do *SPARQL Endpoint* para acesso às interligações entre conjuntos de dados via etiquetas em comum.

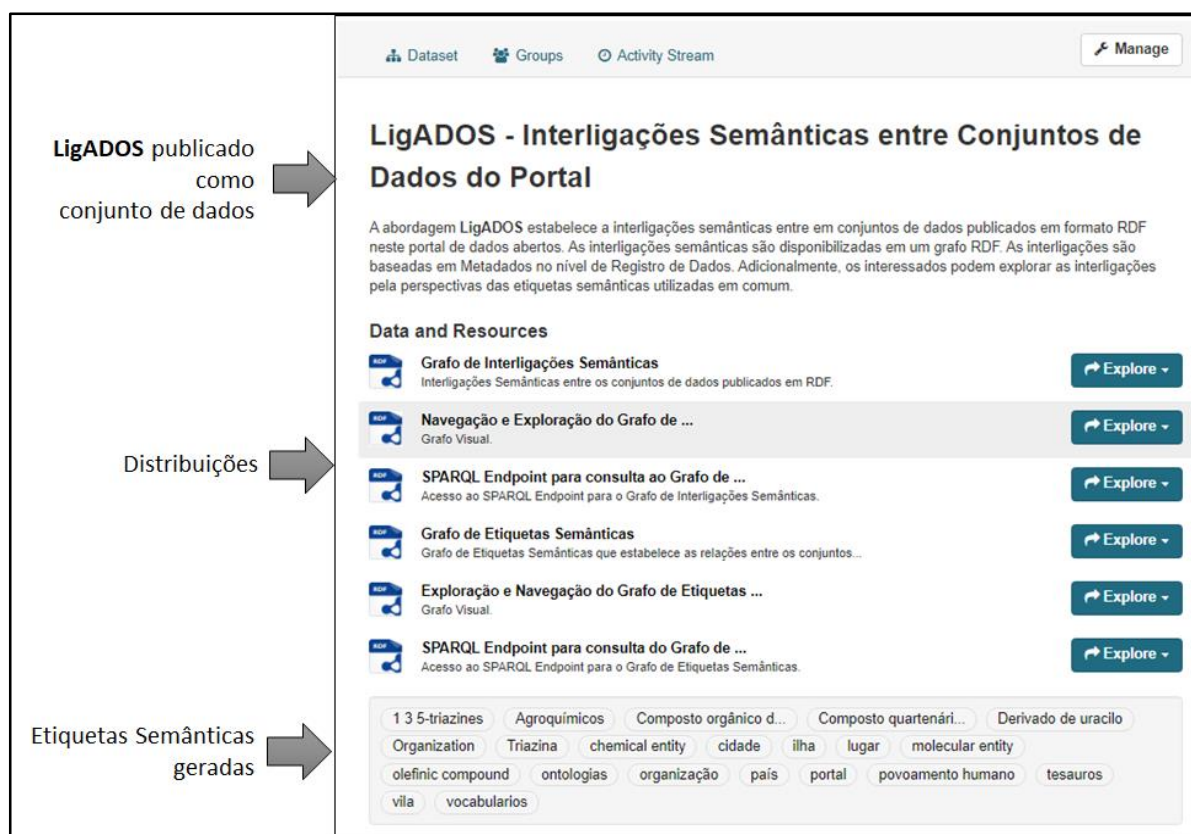


Figura 35: Extrato da instância CKAN mostrando os resultados gerados por LigADOS

A **terceira camada**, composta por duas partes, engloba os componentes da abordagem proposta – LigADOS – que são agrupados em dois módulos: LigADOS-Portal e LigADOS-FDP. LigADOS-Portal é responsável pela geração e disponibilização das interligações e de outros recursos de suporte à interligação. LigADOS-FDP é responsável pela geração de subsídios para esquemas de metadados em um FDP. Significa que LigADOS-

Portal executa funções desde a preparação da estrutura para seu uso até a disponibilização de seus resultados finais. Em contrapartida, LigADOS-FDP faz uso dos resultados gerados por LigADOS-Portal para gerar seus resultados.

O módulo LigADOS-Portal é composto por seis componentes: Extrator de Metadados, Extrator de Classes, Gerador de Interligações, Gerador de Etiquetas, Gerador de Recursos Semânticos e Publicador de Dados. O módulo LigADOS-FDP possui somente o componente Gerador de Metadados FDP.

O componente “Extrator de Metadados” faz a extração dos metadados da instância do portal utilizando a extensão “*ckanext-dcat*”, detalhada no Apêndice B. O componente “Extrator de Classes” executa buscas na quarta camada. A **quarta camada** é composta pelas “Fontes de Dados Externas”, representada pelos vocabulários, ontologias e tesouros para obtenção das classes e/ou conceitos a serem utilizados por LigADOS. Assim, a terceira e quarta camadas interagem por meio de um fluxo de consultas federadas que partem da terceira para a quarta camada; além de um fluxo da quarta para a terceira camada que fornece os dados recuperados.

O componente “Gerador de Interligações” é o responsável pela construção do *Grafo de Interligações Semânticas*, o cerne de LigADOS. O “Gerador de Etiquetas” monta as triplas do *Grafo de Etiquetas Semânticas* assim como provê os subsídios para registro das etiquetas semânticas geradas quando o *Grafo de Interligações Semânticas* for publicado. O componente “Gerador da Biblioteca de Recursos Semânticos” é o responsável pela construção do grafo que contém a *Biblioteca de Recursos Semânticos* e pelo ranqueamento das classes e/ou conceitos mais utilizados. E, por fim, o componente “Publicador de Dados” publica os grafos gerados em uma instância CKAN, fornecendo mecanismos de acesso às interligações e à biblioteca.

4.3.2. Implementação da Arquitetura

Os protótipos dos componentes Extrator de Metadados, Extrator de Classes, Gerador de Interligações, Gerador de Etiquetas e Gerador de Recursos Semânticos do módulo LigADOS-Portal foram desenvolvidos em Java que fazem uso do framework RDF4J, explicado no Apêndice B. Na arquitetura apresentada na Figura 31, “*Triplestore* de transição” refere-se ao suporte de banco de triplas oferecido pelo RDF4J; e “*Triplestore* permanente” é o GraphDB™. O GraphDB™ possui funcionalidades para que o usuário explore e navegue por grafos gerados a partir do resultado de consultas utilizando a forma CONSTRUCT de SPARQL, disponibilizando, inclusive, funcionalidades de interatividade.

O desenvolvimento e experimentação da abordagem LigADOS foram realizados em instâncias de portais de dados abertos baseados no *Comprehensive Knowledge Archive*

Network (CKAN), a plataforma de portal de dados de código aberto e gratuita detalhada no item 2.2.1, que tem se apresentado como padrão para portais de dados abertos (COSTA et al., 2017; NEUMAIER et al., 2017). Por conseguinte, foi necessária a preparação de um ambiente de desenvolvimento que inclui o CKAN como recurso de *software*.

O componente “Publicador de Dados” do módulo LigADOS-Portal depende da instalação do CKAN. O “Publicador de Dados” utiliza a CKANAPI, API oficialmente disponibilizada pelos desenvolvedores do CKAN, para publicar todos os grafos gerados por LigADOS em uma instância CKAN. Utilizando a API, *scripts* em *Python* são executados a partir da interface de linha de comando de um sistema operacional Linux. A execução dos *scripts* é organizada em um arquivo de lote para simplificar a ordenação da execução. Os *scripts* criam a organização, o conjunto de dados (*Grafo de Interligações Semânticas* e *Biblioteca de Recursos Semânticos*), suas distribuições e os metadados a serem cadastrados na instância de portal, incluindo as etiquetas semânticas.

Cabe ressaltar que uma instância CKAN pode ser configurada para permitir que conjuntos de dados sejam criados sem estarem vinculados a nenhuma organização. Nesse caso, qualquer usuário administrador pode alterar o conjunto de dados, o que não parece ser adequado para portais de dados abertos, onde o registro da proveniência dos dados bem como a identificação do responsável pela publicação são fatores importantes. Assim, para que haja somente um ponto de gerenciamento do grafo, faz-se necessário vincular o *Grafo de Interligações Semânticas* e a *Biblioteca de Recursos Semânticos* a uma organização. Por isso, uma organização genérica é criada pelos *scripts* em *Python*.

O Apêndice B detalha os recursos de *hardware* e *software* utilizados para o desenvolvimento e experimentação de LigADOS. Os artefatos utilizados por LigADOS, assim como o código-fonte dos módulos que compõem a abordagem estão disponíveis no GitHub⁵⁰.

4.4. Considerações Finais

Este capítulo apresentou as principais contribuições da dissertação, uma abordagem para interligação de conjuntos de dados publicados em plataformas de dados abertos, chamada LigADOS.

As atividades, arquitetura e estratégias de LigADOS foram detalhadas. Foi apresentado como elementos que classificam as instâncias de dados modelados em triplas, podem ser utilizados como pontos de interligações entre conjuntos de dados publicados. Baseado nisso, LigADOS adota a estratégia de extração de conceitos classificadores, utilizados para descoberta das interligações, para geração de metadados no nível de registro

⁵⁰ <https://github.com/glaugb/LigADOS>

de dados. O nível de metadados de registro de dados está previsto na proposta de camadas de metadados para um FDP.

Neste sentido, é interessante observar que o *Grafo de Interligações Semânticas* é um grafo de metadados. Os metadados do *Grafo de Interligações Semânticas* juntamente com os metadados extraídos da instância do portal, e anotados com o vocabulário DCAT, formam insumos para um FDP. De fato, os metadados supracitados fornecem dados para quatro dos cinco níveis de esquemas de metadados previstos para um FDP. Assim, os metadados para os níveis de catálogo (a instância da plataforma de portal), de conjuntos de dados, suas distribuições e registros de dados representam, respectivamente, segundo, terceiro, quarto e quinto níveis. O quinto e último nível é representado pelas classes/conceitos extraídos de vocabulários, ontologias e tesouros a partir das triplas das distribuições, ou seja, alinhado ao domínio específico dos dados. A lacuna para o esquema do FDP é representada pela falta do primeiro nível, representando os metadados atinentes ao FDP. Entretanto, o primeiro nível deve ser fornecido pelo responsável pela instalação do FDP.

Os metadados no nível de registro de dados carregam a base para as interligações semânticas indiretas, seja entre os conjuntos de dados publicados em uma mesma instância de portal, seja entre conjuntos de dados publicados em diferentes instâncias de portais. O *Grafo de Interligações Semânticas* sozinho contém as interligações dentro de um portal. Quando *Grafos de Interligações Semânticas*, pertencentes a instâncias de portais distintas, são carregados em um FDP, eles viabilizam o estabelecimento das interligações entre portais. Como a estrutura de FDP é pautada na catalogação de metadados para alcançar dados federados, as interligações entre portais se dão por meio de metadados semânticos. Essa concentração de metadados em um ponto central viabiliza que a consulta às interligações seja realizada no ambiente do FDP, sem a necessidade de uso de consultas federadas aos portais de origem.

Como o LigADOS extrai classificações de instâncias para gerar metadados de registros de dados, ele fornece suporte para a criação de índices. Como cada classe/ conceito extraído representa vários recursos (nós) dos conjuntos de dados originais (grafos), pode-se inferir determinado tipo de informação, sem realmente consultar os conjuntos de dados. Embora alguns benefícios gerados possam ser semelhantes aos da sumarização de RDF, LigADOS não aplica nenhum tipo de método de sumarização de RDF. Čebirić et al. (2019) descreveu uma taxonomia de alto nível das abordagens de sumarização de RDF, agrupando os principais métodos dos algoritmos em quatro categorias principais (estrutural, estatístico, mineração de padrões e híbrido), identificando subcategorias sempre que possível. Čebirić et al. também classificaram as abordagens de acordo com sua entrada, saída, disponibilidade na internet e finalidade, mostrando a dinâmica em rápida evolução na área. Em contraste,

LigADOS cria um novo conjunto de dados para representar interligações de dados não explícitos entre todos os conjuntos de dados publicados em um portal. LigADOS considera apenas recursos anotados, ou seja, recursos definidos em vocabulários, ontologias ou tesauros, em qualquer posição da tripla, sem se preocupar em reduzir o tamanho dos conjuntos de dados.

Portanto, LigADOS fornece uma estratégia de interligação de dados que estende algumas das abordagens comentadas na seção 3.2. LigADOS viabiliza que interligações entre conjuntos de dados, outrora implícitas, possam ser explicitamente exploradas pelos consumidores de dados. O produto da aplicação da estratégia em uma instância de portal é visualmente disponibilizado aos seus consumidores de dados, por meio de uma interface interativa, que fornece suporte para busca, consulta e navegação pelas interligações de dados. É importante reiterar que não foram encontradas na literatura soluções de integração de tecnologias de plataformas de dados abertos e *triplestores* com funcionalidades de construção e exploração visual de grafos de forma integrada conforme proposto por LigADOS.

No que tange à geração das interligações, primordialmente, toda estratégia é baseada no processamento automatizado, isto é, sem intervenção por parte de quaisquer tipos de usuários. Em função disso, cabe citar que técnicas de alinhamento de ontologias, que muitas das vezes requerem intervenção de especialistas do domínio associada ao uso de ferramental especializado, não foram consideradas no escopo de LigADOS. Por isso, somente relações simples entre vocabulários, ontologias e tesauros e a ontologia DBpedia foram consideradas. Sendo a DBpedia de domínio bastante amplo, ela pode servir como *hub* para outras bases de conhecimento. Ademais, diversos trabalhos propõem técnicas de alinhamento de ontologias, inclusive alinhamento especificamente com a DBpedia (DOJCHINOVSKI et al., 2016; KOLYVAKIS; KALOUSIS; KIRITSIS, 2018; AZMY et al., 2018).

5. APLICAÇÃO DE LigADOS

Conforme detalhado no capítulo 4, LigADOS interliga conjuntos de dados em RDF, gera metadados semânticos e corrobora para a geração dos esquemas de metadados previstos para um FDP, por meio de dois módulos: LigADOS-Portal e LigADOS-FDP. O módulo LigADOS-Portal tem como escopo uma instância de portal. O módulo LigADOS-FDP requer a existência prévia de um FDP, fora do escopo de uma instância de portal. Portanto, um ambiente deve ser configurado para aplicação de ambos os módulos. Como LigADOS-FDP requer a aplicação prévia de LigADOS-Portal, neste capítulo será apresentada a aplicação de LigADOS-Portal, o módulo responsável pelo cerne da abordagem. Sendo assim, na seção 5.1 é apresentada a configuração do ambiente para aplicação de LigADOS-Portal. As seções 5.2 e 5.3 apresentam a aplicação de LigADOS-Portal em duas instâncias de portais e as análises possíveis de serem efetuadas. Por fim, a seção 5.4 discute a aplicação de LigADOS-Portal pela perspectiva de uso por parte do consumidor de dados humano.

5.1. Configuração do ambiente

A aplicação de LigADOS-Portal baseia-se em algumas premissas. Conforme apresentado na seção 4.3.2, parte de LigADOS-Portal foi desenvolvida utilizando a CKANAPI. Em função disso, o ambiente para aplicação de LigADOS requer um portal de dados abertos construído sob a plataforma de dados abertos CKAN.

A necessidade do portal de dados abertos ter sido construído sobre CKAN é pautada na popularidade dessa plataforma pelo setor público de diversas localidades e, principalmente, na disponibilidade de ferramentas para desenvolvimento de extensões, que atendam demandas específicas de cada solução, conforme discutido nas seções 2.2.1 e 2.2.4.

Para demonstração da aplicação de LigADOS-Portal foram selecionados dois portais dentre os portais construídos sob a plataforma de dados abertos CKAN: o Portal Brasileiro de Dados Abertos e o Portal de Dados Abertos sobre Agrotóxicos. A aplicação de LigADOS nos portais visa: (i) verificar a capacidade da abordagem em corroborar com a inferência de novas informações por meio da navegação pelas interligações de um recurso a outro; e (ii) verificar como as interligações geradas pela abordagem contribuem para tornar explícitas as descrições das instâncias de dados.

O Portal Brasileiro de Dados Abertos foi selecionado como um exemplo de portal com múltiplas temáticas, destinado primordialmente à esfera federal. A variedade de assuntos dos conjuntos de dados publicados por diferentes entidades, unidades e organizações públicas

gera a possibilidade de construção de conhecimentos a partir da observação dos dados por diferentes perspectivas.

O Portal de Dados Abertos sobre Agrotóxicos foi selecionado como um exemplo de portal com foco em uma única temática, especificamente, nos assuntos que envolvem comercialização, uso, consequências de uso e relações de agrotóxicos. Como o PPGI-UFRJ é parceira na manutenção do portal, há um esforço em torná-lo cada vez mais robusto na quantidade de conjuntos de dados publicados, bem no alinhamento com tecnologias que habilitam compreensão dos significados dos dados.

Como os portais estão em produção, foi necessária a criação de duas instâncias CKAN distintas para representar extratos de ambos os portais. Os detalhes da criação das instâncias CKAN estão descritos no Apêndice B.

Para seleção dos extratos de ambos os portais foram considerados conjuntos de dados publicados por diferentes organizações. Como LigADOS-Portal é uma abordagem aplicável a conjuntos de dados modelados em RDF, inicialmente foram pesquisados quais eram os conjuntos de dados publicados em RDF nas instâncias de portais em produção.

No Portal Brasileiro de Dados Abertos foram encontrados poucos conjuntos de dados publicados em RDF. Em função disso, foram extraídas quatro distribuições de conjuntos de dados em seus formatos originais que, posteriormente, foram triplificadas, anotadas com diferentes vocabulários, ontologias e tesouros, e publicadas na instância CKAN representante do portal, destinada à aplicação da abordagem. Do Portal Brasileiro de Dados Abertos somente uma distribuição modelada em RDF foi extraída e utilizada sem pré-processamento.

Como no Portal de Dados Abertos sobre Agrotóxicos não foram encontradas publicações de conjuntos de dados em RDF, foram extraídas seis distribuições de conjuntos de dados em seus formatos originais. As distribuições também foram triplificadas, anotadas e publicadas na outra instância CKAN representante do portal e destinada à aplicação de LigADOS-Portal.

Informações sobre o processo de triplificação e anotação semântica das distribuições de dados extraídas de ambos os portais está descrito no Apêndice C. Sendo assim, os extratos dos portais contêm dados reais, porém modelados como conjuntos de dados em RDF.

Outros portais de dados abertos construídos sobre o CKAN foram verificados⁵¹. Contudo, ou não haviam dados modelados em RDF e a triplificação e anotação exigia a intervenção de especialistas do negócio; ou quando dados modelados em RDF estavam

⁵¹ Exemplos de portais analisados são: European Data Portal: <https://www.europeandataportal.eu/en>; Dados da Diretoria e Pesquisas do Jardim Botânico do Rio de Janeiro: <https://ckan.jbrj.gov.br/>

disponíveis, eles utilizavam a mesma ontologia, desenvolvida especificamente para atender a um processo de negócio específico; ou ainda não estavam anotados.

5.2. Portal Brasileiro de Dados Abertos

O Portal Brasileiro de Dados Abertos é uma instância CKAN, utilizada pelo setor público brasileiro para disponibilizar dados relativos às mais variadas temáticas. Embora a proposta inicial seja a publicação de dados relativos à esfera federal, é possível encontrar dados de outras esferas da organização administrativa do país.

Dentre os conjuntos de dados que foram triplificados, como forma de pré-processamento para uso de LigADOS, dois deles serão inicialmente apresentados como exemplos de diferentes formas de publicação encontradas no portal. Inicialmente, não há relações estabelecidas entre esses conjuntos de dados que, inclusive, foram publicados por organizações distintas.

A Figura 36 apresenta a publicação original do conjunto de dados “Vegetação de Mangue Brasileira”, contendo três distribuições de conteúdo similar, alterando somente o formato dos arquivos de dados. A distribuição “CSV – Vegetação de mangue” foi triplificada para utilização de LigADOS, conforme comentado na introdução deste capítulo.

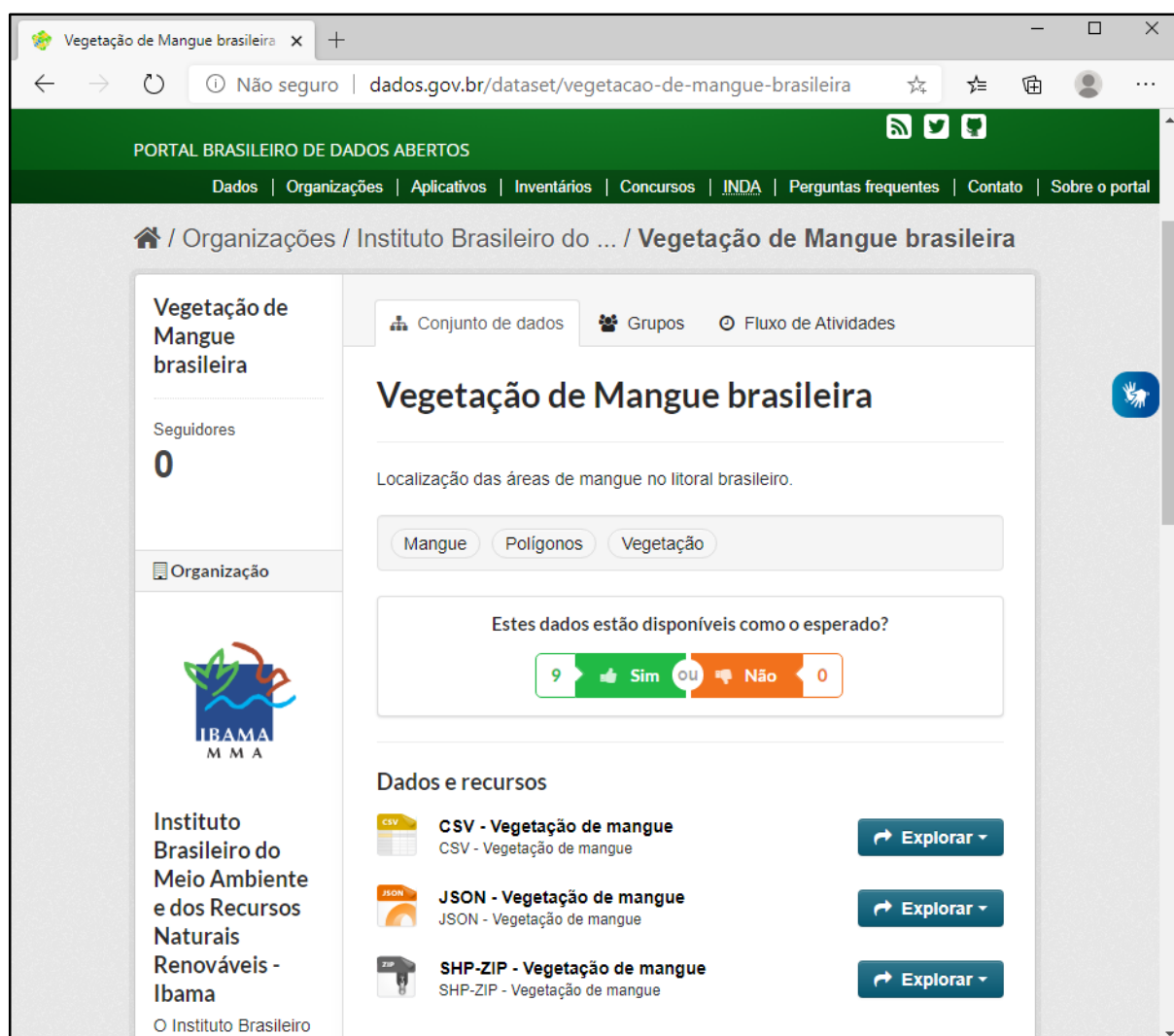


Figura 36: Publicação do conjunto de dados “Vegetação de Mangue Brasileira” e sua distribuição “CSV - Vegetação de mangue” entre outras. Fonte: PORTAL BRASILEIRO DE DADOS ABERTOS

A Figura 37 apresenta a publicação original do conjunto de dados “Sistema Nacional de Informações Florestais - SNIF”. Na Figura 37-a são apresentados parte dos metadados descritivos, conforme a classificação de Miller (2011). Para esse conjunto de dados, há vinte e oito distribuições de conteúdos distintos. Conforme citado na seção 2.2.1, o CKAN não impõe regras sobre como os conjuntos de dados devem ser organizados. Sendo assim, na Figura 37-b são apresentados cinco das vinte e oito distribuições, destacando-se a distribuição “Desmatamento - Bioma - estados - PMDBBS”, que foi triplicada e utilizada por LigADOS.



(a) Parte inicial da publicação do conjunto de dados “Sistema Nacional de Informações Florestais - SNIF” Fonte: PORTAL BRASILEIRO DE DADOS ABERTOS



(b) Cinco das vinte e oito distribuições, destacando-se a distribuição “Desmatamento - Bioma - estados - PMDBBS”

Figura 37: Publicação original do conjunto de dados “Sistema Nacional de Informações Florestais - SNIF” Fonte: PORTAL BRASILEIRO DE DADOS ABERTOS

Na Figura 38-a é apresentada uma parte do *Grafo de Interligações Semânticas* da instância CKAN que contém o extrato do Portal Brasileiro de Dados Abertos. No grafo são apresentadas as distribuições de dois conjuntos de dados distintos, publicados por organizações distintas e que foram citadas no início da seção: “Desmatamento - Bioma - estados - PMDBBS” e “Vegetação de Mangue”. As duas utilizam o tesauro “aims-agrovoc”. Apesar de pertencerem a conjuntos de dados distintos, ambas utilizam quatro conceitos em comum: “Vegetação”, “Mato”, “Floresta tropical” e “Floresta de regiões húmidas”⁵². Porém, somente a distribuição “Desmatamento - Bioma - estados - PMDBBS” utiliza o conceito “Terra de pastagem”, conforme destacado na imagem. Na Figura 38-b os conceitos utilizados foram removidos da visualização, utilizando recursos nativos de navegação em grafo. Mesmo assim, ainda é possível observar que ambas as distribuições possuem interligação, pois tratam do mesmo tema.

Cabe ressaltar que as interligações visualizadas na Figura 38, e estabelecidas por LigADOS, não são visualizadas sem seu uso. Em suas publicações originais, no Portal Brasileiro de Dados Abertos, não há indícios de relações entre as distribuições dos conjuntos de dados. Não se observam etiquetas em comum, nem descrições coincidentes que levem a considerar que haja algum tipo de relação entre eles. Mais que isso, não há metadados coincidentes que façam com que o motor de busca nativo do CKAN os retorne como resultado de uma mesma consulta.

⁵² Grafia original utilizada pelo tesauro.

Supressão de Vegetação” e sua distribuição “Autorização de Supressão de Vegetação” publicado pelo “Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis – Ibama”. Embora os rótulos de ambas as publicações levem a uma mesma situação, isto é, de remoção de vegetação, elas não estão interligadas quando se observam suas publicações originais. De fato, a primeira considera remoções de vegetação não autorizadas; enquanto a segunda trata de remoções autorizadas. Contudo, seria possível analisar autorizações de supressão de vegetação em locais de conservação federal? Inicialmente, pode-se considerar uma análise viável, visto que as duas distribuições de dados utilizam mais de um conceito em comum de um mesmo tesauro. Por exemplo, os conceitos de “Terra de pastagem” e “Floresta tropical” são pontos em comum, entre outros. Essa observação funciona como um sinalizador para analisar as instâncias de dados e responder à questão supramencionada. A Figura 39 apresenta os conjuntos de dados, as distribuições e as interligações que os une.

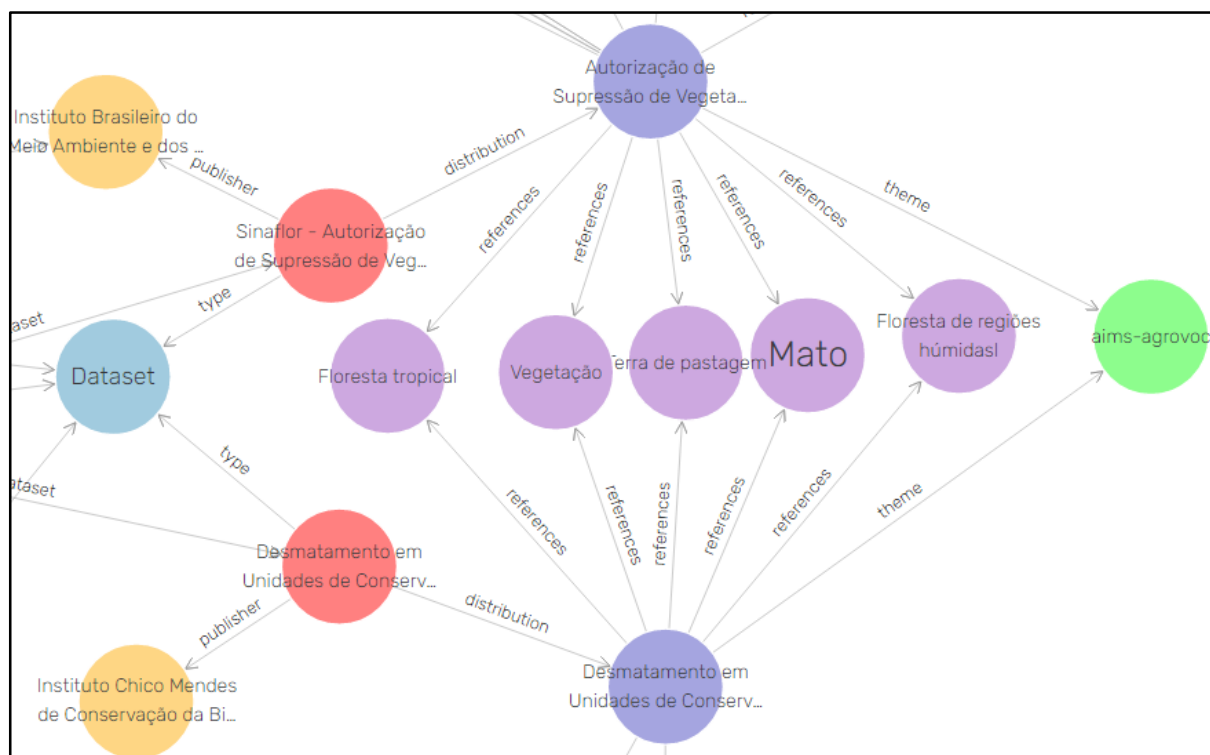


Figura 39: Interligações entre distribuições de conjuntos de dados que podem sinalizar possíveis pontos de interesse de análises

Uma outra possibilidade de análise envolve a compreensão dos assuntos contidos nas distribuições dos conjuntos de dados, sem a necessidade de extraí-los do ambiente do portal. Isso é possível graças aos metadados no nível de registro de dados, representados pelas classes e/ou conceitos dos vocabulários, ontologias e tesauros. Metadados no nível de registro de dados também fornecem as perspectivas pelas quais as distribuições dos conjuntos de dados podem ser analisadas. Na Figura 40 pode-se observar que a distribuição

“Autorização de Supressão Vegetal” trata de assuntos correlatos a “Florestas Tropicais” assim como a distribuição “Desmatamento em Unidades de Conservação”. No entanto, somente a primeira distribuição correlaciona o conceito “Florestas Tropicais” definido pelo tesauro “aims-agrovoc”, à classe “Unidades Federativas do Brasil” definida na ontologia “dbpedia-ontology” e ao conceito “redes rodoviárias” definido no tesauro “eionet-gemet”. Essa observação associada a ligação a outra distribuição de dados, publicada por outra organização, suscita novos questionamentos tal como: “Em qual escala de espaço e tempo têm sido autorizadas supressões de florestas tropicais, coincidentes ou próximas a áreas onde ocorreram desmatamentos desse tipo de vegetação, sob o interesse de construção de novas redes rodoviárias?”. Enfim, as possibilidades de análise surgem apoiadas pelo benefício da união de vários níveis de metadados em um mesmo grafo, que é disponibilizado para navegação e exploração de forma interativa. A interatividade aqui tem papel fundamental, pois permite que o consumidor de dados acesse a página da publicação do conjunto de dados ou a própria distribuição a partir do nó do grafo, sem novas buscas na interface do portal.

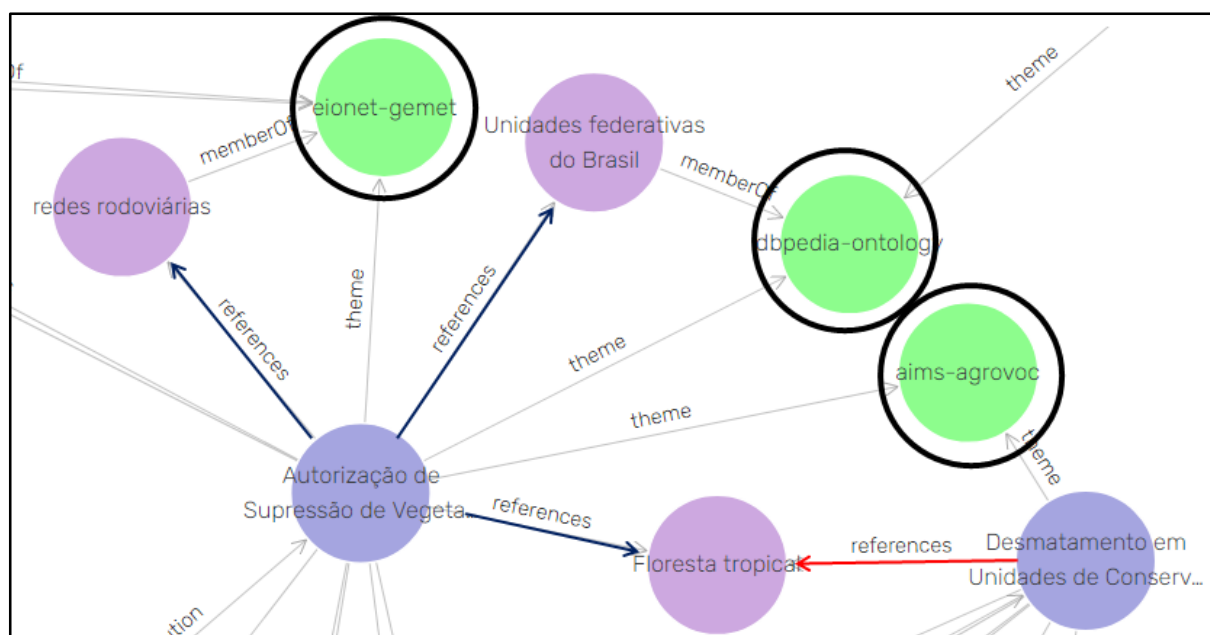


Figura 40: Possibilidades de análise pela compreensão dos assuntos contidos nas distribuições dos conjuntos de dados

Até este ponto foram exemplificadas as interligações e análises por meio de grandes tesouros multilinguais. Entretanto, é natural que alguns conjunto de dados façam uso de ontologias específicas de um domínio, que foram construídas alinhadas a um processo de negócio em particular. Neste contexto, em outro ponto do mesmo *Grafo de Interligações Semânticas*, observa-se a ontologia “vocab-loa”, utilizada pela distribuição “Orçamento 2018 em RDF” pertencente ao conjunto de dados “Orçamento Federal - Orçamento

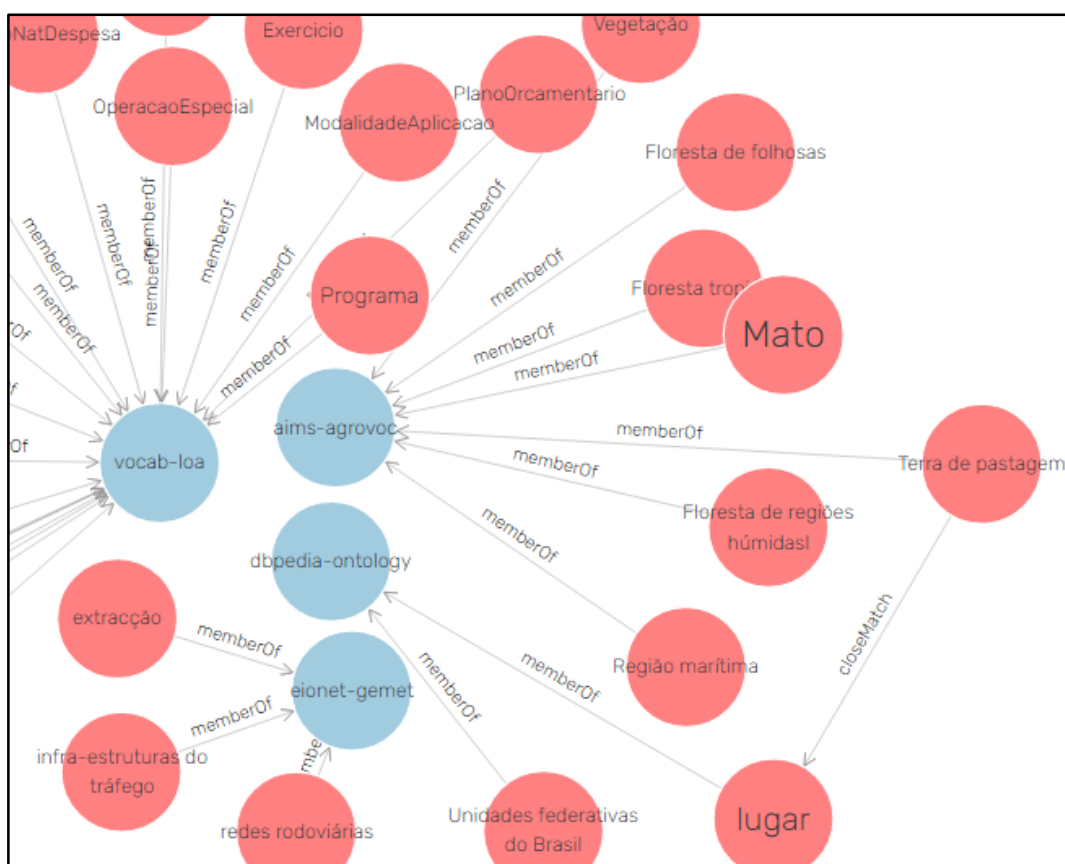


Figura 42: Ontologias “vocab-loa”, “dbpedia-ontology” e os tesouros “aims-agrovoc” e “eionet-gemet” e as classes/conceitos utilizadas pelas distribuições de dados

5.3. Portal de Dados Abertos sobre Agrotóxicos

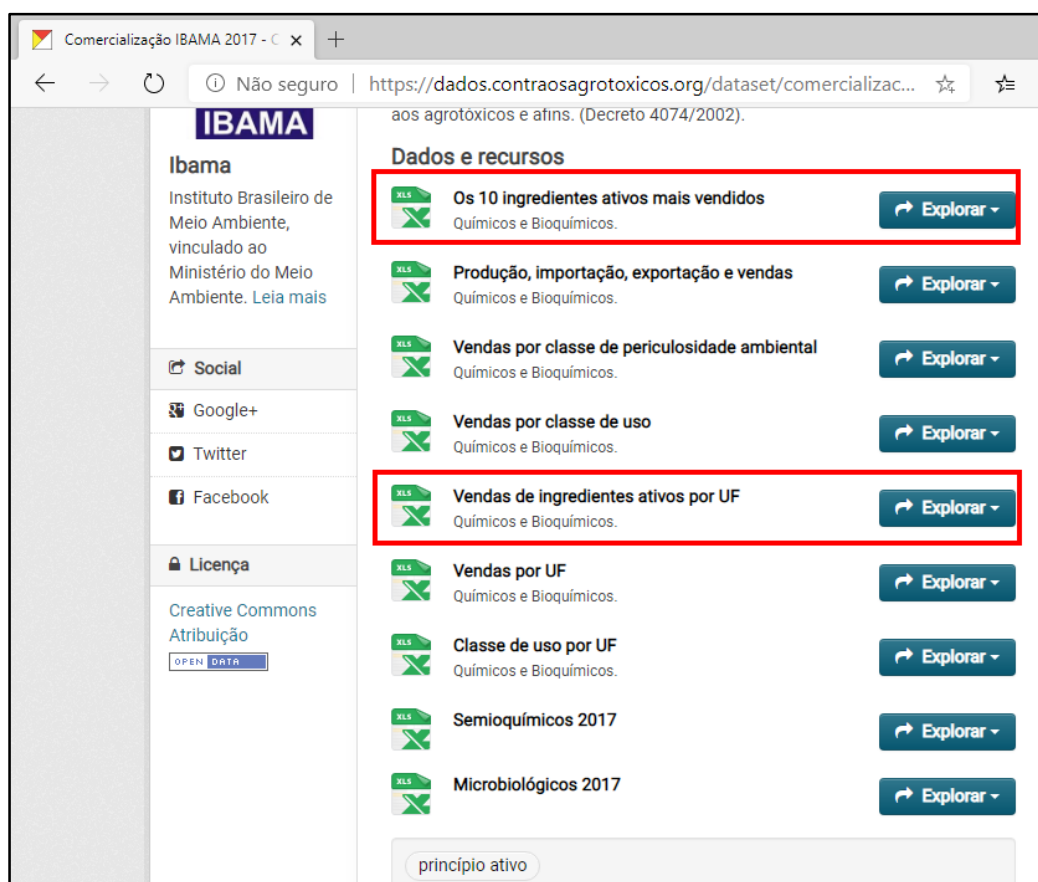
O Portal de Dados Abertos sobre Agrotóxicos é uma instância CKAN, desenvolvida no Programa de Pós-Graduação em Informática da UFRJ, como resultado do projeto “Observatório de Atenção Permanente ao Uso de Agrotóxicos – Portal de Informações Interligadas sobre Agrotóxicos e seus Efeitos sobre a Saúde e Meio Ambiente”, financiado pela FAPERJ no Edital Ext-Pesq n. E-26/111.544/2012⁵³.

Dentre as distribuições pré-processadas, duas delas pertencem a um mesmo conjunto de dados. Destarte, naturalmente compartilham quase todos os metadados, diferindo apenas os nomes, já que o publicador utilizou a mesma descrição para ambas. Assim, o conjunto de dados “Comercialização Ibama 2017” tem nove distribuições, das quais duas foram triplificadas e anotadas para o experimento: “Os 10 ingredientes ativos mais vendidos” e “Vendas de ingredientes ativos por UF”. A Figura 43 apresenta a publicação original.

⁵³ <https://dados.contraosagrototoxicos.org/about>



(a) Parte inicial dos metadados do conjunto de dados “Comercialização IBAMA 2017”



(b) Distribuições do conjunto de dados “Comercialização IBAMA 2017”, destacadas as que foram triplicadas, anotadas e utilizados no experimento

Figura 43: Publicação do conjunto de dados “Comercialização IBAMA 2017”. Fonte: PORTAL DE DADOS ABERTOS SOBRE AGROTÓXICOS

ligam a “Vendas de ingredientes ativos por UF”. O que não é recíproco para esse último, haja vista que a maior partes dos conceitos não a liga a primeira distribuição. LigADOS proporciona esse prático entendimento sobre o escopo de dados publicados.

Na Figura 45 é apresentado outro extrato das interligações semânticas. O lado esquerdo da figura apresenta as classes/conceitos utilizados pela distribuição “Classificação do IARC – Perigos do Câncer”. O lado direito apresenta as classes/conceitos utilizados pela distribuição “Ingredientes Ativos” do conjunto de dados “Agrofit”. Utilizando as publicações originais, os conjuntos de dados e suas distribuições aparecem como resultados de uma pesquisa somente se o termo “agrotóxico” isoladamente for utilizado como *string* de busca. Em se tratando de um portal de dados sobre agrotóxicos, parece bastante óbvio que os dados publicados de alguma forma façam referência a agrotóxicos. Sendo assim, não há clareza sobre o motivo pelo qual ambos aparecem como resultado na mesma busca. Neste sentido, cabe ressaltar que 60% dos conjuntos de dados publicados originalmente no portal respondem positivamente para uma busca pelo termo “agrotóxico”. Assim, LigADOS mostra sua outra faceta: transparência no porquê os dados estão interligados. No exemplo em questão, embora ambas as distribuições utilizam inúmeras classes/conceitos de “ChEBI Ontology”, é possível identificar quais exatamente conectam os dados. Assim, podem ser respondidas questões tais como: “Quais as classificações de ingredientes ativos que ainda estão sendo utilizados na agricultura e que podem ter algum tipo de relação com o desenvolvimento de câncer em seres humanos?”; e “Os agrotóxicos utilizados na agricultura, que não têm conexões com aqueles já identificados com algum tipo de relação com o desenvolvimento de câncer, já foram testados e realmente não têm relações?”. Na Figura 45 a elipse destaca as interligações entre os dados supramencionados.

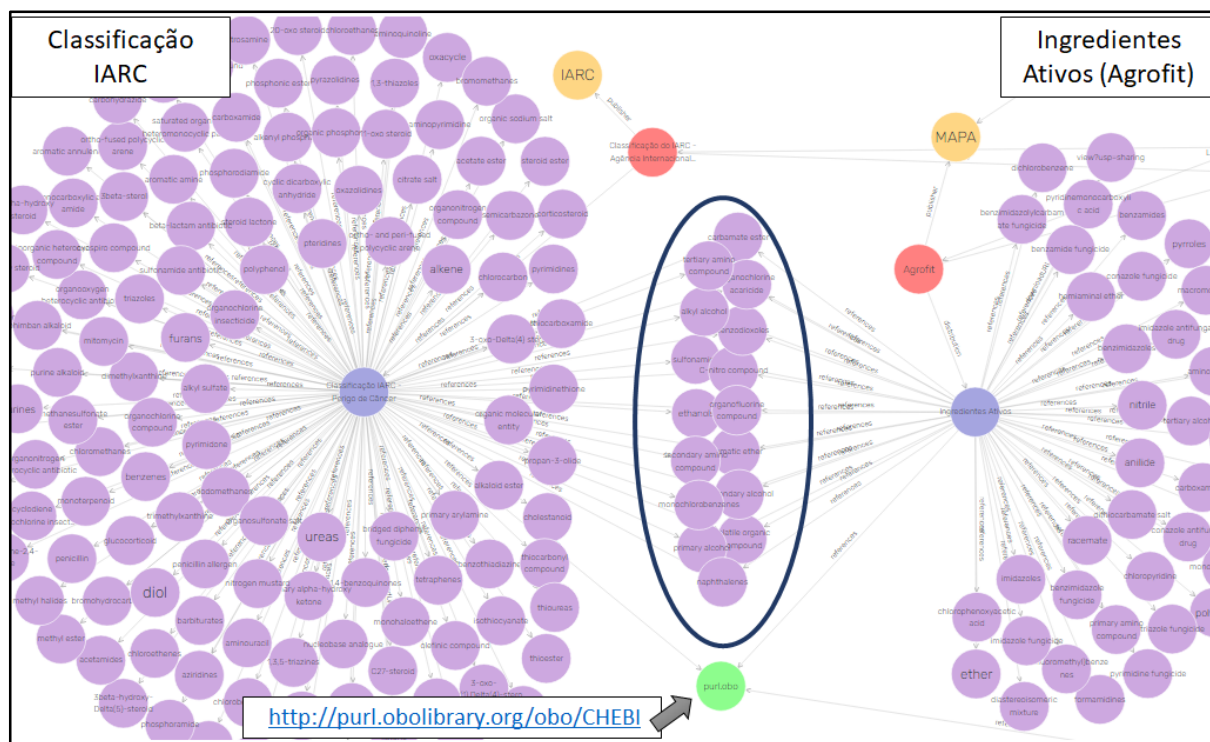


Figura 45: Duas distribuições de dados que utilizam “ChEBI Ontology” com destaque para as classes/conceitos que ligam ambas as distribuições.

O Apêndice D apresenta um extrato do *Grafo de Interligações Semânticas* serializado em N-Triples para a instância CKAN que contém um extrato do Portal de Dados Abertos sobre Agrotóxicos.

Por fim, uma outra forma de visualizar as interligações semânticas é viabilizada pelo Grafo de Etiquetas Semânticas. Diferente das interligações apresentadas até o momento, que apresentam as interligações pelas perspectivas das distribuições de dados, nesse grafo as interligações são apresentadas no nível de conjuntos de dados. Assim, as etiquetas semânticas em comum — geradas por LigADOS como uma extensão a uma abordagem existente (vide seção 4.1.3) — são os pontos de interconexões. A Figura 46 apresenta um extrato do grafo gerado, no qual a etiqueta “benzimidazoles” é o caminho de interligação entre os conjuntos de dados “Agrofit” e “PARA – Programa de Análise de Resíduos de Agrotóxicos em Alimentos”. Esse último conjunto de dados ainda é interligado a “Comercialização IBAMA 2017” pelas etiquetas “Piretrina sintética” e “Agroquímicos”.

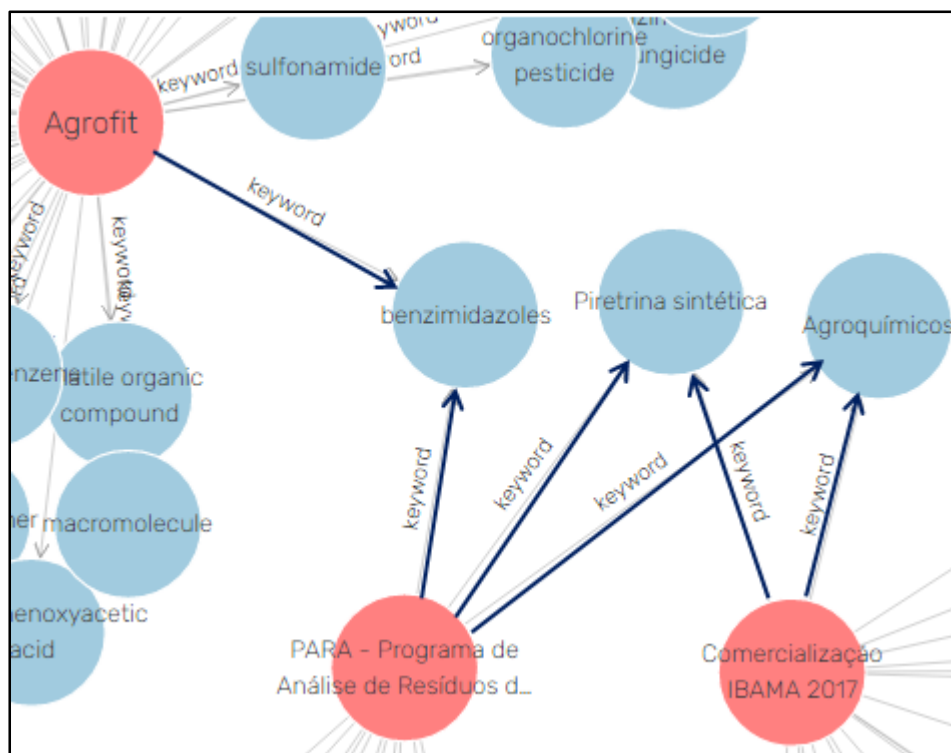


Figura 46: Etiquetas semânticas que interligam conjuntos de dados publicados na instância de portal

5.4. Discussão

Nas seções 5.2 e 5.3 foram apresentadas as formas pelas quais LigADOS-Portal estabelece e publica interligações semânticas indiretas entre distribuições de dados, publicados em portais de dados abertos construídos sobre o CKAN. A publicação das interligações como dados conectados fornece os meios necessários para que sejam efetuadas buscas, consultas e navegação pelas interligações, permitindo a inferência de novas informações, assim como compreensão dos significados das instâncias de dados.

Sem o uso de LigADOS-Portal, são encontradas relações não-semânticas entre os conjuntos de dados publicados, isto é, elementos de metadados coincidentes dentre as publicações de conjuntos de dados em uma instância de portal. A organização que publicou o conjunto de dados, o título ou algumas palavras específicas utilizadas na descrição do conjunto de dados, ou mesmo nomes de distribuições podem possuir expressões textuais coincidentes. Sem o uso de LigADOS-Portal ou de outra abordagem de enriquecimento de etiquetas, etiquetas também podem ser consideradas elementos de metadados coincidentes, visto que representam textos simples por não estarem vinculadas a vocabulários, ontologias ou tesouros. Apesar disso, as relações não-semânticas são utilizadas pelo motor de busca nativo do CKAN para gerar resultados em sua interface de uso.

Com o uso de LigADOS-Portal, as relações não-semânticas continuarão a existir, no entanto conjuntos de metadados semânticos, tanto no nível de conjunto de dados, como no

nível de registros de dados, são gerados. Os metadados semânticos no nível de conjuntos de dados são representados por etiquetas semânticas. Os metadados semânticos no nível de registro de dados são representados pelas interligações semânticas indiretas que são geradas e publicadas.

LigADOS-Portal publica as interligações semânticas indiretas na instância do portal de onde foram geradas, fornecendo capacidades de buscas e consultas seja por humanos, seja por aplicações de *software*.

Em alguns casos, o uso de LigADOS-Portal aumenta o número de interligações semânticas indiretas. Para o Portal Brasileiro de Dados Abertos foi possível observar um aumento de 300% após o uso de LigADOS-Portal. No entanto, o percentual de aumento não é o aspecto mais importante. O aspecto principal diz respeito ao fato de que os novos pontos de interligação passam a considerar as instâncias de dados em RDF, ou seja, os recursos cujos significados foram definidos pelo uso de vocabulários, ontologias ou tesauros, desconsiderando as relações não-semânticas.

O aumento significativo de interligações ocorre quando os conjuntos de dados publicados em uma instância de portal utilizam os mesmos vocabulários, ontologias ou tesauros. Esse fator alinha-se às boas práticas de publicação de dados na Web, tais como reutilizar vocabulários, preferencialmente os padronizados, de forma que a interoperabilidade e consenso entre os publicadores e consumidores de dados sejam reforçados (LÓSCIO et al., 2017). Neste sentido, LigADOS-Portal também corrobora para reutilização de vocabulários, ontologias e tesauros ao publicar, na instância do portal, uma biblioteca contendo as classes e/ou conceitos, que representam os elementos e/ou termos, que foram utilizados para anotação semântica das triplas dos conjuntos de dados em RDF publicados.

LigADOS-Portal também executa parte do processo de FAIRficação, pois define um modelo semântico para os metadados, assim como os transforma para o formato legível por máquinas. Outrossim, os metadados semânticos processáveis por máquina são utilizados pelo módulo LigADOS-FDP para serem alinhados ao esquema de níveis de metadados especificado para um FDP.

Neste contexto, cabe ressaltar que LigADOS-FDP fornece os metadados semânticos até o nível de registro de dados especificado para um FDP. A forma como LigADOS-FDP fornece metadados no nível de registro de dados permite a exploração das interligações semânticas indiretas entre distribuições de dados, de diferentes conjuntos de dados, publicados em diferentes portais. Esse benefício é mais uma consequência da reutilização de vocabulários, ontologias e tesauros para anotação de triplas de conjuntos de dados em RDF.

Em outros casos, LigADOS-Portal pode gerar um número reduzido de interligações semânticas indiretas. Às vezes, menor do que o número de relações não-semânticas. O uso

de LigADOS-Portal no Portal de Dados Abertos Sobre Agrotóxicos gerou esse tipo de situação: poucas interligações semânticas indiretas. De fato, entre alguns conjuntos de dados existiam relações não-semânticas, antes do uso de LigADOS-Portal. Em contraste, com o uso de LigADOS-Portal nenhuma interligação semântica indireta foi estabelecida. Significa dizer que os conjuntos de dados em RDF, publicados na instância de portal, podem até ter alguma relação. Porém, do ponto de vista semântico, os conjuntos de dados não utilizaram vocabulários, ontologias ou tesauros em comum. Isso pode ocorrer por dois motivos: (i) os responsáveis pelos conjuntos de dados *não* fizeram anotação semântica dos recursos em RDF; ou (ii) os responsáveis pelos conjuntos de dados fizeram anotação semântica dos recursos em RDF, mas utilizaram vocabulários, ontologias ou tesauros distintos.

O fato de conjuntos de dados utilizarem diferentes vocabulários, ontologias ou tesauros não necessariamente significa que tratem de assuntos diferentes. A utilização de técnicas de alinhamento de vocabulários, ontologias ou tesauros, no nível de esquema, no nível de instâncias ou de ambos os níveis, como ferramenta complementar à abordagem, pode contribuir para a geração de interligações semânticas indiretas entre conjuntos de dados, já que pela navegação será possível alcançar conjuntos de dados que tratem dos mesmos assuntos.

Destarte, diante dos casos de aplicação, LigADOS, por meio do módulo LigADOS-Portal, mostra-se como uma abordagem apta a gerar interligações entre dados, a partir das quais é possível acesso às publicações originais desses dados. Mais que isso, LigADOS democratiza o entendimento de dados modelados em RDF, a partir da própria plataforma do portal, ao fornecer uma interface visual e interativa que provê uma camada de metadados no nível de registro de dados.

6. CONCLUSÃO

Este trabalho de pesquisa apresentou LigADOS, uma abordagem para interligação de conjuntos de dados em plataformas de dados abertos, por meio de tecnologias da Web Semântica. LigADOS utiliza os princípios de dados conectados como estratégia para dar suporte a esta abordagem, contribuindo para uma melhor exploração das possibilidades de acesso que a representação em RDF proporciona. A principal estratégia consiste em definir um nível de expressividade semântica para os conjuntos de dados publicados, de forma a estabelecer e expor interligações existentes e não explícitas entre eles.

Os resultados indicam que as interligações são estabelecidas quando os conjuntos de dados em RDF fazem uso de vocabulário, ontologias ou tesauros em comum. Mais que isso, os resultados indicam que conjuntos de dados publicados por diferentes organizações e com objetivos distintos podem ter interligações. As interligações ficam ocultas em função, principalmente, da forma como as plataformas de dados abertos atualmente tendem a tratar dados conectados e, principalmente, por causa da falta de metadados semânticos. Os resultados também indicam que conjuntos de dados que, inicialmente, pareciam tratar de assuntos correlatos, ao se observar os metadados catalogados por ocasião de suas publicações, podem não ter interligações semânticas quando se considera as instâncias dos dados.

Ademais, LigADOS gera subsídios para o esquema de metadados conforme especificado para FAIR *Data Points* (FDP). Como metadados no nível de registros de dados são fornecidos ao FDP, LigADOS também possibilita que sejam observadas as ligações entre conjuntos de dados publicados em diferentes portais, cujos metadados até o nível de registro de dados estejam cadastrados no FDP.

Por fim, o protótipo da aplicação mostrou que é possível integrar plataformas de dados abertos a *triplestores* com interfaces para navegabilidade de grafos em RDF. Ressalta-se que LigADOS foi desenvolvida pautada em metadados com características propostas pela iniciativa FAIR e por boas práticas de publicação de dados na Web.

6.1. Contribuições

A solução descrita nesta dissertação foca na extração das classes e/ou conceitos dos elementos e/ou termos utilizados para anotações semânticas das distribuições dos conjuntos de dados, modeladas em RDF, e publicados em uma instância de portal de dados abertos. A partir desse ponto, LigADOS gera metadados no nível de registro de dados. Por meio desses metadados são estabelecidas as interligações semânticas entre os conjuntos

de dados, que são persistidas em um grafo RDF. Todas as interligações estabelecidas são franqueadas ao público em geral, pela disponibilização de uma interface para exploração e navegação do grafo na própria instância de portal de dados, além de, naturalmente, estarem disponíveis e aptas para coleta e processamento por máquina.

Adicionalmente aos metadados no nível de registro de dados, a solução descrita gera metadados no nível de conjuntos de dados ao gerar um grafo RDF de etiquetas semânticas extraídas a partir da classificação e/ou conceitualização das instâncias de dados estruturadas na forma de triplas.

Soma-se a isso a geração de subsídios para composição dos esquemas de metadados conforme especificações existentes para um FDP.

Em suma, as contribuições diretas desta dissertação são:

- apoio à geração de metadados semânticos em dois níveis: (i) registro de dados e (ii) conjuntos de dados;
- apoio à geração de interligações semânticas entre conjuntos de dados abertos já publicados; e
- geração de subsídios para quatro de cinco níveis de esquemas de metadados especificados para FDP.

Ademais, mais duas contribuições indiretas podem ser observadas:

- comparativo de plataformas de dados abertos, de código aberto, sob uma perspectiva distinta daquelas verificadas na literatura científica até o momento;
- comparativo entre o perfil (*profile*) de metadados proposto por LigADOS e o perfil de níveis de metadados previstos para um FDP.

Os produtos gerados por esta dissertação são:

- um protótipo que utiliza dados extraídos de dois portais de dados abertos em produção para utilização na demonstração da aplicação e LigADOS;
- publicação de um estudo sobre os princípios de gerenciamento de dados em um artigo apresentado no *Workshop on Conceptual modeling, Ontologies and Metadata Management for FAIR Data*, parte do *38th International Conference on Conceptual Modeling* (ER 2019) (FIGUEIREDO et al., 2019);
- publicação dos detalhes de LigADOS um artigo apresentado no *14th International Conference on Metadata and Semantics Research* (MTSR 2020) (FIGUEIREDO; CORDEIRO; CAMPOS, 2020).

6.2. Limitações

Uma limitação significativa da abordagem proposta nesta dissertação foi a necessidade de estabelecimento de uma premissa para que LigADOS tivesse seu espaço de experimentação. A premissa dos conjuntos de dados estarem modelados em RDF, associada ao acesso limitado a esses conjuntos de dados em RDF a partir de instâncias de portais reais demandou a necessidade de execução prévia de processos de triplificação e anotação de conjuntos de dados reais. Por isso, foi necessária a pesquisa, análise e seleção de vocabulários, ontologias e tesouros de domínios específicos a serem utilizados nas anotações semânticas. Essas etapas foram realizadas com base nos trabalhos científicos anteriores desenvolvidos no grupo GRECO, assim como na busca por literatura científica que citasse e/ou utilizasse vocabulários controlados. Todavia, o risco de geração de dados simulados, desconexos dos dados reais, foi minimizado ao concentrar a busca por conjuntos de dados de domínios afins, de forma a se especializar nesses domínios.

No que tange aos experimentos, não foram definidas ou aplicadas métricas para validação do alinhamento da abordagem com os princípios FAIR, focando-se, primordialmente, nas discussões e possibilidades de implementação de parte dos princípios.

Com relação à implementação do protótipo, não foram considerados aspectos de performance, nem desenvolvidas rotinas de atualização automática dos grafos semânticos gerados, que monitorem quando conjuntos de dados forem atualizadas ou novos conjuntos forem adicionados. Outrossim, o protótipo fez uso de soluções aplicáveis ao CKAN. No entanto, o código pode ser adaptado para atender outras plataformas de dados abertos.

6.3. Trabalhos Futuros

A exemplo da nuvem LOD, dados conectados podem se tornar bastante complexos de exploração, mesmo com apoio de recursos visuais e interativos. Dada a quantidade crescente de abertura de conjuntos de dados de diversos segmentos, a quantidade de interligações semânticas, mesmo em uma única instância de portal de dados, pode tornar complexa a exploração e navegação visual. Em função disso, a proposta futura envolve a segmentação do *Grafo de Interligações Semânticas* em subgrafos semânticos. A proposta envolve a utilização de medidas de centralidade aplicadas a grafos para identificação de classes e/ou conceitos a partir dos quais um subgrafo pode ser gerado. Sendo assim, as interligações semânticas poderão ser inicialmente explorados de diversos pontos distintos e não de um único ponto (atualmente os pontos iniciais representam o portal e o tipo “`dcat:Dataset`”). Medidas de centralidade que quantifiquem os nós que mais servem como interligação entre

conjuntos de dados, excluindo-se sua própria tipificação como “`dcat:Dataset`” ou “`dcat:Distribution`” poderiam ser o cerne de subgrafos semânticos.

De modo a observar o crescimento do *Grafo de Interligações Semânticas*, faz-se necessária a execução de mais experimentações, validações mais robustas envolvendo, inclusive, usuários da abordagem.

Por fim, outro aspecto que pode ser estendido, é a complementação do conjunto de metadados que subsidia um FDP. Atualmente, LigADOS fornece os elementos de metadados obrigatórios para cada nível. Poucos itens opcionais são fornecidos. A intenção é cobrir a maior parte dos elementos de metadados previstos pela especificação de um FDP.

REFERÊNCIAS

ACHICHI, M. et al. **Linking and disambiguating entities across heterogeneous RDF graphs.** Journal of Web Semantics, 55, 2019. p.108-121.

ALVES, M. **Como escrever teses e monografias - Um roteiro passo a passo.** Rio de Janeiro: Campus, 2003.

AZMY, M. et al. **Farewell freebase: Migrating the SimpleQuestions dataset to DBpedia.** In: Proceedings of the 27th International Conference on Computational Linguistics, 2018. p. 2093-2103.

BELLOZE, K. T. et al. **An evaluation of annotation tools for biomedical texts.** In: CITESEER. ONTOBRAS-MOST. 2012. p. 108–119.

BERENDS, J., et al. **Recommendations for Open Data Portals: from setup to sustainability.** European Data Portal. Luxembourg: Publications Office of the European Union, 2020

BIZER, C., HEATH, T., BERNERS-LEE, T. **Linked data: The story so far.** In: Semantic services, interoperability and web applications: emerging concepts. 2011. p. 205-227. IGI Global.

CARABALLO, A. A. M., et al. **Automatic creation and analysis of a linked data cloud diagram.** In: International Conference on Web Information Systems Engineering. 2016. p. 417-432. Springer, Cham.

CASTRO, B. P. C. de. **Semantic Enrichment and Exploration: An approach developed for dataset tags of Open Data Portals.** Dissertação (Mestrado em Informática) – Universidade Federal do Rio de Janeiro, 2019.

ČEBIRIĆ, Š., et al. **Summarizing semantic graphs: a survey.** The VLDB Journal v. 28 n.3, p. 295-327, 2019.

CIMIANO, P. **Ontologies.** Springer US; 2006.

CKAN + DCAT - CKAN Extensions. Disponível em: <<https://extensions.ckan.org/extension/dcat/>>. Acesso em 18 fev. de 2020.

CKAN - The open source data portal software. Disponível em: <<https://ckan.org/>>. Acesso em: 16 fev. de 2020.

User guide CKAN - 2.9.0a documentation. Disponível em: <<https://docs.ckan.org/en/latest/user-guide.html>>. Acesso em: 08 mar. de 2020.

CORDEIRO, K. **aDapTA: Adaptive Approach for Information Integration to Support Decision Making in Complex Environments.** Tese (Doutorado em Informática) – Instituto de Matemática, Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2015.

CORDEIRO, K. de F., et al. **An approach for managing and semantically enriching the publication of Linked Open Governmental Data**. In: Proceedings of the 3rd workshop in applied computing for electronic government (WCGE), SBBD. 2011. p. 82-95.

CORREA, A. S., ZANDER, P., SULVA, F. S. C.: **Investigating open data portals automatically: a methodology and some illustrations**. In: Proceedings of 19th Annual International Conference on Digital Government Research Governance in the Data Age. DG.O 2018. p. 1–10.

COSTA, L. R., et al. **Ckan user guide**. Ibict, 2017.

CROSAS, M., KING, G., HONAKER, J., SWEENEY, L. (2015). **Automating Open Science for Big Data**. In: The ANNALS of the American Academy of Political and Social Science. 2015. p. 260–273.

CROSS, V. **Semantic similarity: a key to ontology alignment**. In: Proceedings of the 13th International Workshop on Ontology Matching. 2018. p. 61-65.

Data - W3C. Disponível em: <<https://www.w3.org/standards/semanticweb/data>>. Acesso em 16 fev. de 2020.

DCAT - Data Catalog Vocabulary - Version 2. Disponível em: <<https://www.w3.org/TR/vocab-dcat-2/>>. Acesso em 25 ago. de 2020.

DOJCHINOVSKI, M. et al. **DBpedia Links: The Hub of Links for the Web of Data**. In: Proceedings of the SEMANTiCS Conference (SE-MANTiCS 2016). 2016.

DRAGAN, A. **DCAT Application Profile for data portals in Europe**. Disponível em: <<https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe>>. Acesso em 28 set. de 2020.

DSpace 6.x Documentation. Disponível em: <<https://wiki.lyrasis.org/display/DSDOC6x/DSpace+6.x+Documentation>>. Acesso em 30 jun. de 2020.

EPSI PLATFORM. **European Public Sector Information (PSI) Platform Topic Report No. 2016/03**. February 2015.

FAIRDataPoint-Spec - FAIR Data Point design specification. Disponível em: <<https://github.com/FAIRDataTeam/FAIRDataPoint-Spec>>. Acesso em 25 ago. de 2020.

Features | The Dataverse Project - Dataverse.org. Disponível em: <<https://dataverse.org/software-features>>. Acesso em 17 mar. de 2020.

FERREIRA, A. B. de H. **Miniaurélio: o minidicionário da língua portuguesa**. 6. ed. Curitiba: Positivo, 2004. ISBN: 857472416-5

FIGUEIREDO, G. B. de, MOREIRA, J.L.R., CORDEIRO, K. de F., CAMPOS, M. L. M. **Aligning DMBOK and open government with the FAIR data principles**. In: International Conference on Conceptual Modeling. 2019. p. 13-22. Springer, Cham.

FREIRE, N. et al. **Aggregation of cultural heritage datasets through the Web of Data**. Procedia Computer Science 137, 120-126, janeiro 2018.

GABRIEL JUNIOR, et al. **Acesso aberto a dados de pesquisa no Brasil: mapeamento de repositórios, práticas e percepções dos pesquisadores e tecnologias**. Ci.Inf., Brasília, DF, v.48

n.3 (Supl.), p.87-101, set./dez. 2019. Disponível em: <<http://revista.ibict.br/ciinf/article/view/4958>>. Acesso em 03 jun. de 2020.

GOMES, R. do V. A., CASANOVA, M. A., LOPES, G. R., PAES LEME, L. A. P. **A metadata focused crawler for linked data**. In Proceedings of the 16th International Conference on Enterprise Information Systems-Volume 2. 2014. p. 489-500. SCITEPRESS-Science and Technology Publications, Lda.

GRAPHDB - ONTOTEXT. Disponível em: <<https://www.ontotext.com/products/graphdb/>>. Acesso em 22 jun. de 2020.

GUARINO, N., OBERLE, D., STAAB, S. **What Is an Ontology?**. 2009. 10.1007/978-3-540-92673-3_0.

HARTIG, O., THOMPSON, B. **Foundations of an Alternative Approach to Reification in RDF**. arXiv preprint arXiv:1406.3399v2. 2019. Disponível em: <<https://arxiv.org/pdf/1406.3399.pdf>>.

HENDERSON, D., EARLEY, S. **The DAMA Guide to the Data Management Body of Knowledge (DAMA-DMBOK2)**. 2nd edition. New Jersey: Technics Publications, 2017.

HENDLER, J., BERNERS-LEE, T., MILLER, E. **Integrating Applications on the Semantic Web**. In Journal of the Institute of Electrical Engineers of Japan, Vol. 122, n.10, 2002. p. 676-680.

JACOBSEN, A., et al. **A generic workflow for the data FAIRification process**. In Data Intelligence, n. 2, 2020. p. 56–65. doi: 10.1162/dint_a_00028

KETTOUCH, M., LUCA, C., HOBBS, M. **SemiLD: mediator-based framework for keyword search over semi-structured and linked data**. In Journal of Intelligent Information Systems, Vol. 52, n.2, 2019. p. 311-335.

KIRSTEIN, F., et al. **Linked Data in the European Data Portal: A Comprehensive Platform for Applying DCAT-AP**. In Electronic Government. 2019. v. 11685. Springer, Cham. https://doi.org/10.1007/978-3-030-27325-5_15

KOLYVAKIS, P., KALOUSIS, A., & KIRITSIS, D. **Deepalignment: Unsupervised ontology matching with refined word vectors**. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, 2018. p. 787-798.

KUBITZA, D. O., BÖCKMANN, M., GRAUX, D. **SemanGit: A linked dataset from git**. In International Semantic Web Conference. 2019. p. 215-228. Springer, Cham.

LACHTIM, F. A., MOURA, A. M. de C.; CAVALCANTI, M. C. **Ontology matching for dynamic publication in semantic portals**. J. Braz. Comp. Soc., Campinas, v. 15, n. 1, p. 27-43, Mar. 2009. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-65002009000100004&lng=en&nrm=iso>. Acesso em 30 jun. de 2020. <https://doi.org/10.1007/BF03192575>.

LEME, L. A. P. P., et al. **Identifying Candidate Datasets for Data Interlinking**. In: Daniel F., Dolog P., Li Q. (eds) Web Engineering. ICWE 2013. Lecture Notes in Computer Science, vol 7977. Springer, Berlin, Heidelberg.

LDP Implementations - W3C Wiki. Disponível em: <<https://www.w3.org/TR/ldp/>>. Acesso em 18 fev. de 2020.

LIRA, M. A. B. de. **Uma Abordagem Para Enriquecimento Semântico de Metadados Para Publicação de Dados Abertos**. 2014. 95 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal de Pernambuco, 2014.

LÓSCIO, B. F.; BURLE, C.; CALEGARI, N. **Data on the Web best practices**. W3C Recommendation 31 January 2017, World Wide Web Consortium (W3C), Janeiro 2017. Disponível em: <<https://www.w3.org/TR/dwbp/>>. Acesso em 18 set. de 2020.

MARTINEZ-GONZÁLES, M.M., ALVITE-DIEZ, M.L. **The support of constructs in thesaurus tools from a Semantic Web perspective: Framework to assess standard conformance**. Computer Standards & Interfaces, v. 65, pp.79-91, 2019.

MARTINS, Y. C., MOTA, F. F. da, CAVALCANTI, M. C.. **DSCrank: a method for selection and ranking of datasets**. In: Research Conference on Metadata and Semantics Research. 2016. p. 333-344. Springer, Cham.

MILLER, S. J. **Metadata for digital collections: a how-to-do-it manual**. New York, NY: Neal-Schuman Publishers, 2011.

MILIĆ, P., VELJKOVIĆ, N. STOIMENOV, L. **Linked relations architecture for production and consumption of linksets in open government data**. In: Conference on e-Business, e-Services and e-Society. 2015. p. 212-222. Springer, Cham.

MOREIRA, A. **Tesauros e ontologias: estudo de definições presentes na literatura das áreas das Ciências da Computação e da Informação, utilizando-se o método analítico-sintético**. Dissertação (Mestrado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2003.

MOREIRA, J. L. R., et al. **Towards Findable, Accessible, Interoperable and Reusable (FAIR) Data Repositories: Improving a Data Repository to Behave as a FAIR Data Point**. Liinc em Revista, Rio de Janeiro, v.15, n.2, p.244-258, novembro 2019.
<https://doi.org/10.18617/liinc.v15i2.4817>

NEUMAIER, S., et al. **Data Integration for Open Data on the Web**. In: Reasoning Web. Semantic Interoperability on the Web. Reasoning Web 2017. LNCS, v. 10370. 2017. p. 1-28. Springer, Cham. https://doi.org/10.1007/978-3-319-61033-7_1

NEUMAIER, S., UMBRICH, J., POLLERES, A. **Lifting data portals to the web of data**. In LDOW@ WWW. 2017.

OLIVEIRA, M. I. S., et al. **Enabling a Unified View of Open Data Catalogs**. In ICEIS (2), pp. 230-239. 2016.

OMG (Object Management Group), **Business Process Model and Notation (BPMN)**, 2011. Disponível em: <<http://www.omg.org/spec/BPMN/2.0>>. Acesso em 08 jun. de 2020.

OPEN KNOWLEDGE FOUNDATION. **Open Data Handbook**. Disponível em: <<http://opendatahandbook.org/>>. Acesso em 07 set. de 2020.

OREN, E., et al. **What are semantic annotations?** Relatório técnico. DERI Galway, 9, 62. 2006.

PESCE, V., et al. **Setting up a global linked data catalog of datasets for agriculture**. In Research Conference on Metadata and Semantics Research. 2015. p. 357-368. Springer, Cham.

PIETRIGA, E., et al. **Browsing Linked Data Catalogs with LODAtlas**. In: ISWC 2018 - 17th International Semantic Web Conference. 2018. p.137-153. Monterey, United States. fhal-01827766

RDF - Semantic Web Standards. Disponível em: <<https://www.w3.org/RDF/>>. Acesso em 16 fev. de 2020.

REMY, L., et al. **Building an integrated enhanced virtual research environment metadata catalogue**. The Electronic Library, Vol. 37 No. 6, pp. 929-951, dezembro 2019.
<https://doi.org/10.1108/EL-09-2018-0183>

ROCHA, R. P., et al. **Acesso aberto a dados de pesquisa no Brasil: soluções tecnológicas: relatório 2018**. Porto Alegre, RS : UFRGS, 2018. Disponível em:
<<http://hdl.handle.net/10183/185126>>.

SALES, R. de; CAFÉ, L.. **Diferenças entre tesauros e ontologias**. Perspect. ciênc. inf., Belo Horizonte , v. 14, n. 1, p. 99-116, abril 2009 . Disponível em:
<https://www.scielo.br/scielo.php?pid=S1413-99362009000100008&script=sci_arttext&tlng=pt>. Acesso em 03 jul. de 2020.
<https://doi.org/10.1590/S1413-99362009000100008>.

SAYÃO, L., SALES, L. (2016). **Algumas considerações sobre os repositórios digitais de dados de pesquisa**. Informação & Informação. 21(2). 90, dezembro 2016.
<https://doi.org/10.5433/1981-8920.2016v21n2p90>

SCHMACHTENBERG, M., BIZER, C., PAULHEIM, H. **Adoption of the linked data best practices in different topical domains**. In: International Semantic Web Conference. 2014. p. 245-260. Springer, Cham.

SCHMIDT, D., TROJAHN, C., VIEIRA, R. **Analysing top-level and domain ontology alignments from matching systems**. In 11th International Workshop on Ontology Matching co-located with the 15th ISWC (OM 2016). 2016. Kobe, Japan.

SHINTAKU, M., MEIRELLES, R. **Manual do DSPAcE : administração de repositórios**. Salvador : EDUFBA, 2010. ISBN 978-85-232-0732-8

SKOS Simple Knowledge Organization System Reference. Disponível em:
<<https://www.w3.org/TR/skos-reference/>>. Acesso em 09 jun. de 2020.

SOUZA, R. A. T. **Predição de tags usando linked data: um estudo de caso no banco de dados Arquiografia**. Tese (Doutorado) – Universidade de São Paulo, 2013.

SPAHIU, B. **Profiling Linked Data**. Ph.D Thesis – Doctoral School in Computer Science, University of Milano, Bicocca, 2017.

SPARQL Query Language for RDF. Disponível em: <<https://www.w3.org/TR/rdf-sparql-query/>>. Acesso em 24 jun. de 2020.

STELLATO, A. **Dictionary, thesaurus or ontology? Disentangling our choices in the semantic web jungle**. In: Journal of Integrative Agriculture, 11(5). 2012. p.710-719.

TYGEL, A. **A Semantic Tags for Open Data Portals: Metadata Enhancements for Searchable Open Data**. Tese (Doutorado em Informática) – Instituto de Matemática, Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2016.

UBALDI, B. **Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives**. OECD Working Papers on Public Governance, No. 22. Paris: OECD Publishing, 2013. <http://dx.doi.org/10.1787/5k46bj4f03s7-en>

XIE, I., MATUSIAK, K. K. **Chapter 5 - Metadata**. **Discover Digital Libraries**. Elsevier, 2016. p.129–170.

WILKINSON, M. D. et al. **The FAIR Guiding Principles for scientific data management and stewardship**. *Scientific Data*, v. 3, 2016. Disponível em: <<https://www.nature.com/articles/sdata201618>>. Acesso em 29 jun. de 2020.

WILKINSON, M. D. et al. **Interoperability and FAIRness through a novel combination of Web technologies**. *PeerJ Computer Science*. 3:e11, 2017. Disponível em: <<https://peerj.com/articles/cs-110/>>. Acesso em 20 jul. de 2020.

APÊNDICES

APÊNDICE A — COMPARATIVO DE *PROFILES*: FAIR DATA POINT (FDP) E LigADOS

A comparação foi dividida de acordo com as camadas de metadados especificadas para um FDP. Ela abrange três de cinco camadas de metadados definidas para um FDP. As camadas comparados são: catálogo, conjunto de dados e distribuição. A camada de metadados do FDP não é gerada pela abordagem, conforme explicitado na seção 4.1.5. A camada de registro de dados, embora gerada pela abordagem, não é comparado, visto que não há um detalhamento dela nas especificação do FDP, conforme também explicado na seção 4.15.

Para cada item pertencente a uma das camadas abrangidas pela comparação foi indicado o alinhamento com o *profile* proposto e utilizado por LigADOS. Assim, cada elemento de metadados foi categorizado:

- Extraído do portal (E) → LigADOS extraiu o elemento de metadado da instância do portal pela execução do macroprocesso de “Extração de Metadados” (seção 4.1.1);
- Gerado pela abordagem (G) → LigADOS não extraiu o elemento de metadado pelo fato dele não estar disponível no conjunto de metadados cadastrados na instância do portal, mas, em atendimento à especificação de obrigatoriedade por parte do FDP, LigADOS gera o metadado como parte da abordagem e conforme detalhado na seção 4.1.5; e
- Não previsto (N) → LigADOS não utiliza o elemento de metadado e também não pode extraí-lo, pois não há um elemento correspondente entre o conjunto de metadados cadastrados na instância do portal.

Os Quadros 10, 11 e 12 apresentam os comparativos para os metadados das camadas de catálogo, conjunto de dados e distribuição especificados para um FDP e o alinhamento com os metadados extraídos, gerados ou não previstos pela abordagem LigADOS.

Quadro 10: Comparativo para os metadados da camada de catálogo para um FDP

Metadados da camada de CATÁLOGO					Alinhamento com LigADOS
Ontologia	Nome do Termo	Tipo de dado	Obrigatório/Opcional	Descrição	Extraído do portal (E) / Não previsto (N) / Gerado pela abordagem (G)
RDF	rdf:type	IRI	Obrigatório	Obrigatório ser do tipo dcat:Catalog	E
DC terms	dct:title	String	Obrigatório	Nome do catálogo com a etiqueta de linguagem	E
	dct:hasVersion	String	Obrigatório	Versão do catálogo	G
	dct:publisher	IRI	Obrigatório	Organização(ões) ou pessoa(s) responsável(is) pelo catálogo	G
	dct:description	String	Opcional	Descrição do catálogo com etiqueta de linguagem	E
	dct:language	IRI	Opcional	-	N
	dct:license	IRI	Opcional	-	N
	dct:issued	DateTime	Opcional	Data de criação do catálogo	N
	dct:modified	DateTime	Opcional	Data da última modificação do catálogo	E
	dct:conformsTo	IRI	Opcional	Especificação do esquema de metadados do catálogo (ex.: ShEx)	N
	dct:rights	IRI	Opcional	-	N
	dct:accessRights	IRI	Opcional	Descrição dos direitos de acesso	N
	dct:isPartOf	IRI	Obrigatório	Relação com os metadados da camada-pai	G
FDP ontology	fdp:metadatalIdentifier	IRI	Obrigatório	Identificador da camada de metadados de catálogo no FDP	G
	fdp:metadatalIssued	DateTime	Obrigatório	Data de criação da camada de metadados de catálogo no FDP	G
	fdp:metadatalModified	DateTime	Obrigatório	Data da última modificação dos metadados de catálogo no FDP	G
RDF Schema	rdfs:label	String	Opcional	Nome do catálogo com etiqueta de linguagem	N
FOAF	foaf:homepage	IRI	Opcional	-	N
DCAT	dcat:dataset	IRI	Obrigatório	Lista de URLs de conjuntos de dados do catálogo	E
	dcat:themeTaxonomy	IRI	Obrigatório	Lista de URLs das taxonomias utilizadas para identificar os conjuntos de dados e serviços do portal conjuntos de dados	G

Quadro 11: Comparativo para os metadados da camada de conjuntos de dados para um FDP

Metadados da camada de CONJUNTO DE DADOS					Alinhamento com LigADOS
Ontologia	Nome do Termo	Tipo de dado	Obrigatório/Opcional	Descrição	Extraído do portal (E) / Não previsto (N) / Gerado pela abordagem(G)
RDF	rdf:type	IRI	Obrigatório	Obrigatório ser do tipo dcat:Dataset	E
DC terms	dct:title	String	Obrigatório	Nome do conjunto de dados com etiqueta de nome	E
	dct:publisher	IRI	Obrigatório	Organização(ões) ou pessoa(s) responsável(is) pelo conjunto de dados	E
	dct:hasVersion	String	Obrigatório	Versão do conjunto de dados	G
	dct:description	String	Opcional	Descrição do conjunto de dados com etiqueta de nome	E
	dct:conformsTo	IRI	Opcional	Especificação do esquema de metadados do conjunto de dados (ex.: ShEx)	N
	dct:issued	DateTime	Opcional	Data de criação do conjunto de dados no catálogo	E
	dct:modified	DateTime	Opcional	Data da última modificação do conjunto de dados no catálogo	E
	dct:language	IRI	Opcional	-	E
	dct:license	IRI	Opcional	-	N
	dct:rights	IRI	Opcional	-	N
	dct:accessRights	IRI	Opcional	Descrição dos direitos de acesso	N
	dct:isPartOf	IRI	Obrigatório	Relação com os metadados da camada-pai	G
FDP ontology	fdp:metadatalIdentifier	IRI	Obrigatório	Identificador da camada de metadados de conjunto de dados no FDP	G
	fdp:metadatalIssued	DateTime	Obrigatório	Data de criação da camada de metadados de conjunto de dados no FDP	G
	fdp:metadatalModified	DateTime	Obrigatório	Data da última modificação dos metadados de conjunto de dados no FDP	G
RDF Schema	rdfs:label	String	Opcional	Nome do conjunto de dados com etiqueta de linguagem	N
DCAT	dcat:distribution	IRI	Obrigatório	Lista de URLs de distribuições de conjuntos de dados	E
	dcat:theme	IRI	Obrigatório	Lista de conceitos que descrevem o conjunto de dados	G
	dcat:contactPoint	IRI	Opcional	-	N
	dcat:keyword	String	Opcional	Palavras-chave relacionadas aos conjuntos de dados com etiqueta da linguagem	E
	dcat:landingPage	IRI	Opcional	Home page do conjunto de dados	N

Quadro 12: Comparativo para os metadados da camada de distribuição para um FDP

Metadados da camada de DISTRIBUIÇÃO					Alinhamento com LigADOS
Ontologia	Nome do Termo	Tipo de dado	Obrigatório/Opcional	Descrição	Extraído do portal (E) / Não previsto (N) / Gerado pela abordagem (G)
RDF	rdf:type	IRI	Obrigatório	Obrigatório ser do tipo dcat:Distribution	E
DC terms	dct:title	String	Obrigatório	Nome da distribuição com etiqueta de nome	E
	dct:conformsTo	IRI	Opcional	Especificação do esquema de metadados da distribuição (ex.: ShEx)	N
	dct:license	IRI	Obrigatório	Link pra descrição da licença	G
	dct:hasVersion	String	Obrigatório	Versão da distribuição	G
	dct:issued	DateTime	Opcional	Data de criação da distribuição no catálogo	E
	dct:modified	DateTime	Opcional	Data da última modificação da distribuição no catálogo	E
	dct:rights	IRI	Opcional	-	N
	dct:description	String	Opcional	Descrição da distribuição com etiqueta de nome	E
	dct:accessRights	IRI	Opcional	Descrição dos direitos de acesso	N
	dct:isPartOf	IRI	Obrigatório	Relação com os metadados da camada-pai	G
FDP ontology	fdp:metadataIdentifier	IRI	Obrigatório	Identificador da camada de metadados de distribuição do FDP	G
	fdp:metadataIssued	DateTime	Obrigatório	Data de criação da camada de metadados de distribuição do FDP	G
	fdp:metadataModified	DateTime	Obrigatório	Data da última modificação dos metadados de distribuição do FDP	G
RDF Schema	rdfs:label	String	Opcional	Nome da distribuição com etiqueta de linguagem	N
DCAT	dcat:accessURL	IRI	Obrigatório (ou dcat:downloadURL)	Uma página web, feed, SPARQL endpoint ou outro tipo de recurso que dá acesso à distribuição do conjunto de dados	E
	dcat:downloadURL	IRI	Obrigatório (ou dcat:accessURL)	Um arquivo que contenha a distribuição do conjunto de dados em um dado formato	E
	dcat:mediaType	String	Obrigatório	O tipo de mídia da distribuição	E
	dcat:format	String	Opcional	-	E
	dcat:byteSize	Decimal	Opcional	-	N

APÊNDICE B — RECURSOS COMPUTACIONAIS DA APLICAÇÃO LigADOS

Os recursos computacionais compreendem hardware e software, criando um ambiente para desenvolvimento e experimentação da solução tecnológica para a proposta apresentada neste trabalho.

Os recursos de *hardware* para desenvolver e experimentar a abordagem de solução envolvem um servidor de rede e uma estação de trabalho. É necessário o uso de um servidor para armazenamento do portal de dados. Para isso, uma máquina virtual com o sistema operacional Ubuntu 14.04 LTS 64-bit foi instalada. A máquina virtual possui 4 GB de memória RAM e 12 GB de espaço em disco.

Para o desenvolvimento do código da aplicação gerada pela abordagem foi utilizada uma estação de trabalho com processador Intel Core i7-xxx, 8 GB de memória RAM e 80 GB de espaço em disco.

Dentre os recursos de *software* necessários, uma instância da plataforma de portal CKAN na versão 2.8 foi instalada no servidor provido por meio de uma máquina virtual, bem como a extensão “*ckanext-dcat*”.

A “*ckanext-dcat*” é uma extensão desenvolvida para o CKAN que possibilita a exposição e o consumo de metadados de outros portais de dados usando documentos em RDF que foram anotados com o DCAT. O *Data Catalog Vocabulary* (DCAT) é um vocabulário em RDF para anotação da estrutura dos conjuntos de dados de modo a facilitar a interoperabilidade entre portais de dados, sendo uma recomendação da W3C. A extensão “*ckanext-dcat*” não triplicifica os dados em si; ela triplicifica os metadados e não fornece nativamente capacidades para manipulação de dados conectados (CKAN + DCAT - CKAN EXTENSIONS, 2020). Ou seja, a abordagem utilizada pela extensão prepara metadados para serem utilizados como insumos para abordagens que objetivam interoperabilidade, atuando como atividade-meio nestes processos.

A instalação do CKAN provê uma instalação da linguagem de programação Python. Códigos fonte em Python são interpretados o que possibilita a execução de *scripts* diretamente da linha de comando de um sistema operacional. Na abordagem proposta, são utilizados *scripts* em Python para configurações no CKAN, sem necessidade de interação direta por parte do usuário.

O ambiente de desenvolvimento da abordagem de solução necessita utilizar um *triplestore* para armazenamento dos conjuntos de dados em RDF e execução de consultas utilizando a linguagem de consulta SPARQL. Adicionalmente uma linguagem de programação

foi necessária para codificar uma aplicação que executa as chamadas dos comandos na linguagem de consulta SPARQL. A aplicação desenvolvida orquestra as execuções das consultas disparadas para o *triplestore*. Para codificação a linguagem de programação utilizada foi o Java, mais especificamente o *framework* Eclipse RDF4J.

O Eclipse RDF4J⁵⁴ é um *framework* Java de código aberto para lidar com dados modelados em RDF. O *framework* fornece dois bancos de dados – um em memória e outro em disco – além compatibilidade com outras diversas soluções de armazenamento de triplas. Para isso, o *framework* fornece uma API que pode ser conectada a vários *triplestores*, bem como a estabelecer conexão com qualquer SPARQL *Endpoint*.

O GraphDB™ é um *triplestore* robusto, compatível com os padrões definidos pela W3C e com suporte para a exploração visual e interativa de grafos RDF. O GraphDB™ fornece uma versão não paga (GraphDB™ *Free*) que inclui 16 (dezesesseis) dos 22 (vinte e dois) benefícios oferecidos pela versão paga mais completa (GraphDB™ *Enterprise*), incluindo a capacidade ilimitada de quantidade de triplas RDF que podem ser gerenciadas, o suporte completo ao SPARQL 1.1 e total compatibilidade com o *framework* RDF4J. Ademais, o uso da API fornecida pelo RDF4J em conjunto com o GraphDB™ está bem documentado. A limitação da versão não paga está na possibilidade de execução de somente duas consultas simultaneamente (GRAPHDB - ONTOTEXT, 2020).

Mesmo com esta limitação, o GraphDB™ foi escolhido como *triplestore* para execução dos experimentos desta pesquisa, pelo suporte ao SPARQL 1.1 e a possibilidade de exploração visual e interativa de grafos RDF. A limitação supracitada não afeta o experimento desta pesquisa, que se dá em condições onde a quantidade de requisições simultâneas ao *triplestore* é controlada. Ademais a possibilidade de explorar grafos em RDF utilizando recursos visuais, associada a possibilidade de disponibilizar acesso a funções do *triplestore* como um serviço HTTP reiterou a escolha do produto.

⁵⁴ <https://rdf4j.org/about/>

APÊNDICE C — TRIPLIFICAÇÃO E ANOTAÇÃO SEMÂNTICA DE CONJUNTOS DE DADOS DA APLICAÇÃO DE LigADOS

Para o processo de triplificação foram considerados conjuntos de dados publicados em portais de dados abertos construídos sobre o CKAN, cujo domínio estivesse ligado à temática de agrotóxicos e meio ambiente.

O Quadro 13 apresenta as distribuições de conjuntos de dados extraídas do Portal Brasileiro de Dados Abertos que foram triplificadas, anotadas e utilizadas para experimentação.

Quadro 13: Distribuições de conjuntos de dados extraídos do Portal Brasileiro de Dados Abertos para execução dos experimentos

Organização responsável: Instituto Chico Mendes de Conservação da Biodiversidade - ICMBio Conjunto de dados: Desmatamento em Unidades de Conservação Federais Distribuição: Desmatamento em Unidades de Conservação ... http://dados.gov.br/dataset/incendios-em-unidades-de-conservacao-federais/resource/d27baf34-5f5f-4956-82ef-8036c58e5015 Anotado com tesouro AIMS-AGROVOC.
Organização responsável: Ministério do Meio Ambiente Conjunto de dados: Sistema Nacional de Informações Florestais - SNIF Distribuição: Desmatamento - Bioma - estados - PMDBBS http://dados.gov.br/dataset/sistema-nacional-de-informacoes-florestais-snif/resource/5e18985a-e215-46d3-89dc-7e271c838780 Anotado com tesouro AIMS-AGROVOC e DBPedia Ontology.
Organização responsável: Instituto Brasileiro de Meio Ambiente Conjunto de dados: Vegetação de Mangue brasileira Distribuição: Vegetação de mangue http://dados.gov.br/dataset/vegetacao-de-mangue-brasileira/resource/ef0450e5-72c7-4497-9bc1-8806d5369d6c Anotado com tesouro AIMS-AGROVOC e DBPedia Ontology.
Organização responsável: Instituto Brasileiro de Meio Ambiente Conjunto de dados: Sinaflor - Autorização de Supressão de Vegetação Distribuição: Autorização de Supressão de Vegetação https://dados.contraosagrototoxicos.org/dataset/classificacao-do-iarc-agencia-internacional-de-pesquisa-em-cancer/resource/be27b35b-2ade-4a02-9f93-df597b79f480 Anotado com tesouro AIMS-AGROVOC, DBPedia Ontology e tesouro GEMET.

O conjunto de dados “Orçamento Federal - Orçamento Fiscal e da Seguridade Social” publicado pelo “Ministério da Economia” foi utilizado sem nenhum tipo de pré-processamento, pois já foi disponibilizado em RDF.

O Quadro 14 apresenta as distribuições de conjuntos de dados extraídas do Portal de Dados Abertos sobre Agrotóxicos que foram triplicadas, anotadas e utilizadas para experimentação.

Quadro 14: Distribuições de conjuntos de dados extraídos do Portal de Dados Abertos sobre Agrotóxicos para execução dos experimentos

Organização responsável: Instituto Brasileiro de Meio Ambiente Conjunto de dados: Comercialização IBAMA 2017 Distribuição: Os 10 ingredientes ativos mais vendidos https://dados.contraosagrototoxicos.org/dataset/comercializacao-ibama-2017/resource/9bf99cc9-bf20-4405-8773-7bc1b22d87a5 Anotado com tesouro AIMS-AGROVOC.
Organização responsável: Instituto Brasileiro de Meio Ambiente Conjunto de dados: Comercialização IBAMA 2017 Distribuição: Vendas de ingredientes ativos por UF https://dados.contraosagrototoxicos.org/dataset/comercializacao-ibama-2017/resource/f1f591a8-9e2c-4088-978e-e131e8d52183 Anotado com tesouro AIMS-AGROVOC.
Organização responsável: Ministério da Agricultura, Pecuária e Abastecimento Conjunto de dados: Cadastro Nacional de Produtores Orgânicos Distribuição: Cadastro Nacional de Produtores Orgânicos https://dados.contraosagrototoxicos.org/dataset/cadastro-nacional-de-produtores-organicos/resource/7d63b8e2-72f8-4ee6-9dbf-01e65922338f Anotado com DBPedia Ontology e FOAF Ontology.
Organização responsável: IARC Conjunto de dados: Classificação do IARC - Agência Internacional de Pesquisa em Câncer Distribuição: Classificação IARC - Perigo de Câncer https://dados.contraosagrototoxicos.org/dataset/classificacao-do-iarc-agencia-internacional-de-pesquisa-em-cancer/resource/be27b35b-2ade-4a02-9f93-df597b79f480 Anotado com ChEBI Ontology.
Organização responsável: Ministério da Agricultura, Pecuária e Abastecimento Conjunto de dados: Agrofit Distribuição: Ingredientes Ativos https://dados.contraosagrototoxicos.org/dataset/agrofit/resource/d7f067c1-2349-45b6-987f-f23e0f413a02 Anotado com ChEBI Ontology.
Organização responsável: Agência Nacional de Vigilância Sanitária Conjunto de dados: PARA - Programa de Análise de Resíduos de Agrotóxicos em Alimentos Distribuição: Dados - PARA 2010 https://dados.contraosagrototoxicos.org/dataset/para-programa-de-analise-de-residuos-de-agrotoxicos-em-alimentos/resource/d2da8960-0b7d-4c42-a5f3-5f6a153e2e16 Anotado com tesouro AIMS-AGROVOC e ChEBI Ontology.

APÊNDICE D — EXTRATO DO GRAFO DE INTERLIGAÇÕES

SEMÂNTICAS PARA O EXTRATO PORTAL DE DADOS

ABERTOS SOBRE AGROTÓXICOS

```
<http://localhost/pt_BR/dataset/para-programa-de-analise-de-
residuos-de-agrotoxicos-em-alimentos>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://www.w3.org/ns/dcat#Dataset> .
```

```
<http://localhost/pt_BR/dataset/para-programa-de-analise-de-
residuos-de-agrotoxicos-em-alimentos/resource/6dd9617f-e86c-400e-
961c-e31d81f8eb8b> <http://www.w3.org/1999/02/22-rdf-syntax-
ns#type> <http://www.w3.org/ns/dcat#Distribution> .
```

```
<http://localhost/pt_BR/dataset/para-programa-de-analise-de-
residuos-de-agrotoxicos-em-alimentos/resource/6dd9617f-e86c-400e-
961c-e31d81f8eb8b> <http://purl.org/dc/terms/title> "Dados - PARA
2010" .
```

```
<http://localhost/pt_BR/dataset/para-programa-de-analise-de-
residuos-de-agrotoxicos-em-alimentos>
<http://www.w3.org/ns/dcat#distribution>
<http://localhost/pt_BR/dataset/para-programa-de-analise-de-
residuos-de-agrotoxicos-em-alimentos/resource/6dd9617f-e86c-400e-
961c-e31d81f8eb8b> .
```

```
<http://aims.fao.org/aos/agrovoc/> <http://www.w3.org/1999/02/22-
rdf-syntax-ns#type> <http://purl.org/vocommons/voaf#Vocabulary> .
```

```
<http://aims.fao.org/aos/agrovoc/>
<http://purl.org/vocab/vann/preferredNamespacePrefix> "aims-
agrovoc" .
```

```
<http://localhost/pt_BR/dataset/para-programa-de-analise-de-
residuos-de-agrotoxicos-em-alimentos/resource/6dd9617f-e86c-400e-
961c-e31d81f8eb8b> <www.w3.org/ns/dcat#theme>
<http://aims.fao.org/aos/agrovoc/> .
```

```
<http://localhost/pt_BR/dataset/para-programa-de-analise-de-
residuos-de-agrotoxicos-em-alimentos/resource/6dd9617f-e86c-400e-
961c-e31d81f8eb8b> <http://purl.org/dc/terms/references>
<http://aims.fao.org/aos/agrovoc/c_24430> .
```

```
<http://aims.fao.org/aos/agrovoc/c_24430>
<http://purl.org/dc/dcam/memberOf>
<http://aims.fao.org/aos/agrovoc/> .
```

```
<http://aims.fao.org/aos/agrovoc/c_24430>
<http://www.w3.org/2004/02/skos/core#closeMatch>
<http://dbpedia.org/ontology/ChemicalCompound> .
```

```

<http://dbpedia.org/ontology/ChemicalCompound>
<http://www.w3.org/2000/01/rdf-schema#label> "composto químico"@pt
.
<http://dbpedia.org/ontology/ChemicalCompound>
<http://purl.org/dc/dcam/memberOf> <http://dbpedia.org/ontology/> .
<http://dbpedia.org/ontology/>
<http://purl.org/vocab/vann/preferredNamespacePrefix> "dbpedia-
ontology" .

<http://dbpedia.org/ontology/> <http://www.w3.org/1999/02/22-rdf-
syntax-ns#type> <http://purl.org/vocommons/voaf#Vocabulary> .

<http://localhost/pt_BR/dataset/para-programa-de-analise-de-
residuos-de-agrotoxicos-em-alimentos/resource/6dd9617f-e86c-400e-
961c-e31d81f8eb8b> <http://purl.org/dc/terms/references>
<http://aims.fao.org/aos/agrovoc/c_3131> .

<http://aims.fao.org/aos/agrovoc/c_3131>
<http://purl.org/dc/dcam/memberOf>
<http://aims.fao.org/aos/agrovoc/> .

<http://aims.fao.org/aos/agrovoc/c_3131>
<http://www.w3.org/2000/01/rdf-schema#label> "Fruta"@pt .

```