MINISTÉRIO DA DEFESA
EXÉRCITO BRASILEIRO
DEPARTAMENTO DE CIÊNCIA E TECNOLOGIA
INSTITUTO MILITAR DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS E COMPUTAÇÃO

RAFAEL NEVES DA SILVEIRA

MÉTODO PARA ROTULAR LIGAÇÕES SEMÂNTICAS NA WEB DE DADOS

RIO DE JANEIRO
2021

RAFAEL NEVES DA SILVEIRA

MÉTODO PARA ROTULAR LIGAÇÕES SEMÂNTICAS NA WEB DE DADOS

Dissertação apresentada ao Programa de Pós-graduação em Sistemas e Computação do Instituto Militar de Engenharia, como requisito parcial para a obtenção do título de Mestre em Ciências em Sistemas e Computação.

Orientador(es): Maria Cláudia Reis Cavalcanti, D.Sc.

Rio de Janeiro

2021

Silveira, Rafael Neves da.
Método para rotular ligações semânticas na Web de Dados / Rafael Neves da Silveira. – Rio de Janeiro, 2021.
103 f.

Orientador(es): Maria Cláudia Reis Cavalcanti.

Dissertação (mestrado) – Instituto Militar de Engenharia, Sistemas e Computação, 2021.

1. Web Semântica. 2. Ontologias. 3. Enriquecimento de Dataset. i. Reis Cavalcanti, Maria Cláudia (orient.) ii. Título

RAFAEL NEVES DA SILVEIRA

Método para rotular ligações semânticas na Web de Dados

Dissertação apresentada ao Programa de Pós-graduação em Sistemas e Computação do Instituto Militar de Engenharia, como requisito parcial para a obtenção do título de Mestre em Ciências em Sistemas e Computação.

Orientador(es): Maria Cláudia Reis Cavalcanti.

Aprovado em Rio de Janeiro, 9 de fevereiro de 2021, pela seguinte banca examinadora:

Prof. **Maria Cláudia Reis Cavalcanti** - D.Sc. do IME - Presidente

Prof. **Giseli Rabello Lopes** - D.Sc. da UFRJ

Prof. **Ronaldo Ribeiro Goldschmidt** - D.Sc. do IME

Rio de Janeiro

2021

Ao Serviço Geológico do Brasil,
pela oportunidade de capacitação.

AGRADECIMENTOS

Agradeço à minha família, por me apoiar e entender o afastamento necessário para o estudo e as atividades de pesquisa. Agradeço também à minha orientadora, professora Maria Cláudia Reis Cavalcanti, pela dedicação e por me guiar de forma tão eficiente em busca do conhecimento, sempre me inspirando com seu exemplo. Por fim, agradeço aos colegas do Instituto Militar de Engenharia, pela ajuda e parceria nos momentos de dificuldade.

"Toda reforma interior e toda mudança para melhor dependem exclusivamente da aplicação do nosso próprio esforço."

(Immanuel Kant)

RESUMO

A Web Semântica, com suas linguagens e padrões, fornece uma estrutura comum que permite que os dados sejam compartilhados e reutilizados. Uma forma de aumentar o conhecimento sobre esses dados é realizando novas interligações entre *datasets*. No entanto, a maioria das abordagens de interligação apresentam ligações do tipo *"Same As"* ou *"Related To"*. Este último tipo deixa vago o significado da relação encontrada. Este trabalho apresenta um método para rotular esse tipo de relação entre *datasets*, por meio da combinação do uso de ontologias e vocabulários controlados com técnicas de *Relation Extraction*. Além do método, apresenta também a aplicação WEB denominada PLAIN que o implementa, e estudos de caso demonstrando a funcionalidade, a viabilidade e o impacto da abordagem proposta.

Palavras-chave: Web Semântica. Ontologias. Enriquecimento de Dataset.

ABSTRACT

The Semantic Web, with its languages and standards, provides a common framework that allows data to be shared and reused. One way to increase knowledge about this data is by making new interconnection between datasets. However, most of the interconnection approaches have connections like "Same As" or "Related To". The latter type leaves vague the meaning of the relationship found. This paper presents a method for labeling this type of relationship between datasets, by combining the use of ontologies and controlled vocabularies with Relation Extraction techniques. Besides the method, it also presents the WEB application called PLAIN that implements it, and case studies demonstrating the functionality, feasibility and impact of the proposed approach.

Keywords: Semantic Web. Ontologies. Dataset enrichment.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BPMN	*Business Process Model and Notation*
CPRM	Companhia de Pesquisa de Recursos Minerais
DEER	*Data Extraction and Enrichment Framework*
IE	*Information Extraction*
IME	Instituto Militar de Engenharia
LOD	*Linked Open Data*
MRAR	*Mining Multi-Relation Association Rules*
OWL	*Web Ontology Language*
NE	*Named Entity*
NLP	*Natural Language Processing*
PHP	*Hypertext Preprocessor*
RE	*Relation Extraction*
RDF	*Resource Description Framework*
SIAGAS	Sistema de Informações de Águas Subterrâneas
SPARQL	*SPARQL Protocol and RDF Query Language*
UF	Unidade Federativa
URI	*Uniform Resource Identifier*
W3C	*World Wide Web Consortium*
WS	*Web* Semântica

LISTA DE SÍMBOLOS

$\wedge$	E
$\exists$	Existe
$\in$	Pertence
$\times$	Produto cartesiano
$\mid$	Tal que
$\cup$	União

SUMÁRIO

1 INTRODUÇÃO

Atualmente, a quantidade de dados gerados pela sociedade aumenta gradativamente a cada dia. Isso é possibilitado pelos avanços tecnológicos e motivado pela demanda crescente da população por informação em tempo real. Em um mundo competitivo, as empresas e governos buscam atender a essa necessidade com a busca por novas formas de lidar com esse grande volume de informações.

Diante disso, a Web destaca-se por sua amplitude e importância global, uma vez que possibilitou o amplo acesso da sociedade a todos os tipos de informações. Nela o fluxo da informação segue de forma não-hierárquica e interativa, na medida em que tanto emissores quanto destinatários possuem funções ativas de produtores e receptores de informações. Devido ao grande volume dessas informações, torna-se necessário organizá-las de forma a torná-las compreensíveis por agentes de software, a fim de que os dados possam gerar algum tipo de conhecimento.

Como uma evolução, a Web Semântica[1] (WS), com suas linguagens e padrões, fornece uma estrutura comum que permite que os dados sejam compartilhados e reutilizados além dos limites de aplicativos, empresas e comunidades. Para alimentar a WS, surgiram iniciativas como a denominada Dados Conectados, do inglês *Linked Data*, que é um conjunto de boas práticas para a publicação de dados na Web. Nesse contexto, os Dados Conectados são dados publicados e ligados utilizando as tecnologias e padrões da WS (YU, 2014).

1.1 Motivação e Caracterização do Problema

Uma forma de aumentar o conhecimento sobre Dados Conectados é realizando uma ampliação dos *datasets* da WS. A partir de um *dataset* de interesse, que pode ser chamado de *dataset Fonte*, a tarefa de ampliação inicia-se pela busca por *datasets* externos que contenham recursos comuns ao *dataset Fonte*. No entanto, essa não é uma tarefa simples porque, após encontrar um *dataset* externo, também chamado de *dataset Alvo*, é preciso obter associações que interliguem recursos de ambos os *datasets*. Já em um processo de enriquecimento, um conjunto de ações mais amplo pode ser utilizado na busca por relações mais complexas e, assim, criar propriedades extras.

Para encontrar essas interligações, abordagens baseadas em mineração de regras de associação são eventualmente utilizadas. O algoritmo MRAR (RAMEZANI et al., 2014) explora regras de associação de multirrelação sobre grafos direcionados, como os

[1] https://www.w3.org/standards/semanticweb/

que são usados para representação de dados na Web de Dados. Já o algoritmo MRAR+ (OLIVEIRA et al., 2019) é uma evolução do algoritmo MRAR e uma das iniciativas que utiliza esta abordagem para encontrar associações entre recursos de *datasets* na Web. Assim como o MRAR+, ferramentas como o Silk (BIZER et al., 2009) e o LIMES (NGOMO; AUER, 2011) também são focadas na tarefa de interligação de *datasets*.

Além disso, a maioria dos trabalhos da literatura são direcionados para a descoberta de relações do tipo *"Same As"*. No entanto, é frequente a utilização equivocada dessa relação para vincular recursos que não são necessariamente o mesmo, conforme discutido em (HALPIN; HAYES, 2010). Nesse sentido, conforme apurado em (SCHMACHTENBERG; BIZER; PAULHEIM, 2014), predicados com semântica pobre, cujo significado é incerto (por exemplo: "veja também"), são amplamente utilizados em *datasets* publicados de diferentes áreas de conhecimento.

Em um dos levantamentos apresentados é possível observar que o predicado *rdfs:seeAlso*[2] foi utilizado em mais de 43% do total de predicados da área de ciências da vida e 52% entre os predicados da área interdisciplinar (do inglês, *crossdomain*). Isso coloca o predicado como o segundo mais utilizado nessas áreas de conhecimento. Ao consultar *datasets* nesse contexto, como a DBpedia (*crossdomain*), pode-se observar que outro predicado, com uma semântica mais rica, poderia definir melhor esse tipo de relação. Um exemplo é a tripla *dbo:Grape rdfs:seeAlso dbo:Wine*, cujo predicado poderia ser substituído por *dbo:ingredient*[3].

Há também trabalhos que sugerem relações do tipo *"Related To"*, como em (SHERIF; NGOMO; LEHMANN, 2015), em que não há clareza semântica na ligação. Uma outra abordagem para aprimorar dados da Web com semântica é o Processamento de Linguagem Natural (do inglês *Natural Language Processing* – NLP), que é o processamento automático de linguagem humana. Ele é composto por uma série de tarefas. Por exemplo, através de técnicas de NLP é possível identificar nos textos das páginas Web referências a entidades do mundo real, que, nesse contexto, são chamadas de Entidades Nomeadas (ou *Named Entities* – NE's, em inglês), e suas relações, e atribuir uma *Uniform Resource Identifier* – URI a cada uma dessas entidades. Essa técnica é conhecida como Extração de Informações (do inglês *Information Extraction* – IE). Além de identificar as NE's, a IE também pode extrair a relação entre elas, tarefa conhecida como Extração de Relações (do inglês *Relation Extraction* – RE). Assim, tem-se uma questão em aberto: como melhorar a semântica das relações encontradas após o processo de enriquecimento de *dataset*?

[2] https://www.w3.org/wiki/UsingSeeAlso
[3] http://dbpedia.org/ontology/ingredient

1.2 Hipótese e Questões de Pesquisa

Ao se detalhar melhor o problema, pode-se levantar as seguintes questões de pesquisa a serem respondidas no decorrer do trabalho:

1. Como é possível rotular predicados em *datasets* da Web de Dados com a ajuda de catálogos de vocabulários controlados?
2. De que forma a combinação do uso de vocabulários controlados com técnicas de *Relation Extraction* auxilia no processo de rotulação?

Diante dessas questões, levanta-se a hipótese de que se utilizarmos ontologias e vocabulários controlados, combinados a técnicas de RE, então poderemos enriquecer semanticamente um *dataset*, rotulando as relações com semântica pobre.

Esta hipótese se justifica pois é possível observar que existe um padrão de relacionamentos formalmente mapeado em vocabulários controlados e em textos escritos em linguagem natural, que pode ser utilizado nessa rotulação.

1.3 Objetivo

Este trabalho visa desenvolver um método para enriquecer as ligações nos *datasets* de dados conectados. Isso se dá pela explicitação da semântica dessas ligações e da rotulação das mesmas. A ideia é partir das relações descobertas por mecanismos de descoberta de relações e/ou enriquecimento de *datasets* na Web de Dados. Então, utilizando catálogos de vocabulários e ontologias, encontrar recursos semânticos que ajudem a rotular as ligações. A proposta inclui também o uso de técnicas de Processamento de Linguagem Natural (do inglês *Natural Language Processing* – NLP), mais especificamente, técnicas de *Relation Extraction - RE*. Os objetivos específicos desta pesquisa são:

1. Implementar um protótipo com base no método proposto, de modo complementar a ferramentas como MRAR+;
2. Realizar estudos de caso que permitam validar o método com o auxílio de especialistas no domínio aplicado.

1.4 Contribuições Esperadas

As contribuições esperadas para este trabalho são:

1. Um método para rotular novas relações em *datasets* da Web de Dados;

2. Um protótipo com base no método proposto;
3. Resultados de estudos de caso sobre dados reais que permitam validar o método com o auxílio de especialistas no domínio aplicado.

1.5 Organização do Trabalho

A presente pesquisa está organizada em seis capítulos. Este capítulo apresenta a introdução do trabalho, com a exposição da motivação e a caracterização do problema, objetivo e contribuições esperadas. No Capítulo 2 será apresentada a fundamentação teórica oportuna. No Capítulo 3 serão apresentados os trabalhos relacionados ao assunto tratado nesta pesquisa. Já no Capítulo 4 será tratada a abordagem proposta. Em seguida, no Capítulo 5, serão apresentados estudos de caso e os resultados de experimentos realizados. Por fim, no Capítulo 6, será apresentada a conclusão do trabalho, pontuando as contribuições e sugestões de trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Nas próximas subseções serão apresentados os conceitos básicos que foram utilizados na pesquisa. Serão descritos os conceitos de Web de Dados, vocabulários controlados, ontologias, repositórios de vocabulários e ontologias, interligação de *datasets* e extração de relações.

2.1 Web de Dados

Em termos técnicos, a Web[1] é um sistema que foi inventado por Tim Berners-Lee e Robert Cailliau, em 1989, com o objetivo de utilizar a Internet para consulta e atualização de documentos organizados em uma estrutura hipertextual. Hipertexto é um texto estruturado, composto por um conjunto interligado de recursos, por exemplo, textos, imagens, vídeos, etc. A arquitetura desse sistema foi criada com base no conceito cliente-servidor, onde uma aplicação (um cliente Web) requisita um documento (um recurso) a uma outra aplicação (um servidor Web) informando a identificação desse documento (LAUFER, 2015).

Além da Web de documentos clássica, o *World Wide Web Consortium* (W3C) está ajudando a criar um conjunto de tecnologias para apoiar uma Web de Dados, com o mesmo tipo de dado encontrado nos bancos de dados tradicionais. A ideia da Web de Dados é permitir que os computadores realizem um trabalho mais útil, possibilitando o desenvolvimento de sistemas que possam facilitar interações confiáveis na Web. Nela, informações adicionais são disponibilizadas para permitir que as máquinas possam compreender melhor os dados contidos nos recursos e o significado de determinada relação entre dois ou mais recursos (ISOTANI; BITTENCOURT, 2015).

O *Resource Description Framework*[2] (RDF) é uma linguagem para representação de recursos na Web. Os links descritos nessa linguagem são utilizados para acessar dados de diversos alvos. O RDF é uma estrutura para representar informações na Web Semântica. Ele permite fazer afirmações sobre recursos que, nesse contexto, são quaisquer coisas, tanto concretas quanto abstratas. Uma determinada empresa, uma pessoa, uma página Web, um sentimento, uma cor, também são considerados recursos (LAUFER, 2015).

Para descrever a relação entre recursos, o RDF oferece uma estrutura de triplas do tipo *<sujeito> <predicado> <objeto>* (ISOTANI; BITTENCOURT, 2015). O conjunto destas estruturas em triplas é chamado de Grafo RDF. A representação de um grafo RDF pode ser visualizada na Figura 1.

[1] https://www.w3.org/History/1989/proposal.html
[2] https://www.w3.org/TR/rdf-primer/

Figura 1 – Tripla RDF (FERRAZ, 2018)

Conforme observado, uma afirmação RDF consiste de três elementos (uma tripla) e tem a seguinte estrutura: *<sujeito> <predicado> <objeto>*. Essa afirmação expressa uma relação entre dois recursos. O sujeito e o objeto representam os dois recursos sendo relacionados. O predicado representa a natureza desta relação, que é formulada de modo direcional (do sujeito para o objeto) e é chamada em RDF de propriedade. Ou seja, as Ligações Semânticas são realizadas por predicados.

A Figura 2 demonstra um exemplo de múltiplas triplas que são disponibilizadas na linguagem RDF.

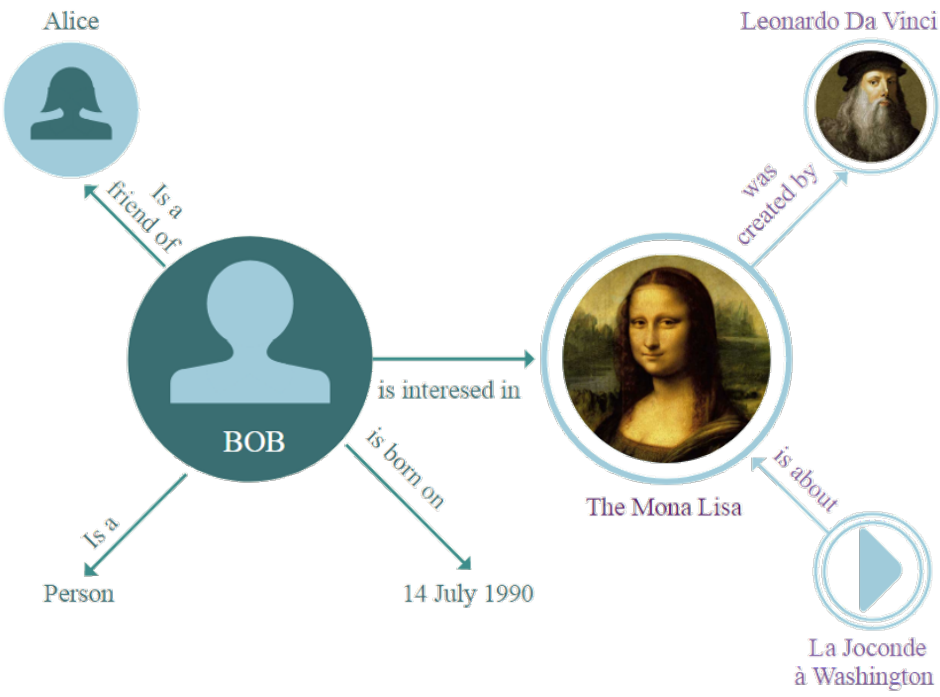


Figura 2 – Cenário descrevendo um conjunto de triplas e seus relacionamentos (ISOTANI; BITTENCOURT, 2015).

Já a Web Semântica[3] estende a Web Clássica, provendo uma estrutura semântica para páginas Web, a qual permite que tanto agentes humanos quanto agentes de software possam entender o conteúdo presente nas páginas. O termo Web Semântica refere-se à visão do W3C da Web de Dados Conectados. Os padrões e tecnologias definidos pelo

[3] https://www.w3.org/standards/semanticweb/

W3C para a Web Semântica permitem que sejam criados repositórios de dados na Web, construam vocabulários e escrevam regras para descrever e lidar com tais dados.

Dados Conectados e Web de Dados são conceitos intimamente relacionados ao conceito de Web Semântica. A ideia de Dados Conectados foi proposta originalmente por Tim Berners-Lee, e seus princípios de Dados Conectados são considerados a introdução oficial e formal do próprio conceito. Atualmente, Dados Conectados são um movimento apoiado pelo W3C que se concentra na conexão de conjuntos de dados na Web e pode ser visto como um subconjunto do conceito de Web Semântica, que consiste em adicionar significados à Web (YU, 2014).

A Web Semântica, através de suas tecnologias e padrões, define a forma como sintaticamente os metadados podem ser agregados às informações. Nesse sentido, deve-se fazer uso de vocabulários que possuam uma semântica bem definida para que o significado pretendido pelo publicador seja o mesmo que o significado entendido pelo consumidor dos dados.

2.2 Vocabulários Controlados e Ontologias

A concepção e o uso das ontologias sempre fizeram parte da proposta da Web Semântica e têm-se mostrado uma das tecnologias-chave na criação de aplicativos mais adequados para lidar com grandes quantidades de informações de maneira inteligente (MCGUINNESS, 2004; HORROCKS, 2008). Uma ontologia pode ser definida como uma especificação formal e explícita de uma conceituação compartilhada (GRUBER, 1995). Mais detalhadamente, uma ontologia deve:

- Ser legível por máquina;
- Ter seus conceitos, propriedades, relações, funções, restrições e axiomas explicitamente definidos;
- Ser um modelo abstrato com visão simplificada de algum fenômeno no mundo que se quer representar;
- Ter conhecimento consensual.

A seguir é apresentado um exemplo de ontologia especificada em triplas RDF e que utiliza também o OWL[4]:

Person	*owl:equivalentClass*	*:Human*	.
Mary	*rdf:type*	*:Person*	.
Mary	*rdf:type*	*:Human*	. *(inferência)*

[4] https://www.w3.org/2001/sw/wiki/OWL

Nos casos em que nenhum vocabulário satisfaz as necessidades de expressividade dos metadados, uma nova ontologia pode ser criada, com o cuidado em reutilizar o maior número possível de elementos de ontologias já existentes, evitando assim a duplicação de referências diferentes aos mesmos conceitos (LAUFER, 2015).

Os vocabulários controlados proporcionam uma organização do conhecimento para posterior recuperação. Na biblioteconomia e na ciência da informação, o vocabulário controlado é uma lista cuidadosamente selecionada de palavras e frases, que são usadas para marcar unidades de informação para que possam ser mais facilmente recuperadas por uma busca.

Um fator de sucesso para se atingir esse objetivo é a utilização de vocabulários de referência. Os termos vocabulário e ontologia são comumente utilizados de forma intercambiável. Não existe na literatura uma separação clara dos dois conceitos, sendo que, em muitos casos, o termo vocabulário é utilizado para o caso de ontologias mais simples (LAUFER, 2015).

A diferença fundamental entre uma ontologia e um vocabulário controlado é o nível de abstração e as relações entre conceitos. Uma ontologia formal é um vocabulário controlado expresso em uma linguagem de representação ontológica. Esta linguagem tem uma gramática para usar termos de vocabulário para expressar algo significativo dentro de um domínio específico de interesse. A gramática contém restrições formais (por exemplo, especifica o que significa ser uma declaração bem formada, afirmação, consulta, etc.) sobre como os termos do vocabulário controlado da ontologia podem ser usados em conjunto.

2.3 Repositórios de Vocabulários e Ontologias

Conforme observado, para que o cenário da Web Semântica fique completo é preciso que se estabeleça um conjunto de vocabulários de referência de forma a facilitar a comunicação dos metadados. Para cada publicação específica, deve ser feita uma busca de vocabulários existentes que possam ser reutilizados. Existem alguns catálogos que podem auxiliar o usuário na busca de ontologias, entre eles o *Linked Open Vocabularies - LOV*[5], o BioPortal[6], o AgroPortal[7] e o JoinUp[8]. A Figura 3 apresenta a tela inicial do serviço de busca oferecido pelo LOV. Nela, cada círculo representa um vocabulário e seu "tamanho" equivale à quantidade de seu reuso por outros vocabulários.

Catálogos de vocabulários controlados possibilitam essa exploração porque agrupam em um local a informação de diversos vocabulários e suas relações. Na Figura 4 é apresentado um exemplo de uma consulta ao LOV por meio de sua interface Web. No

[5] https://lov.linkeddata.es/dataset/lov/
[6] https://bioportal.bioontology.org/
[7] http://agroportal.lirmm.fr/
[8] https://joinup.ec.europa.eu/

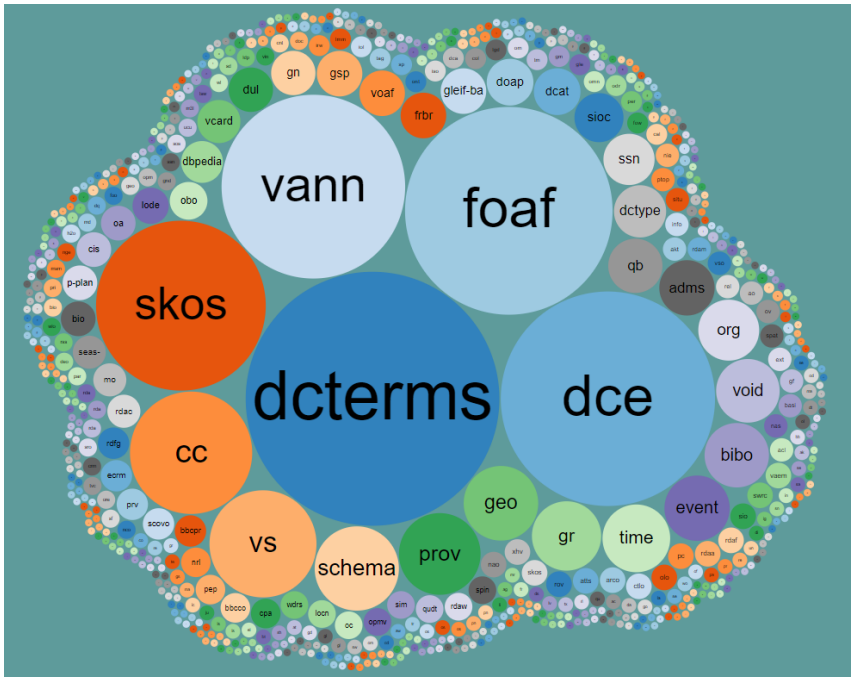


Figura 3 – Tela inicial do LOV.

exemplo é consultado o termo “sport” e filtrado pelos resultados que são classes. Foram retornadas 138 classes como opção.

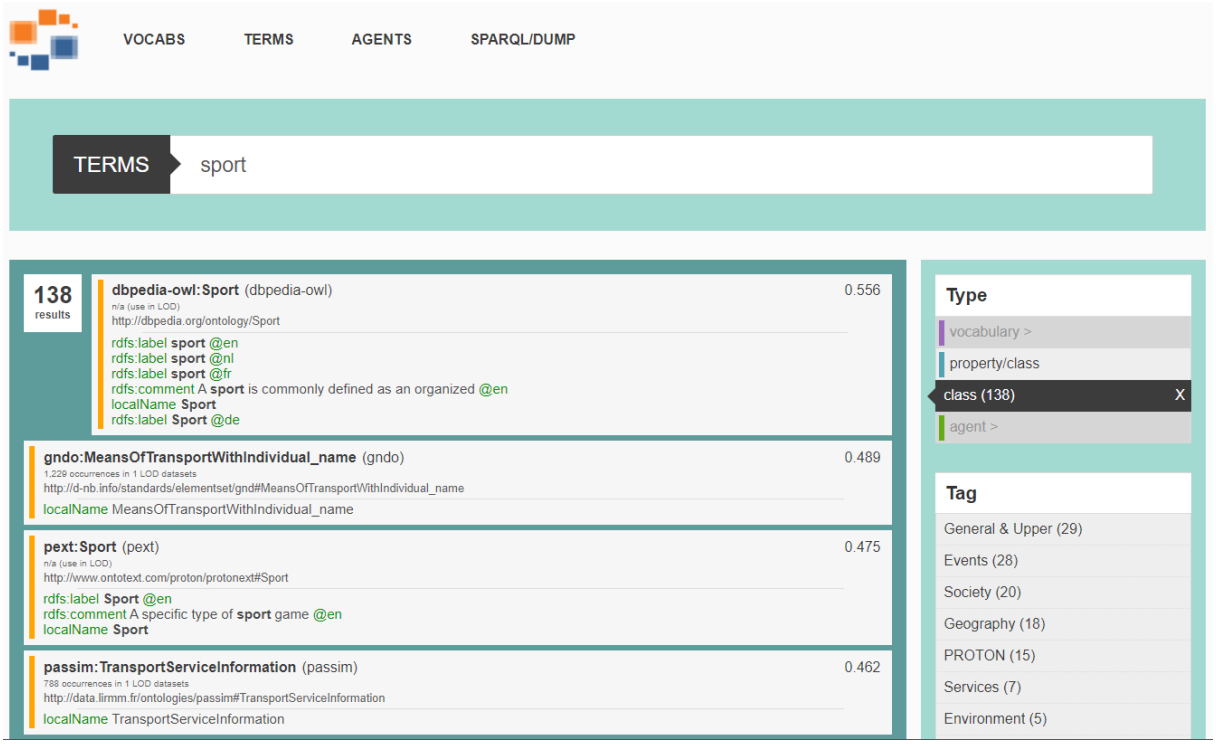


Figura 4 – Exemplo de consulta ao LOV utilizando a interface Web.

Uma busca pelo termo “professor” no LOV retorna os vocabulários que contam com esse termo. É disponibilizada uma lista com os vocabulários nos quais o termo foi

encontrado. Pode-se navegar para um dicionário e explorar a hierarquia de classes do termo. Dando continuidade ao exemplo acima, na lista que resultou da busca realizada, ao selecionar a classe "*Professor*" no vocabulário DBpedia, pode-se observar que esta é uma subclasse de "*Scientist*". Ainda"*Scientist*" é possível observar que esta é uma subclasse da classe "*Person*". Esse fluxo de navegação está representado no diagrama da Figura 5.

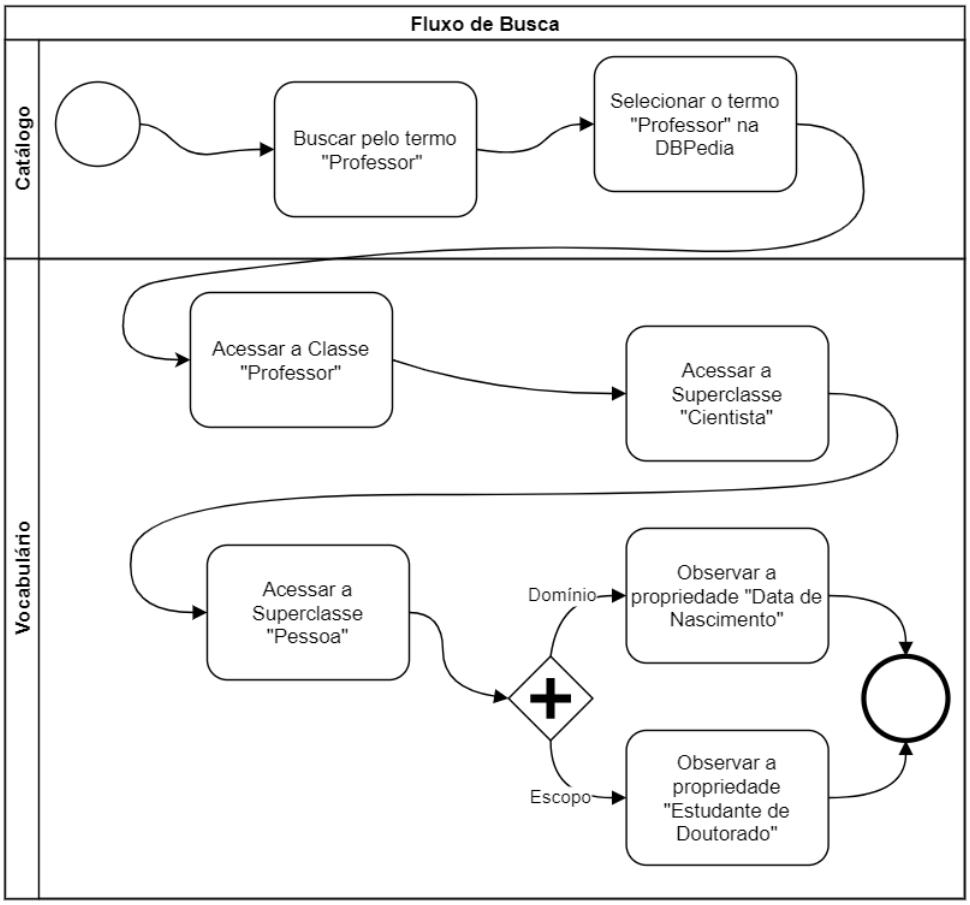


Figura 5 – Um exemplo de fluxo de navegação pelo LOV.

Para cada classe navegada existem propriedades das quais ela é domínio (*domain*) e escopo (*range*). A classe "*Person*", por exemplo, é domínio da propriedade "*birthDate*" e escopo da propriedade "*doctoralStudent*", entre outros. Desta maneira, observa-se que está formalizado que uma pessoa pode ter como propriedade uma data de nascimento e que ser estudante de doutorado é uma propriedade de pessoa, e pode associar tal pessoa à universidade ou ao seu orientador.

2.4 Interligação de *Datasets*

Quando se tem um conjunto de Dados Conectados, pode ser interessante aumentar o conhecimento sobre ele. Uma forma de aumentar esse conhecimento é por meio da análise das ligações entre os recursos disponíveis nesse conjunto (LEHMANN; SCHÜPPEL;

AUER, 2007; HERRERA et al., 2016). Outra forma de aumentar esse conhecimento é realizando uma ampliação de *dataset*, que pode ser chamado de *dataset Fonte*. De forma geral, essa tarefa passa pela busca por *datasets Alvos* que contenham recursos relacionados ao *dataset Fonte*. Uma vez encontrado um *dataset Alvo*, um novo desafio se apresenta que é estabelecer as relações entre os recursos de ambos os *datasets*. Na Figura 6 é apresentada uma sequência genérica de um processo de ampliação.

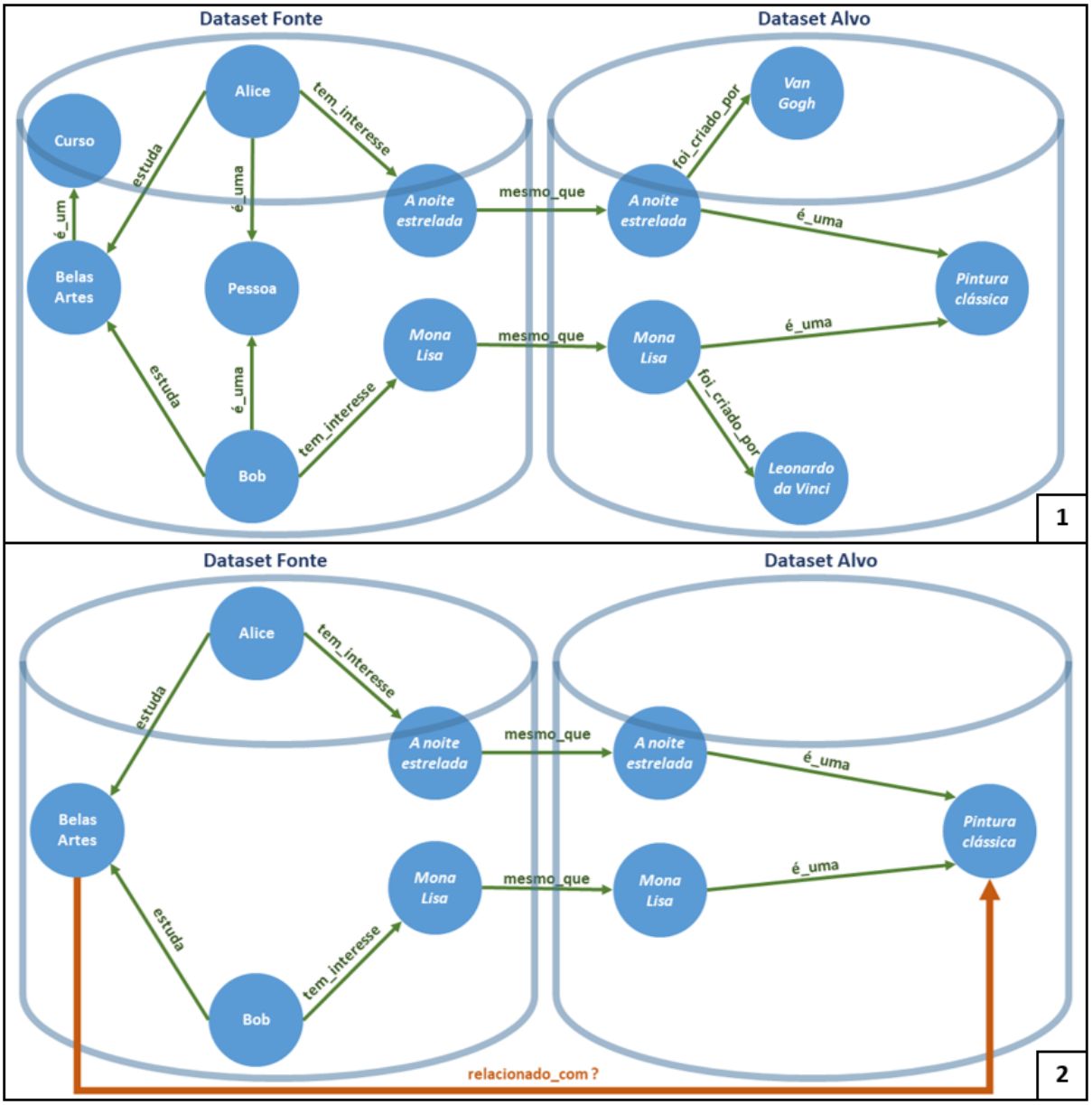


Figura 6 – Interligação de *datasets*. Adaptado de (ISOTANI; BITTENCOURT, 2015).

Na etapa 1 apresentada na Figura 6 assume-se que o *dataset Alvo* já foi encontrado. Essa etapa é responsável por encontrar os recursos semelhantes. Tem-se o *dataset Fonte* (a ser ampliado) com sete triplas, e um segundo *dataset*, que contém recursos em comum com o *dataset Fonte*. No exemplo da figura, realiza-se a vinculação dos recursos em comum representados como: *A noite estrelada* e *Mona Lisa*. Num segundo momento, na

etapa 2, é possível ainda identificar novas relações como a relação entre *Belas Artes* e *Pintura Clássica*. Conforme destacado na figura, nota-se que poderia existir uma tripla relacionando os recursos *Belas Artes* e *Pintura Clássica*. Mas qual de fato é a semântica desse relacionamento? Identificar esta semântica é uma tarefa difícil. Poderia-se sugerir, por exemplo, que no curso de *Belas Artes* ***estuda-se*** *Pintura Clássica*, mas até onde foi possível investigar, o que se consegue gerar automaticamente com as ferramentas atuais é apenas uma relação genérica do tipo ***relacionado com***.

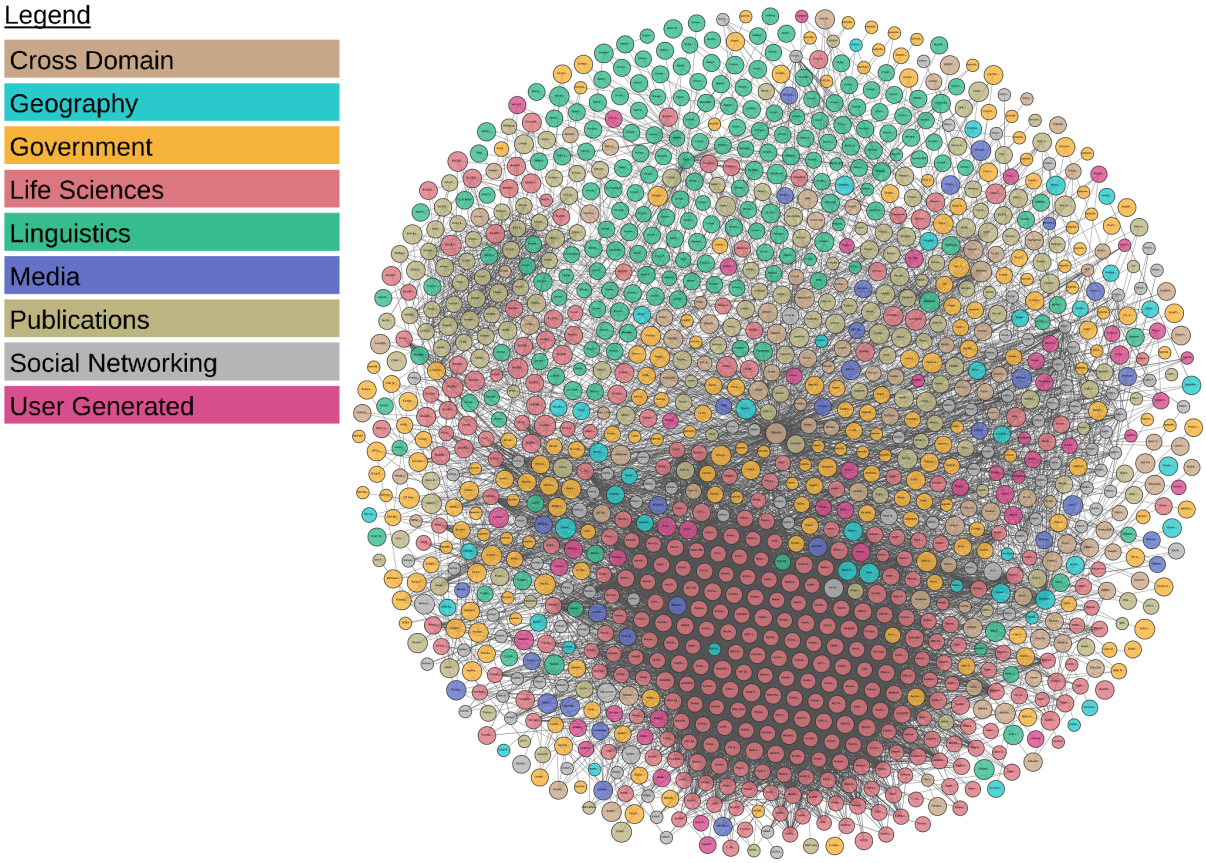


Figura 7 – Diagrama da Nuvem de Dados Abertos Conectados.

Iniciativas de interligação de dados, como a destacada na etapa 1 da Figura 6, têm dado mais atenção ao enriquecimento semântico dos *datasets*. Nos últimos anos, a Nuvem de dados abertos conectados (do inglês *Linked Open Data - LOD*[9]) tem sido a representação mais conhecida dos princípios de Dados Conectados. Até dezembro de 2020, tinha-se 1.269 conjuntos de dados com 16.201 *links*, conforme apresentado na Figura 7. Nessa figura, cada círculo representa um *dataset* distinto presente na nuvem da LOD, já os arcos indicam que existem ligações entre os itens dos conjuntos de dados conectados. Há também a diferenciação de cores entre as áreas de conhecimento de cada *dataset*, como pode ser observado na legenda da figura.

[9] https://lod-cloud.net/

De acordo com o apresentado em (MANDREOLI; MONTANGERO, 2019), a intenção por trás desse conjunto de dados interconectados é criar a semente inicial da extensão legível por máquina da Web atual, apelidada de Web de Dados. Muitos conjuntos com grande volume de dados são adicionados à nuvem LOD regularmente porém são pouco vinculados a outros conjuntos de dados, i.e., não há muitas ligações entre recursos de distintos *datasets*.

Estudos mostram que 44% dos *datasets* na LOD não estão conectados a outros conjuntos de dados (SCHMACHTENBERG; BIZER; PAULHEIM, 2014). Conforme apontado em (NENTWIG et al., 2017), o principal motivo dessa importante falta de links na nuvem LOD está na dificuldade de criá-los, sendo um processo muito custoso quando realizado manualmente.

Muitas ferramentas e estruturas de software já foram desenvolvidas para solucionar o problema da descoberta de link (do inglês *Link Discovery* – LD), especialmente para identificar objetos semanticamente equivalentes em diferentes *datasets*. O funcionamento básico por trás da maioria dessas abordagens é reduzir o problema do LD a um problema de computação de similaridade, como ilustrado pela etapa inicial da Figura 8. Dados dois conjuntos de recursos, *Fonte*, do inglês *Source* – S, e *Alvo*, do inglês *Target* – T, o objetivo é encontrar automaticamente pares de recursos em $S \times T$ que devem ser vinculados entre si. Para realizar essa vinculação, frequentemente se utiliza a propriedade *owl:sameAs*[10] que, em (PARIS, 2018) o autor explica que essa é uma das relações mais importantes da LOD, sendo utilizada para indicar que dois recursos são iguais.

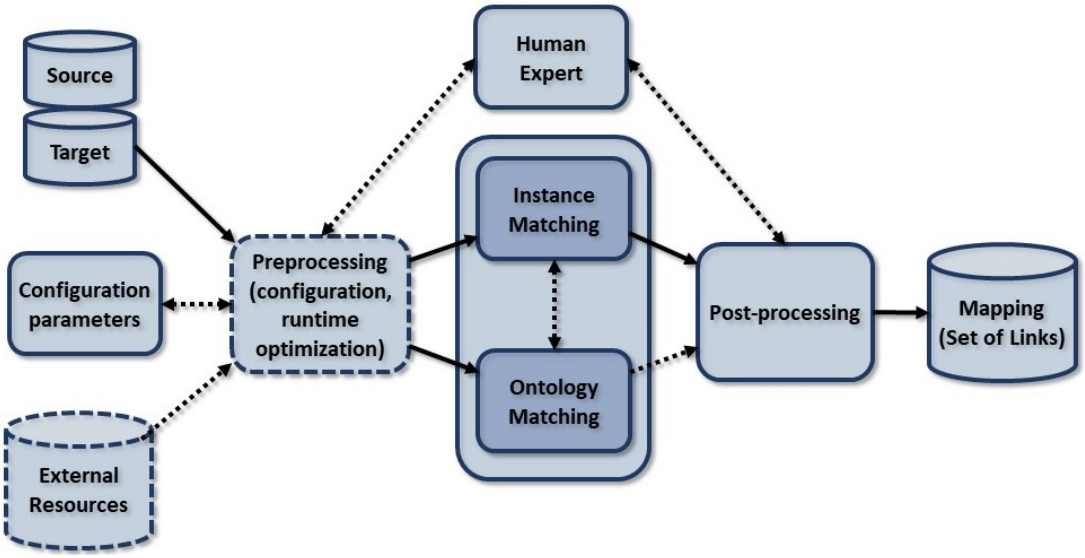


Figura 8 – Fluxo de trabalho genérico dos frameworks de LD (NENTWIG et al., 2017).

Alcançar uma alta eficácia e uma alta eficiência do processo de vinculação são os dois principais desafios que surgem ao lidar com LD. Uma alta eficácia consiste em

[10] https://www.w3.org/2002/07/owl#sameAs

encontrar quase todos os links entre duas fontes fornecidas sem derivar links incorretos. Para atingir esse objetivo, é necessário encontrar uma configuração ou especificação de link adequada, especificando as condições de similaridade que dois recursos $s \in S$ e $t \in T$ devem cumprir, para que sejam considerados similares. Mesmo quando recebemos uma especificação de link adequada, temos que resolver o problema de eficiência, pois uma implementação ingênua que compara todos os elementos de S com todos os elementos de T teria uma complexidade de $O(\mid S \mid \cdot \mid T \mid)$.

Segundo (NENTWIG et al., 2017), o problema de descoberta de link (do inglês *Link Discovery* – LD) pode ser descrito da seguinte maneira:

- Dados dois conjuntos de recursos S e T (por exemplo, sobre filmes) e uma relação Ψ[11](por exemplo, *owl:sameAs* ou *dbo:producer*), encontre todos os pares $(s, t) \in S \times T$ tal que $\Psi(s, t)$ se mantenha;
- O resultado é representado como um conjunto de links chamado mapeamento: $M_{S,T} = \{(a_i, \Psi, b_j) | a_i \in S, b_j \in T\}$;
- Opcionalmente, uma pontuação de similaridade ($sim \in [0, 1]$) calculada por uma ferramenta LD pode ser adicionada às entradas de mapeamentos para expressar a confiança de um link calculado;
- Nesse caso, os links podem ser representados como quádruplos $(a_i, \Psi, b_j, sim(a_i, b_j))$.

LD tem muitas semelhanças com o problema da resolução da entidade, do inglês *Entity Resolution – ER*, também chamado de deduplicação, reconciliação de referência ou correspondência de objetos, que já foi abordado por (ELMAGARMID; IPEIROTIS; VERYKIOS, 2006; KÖPCKE; RAHM, 2010; CHRISTEN, 2012). Para ambos os problemas podem ser aplicadas técnicas semelhantes para avaliar a similaridade entre objetos, com eficácia, e para melhorar a eficiência. A solução do problema de LD é custosa devido ao grande volume e heterogeneidade semântica dos conjuntos de dados.

Apesar das mesmas técnicas se aplicarem a ambos os problemas, existem diferenças significativas entre o LD e a ER que levam ao desenvolvimento de ferramentas específicas para o LD. Isso ocorre porque a maioria das abordagens de ER concentra-se em conjuntos de dados homogêneos de objetos estruturados relativamente simples, descritos por um conjunto de atributos de valor único. Por outro lado, os recursos para o LD podem ser heterogêneos e altamente inter-relacionados com outros recursos dentro dos conjuntos de dados (KÖPCKE; THOR; RAHM, 2010).

[11] Para fins de desambiguação, foi alterada a notação utilizada no trabalho original.

Datasets como DBpedia[12] ou LinkedGeoData[13] geralmente utilizam uma ontologia que descreve seus recursos e os interligam através de propriedades pré-definidas. Conforme pode ser observado na LOD, o fato desses *datasets Alvos* utilizarem ontologias faz com que os *datasets Fontes* passem também a utilizá-las, facilitando a interligação com o uso de propriedades como *owl:sameAs*. Em (OLIVEIRA et al., 2019), quando esses recursos provêm de dois *datasets* diferentes, a interligação é chamada de externa e possibilita a realização de consultas que alcançam ambos os *datasets*.

Além disso, um ponto de cautela são as possíveis imprecisões no uso do *owl:sameAs*. Em (HALPIN; HAYES, 2010) esse assunto é tratado sob diferentes pontos de vista. Podemos destacar um possível uso do predicado para descrever a mesma coisa, porém em contextos diferentes. Para ilustrar, pode-se dizer que João tem dois papéis. Quando ele está em uma reunião de trabalho é o Sr. João, Assistente de Planejamento, com propriedades pertinentes a essa atividade, por exemplo, seu "horário de trabalho". Já em outro contexto, o Sr. João é um membro do clube de tênis, com propriedades como "reserva da quadra". Isso não significa que o Sr. João não reservou uma quadra quando está em uma reunião de trabalho mas, no contexto do trabalho do Sr. João, a "reserva" não importa. Sendo assim, observa-se que não é adequado descrever como sendo o mesmo João, já que existe essa diferença de papel conforme a situação. Então, na falta de um representante neutro para representar a pessoa João, podemos afirmar que a relação entre os papéis que o Sr. João desempenha nos dois *datasets* seria um predicado como "participa como" ou "atua como". Situações como essa acontecem na Web de Dados e mostram a complexidade do processo de LD.

Esse processo de LD pode ser aplicado tanto na etapa 1 quanto na etapa 2 de uma interligação entre *datasets* como a apresentada na Figura 6. Conforme observado na Figura 8, o processo se inicia com a leitura dos *datasets* S e T. Em seguida é realizada a atividade de pré-processamento (***Preprocessing***) com a utilização de parâmetros de configuração e recursos externos. Nessa atividade, são realizadas a preparação e a limpeza dos dados de entrada com a participação de um especialista no domínio. Na sequência acontecem o ***Instance Matching*** - IM e o ***Ontology Matching*** - OM. No OM, dadas duas ontologias O_1 e O_2, é retornado como saída o mapeamento com o resultado de cada elemento de O_1 que corresponde a um elemento de O_2. Já no IM, busca-se resolver o problema do reconhecimento quando diferentes instâncias se referem à mesma entidade do mundo real. Na sequência do processo de LD é realizado o ***Post-processing***, que consiste principalmente em selecionar os links de acordo com sua especificação, ou seja, de acordo com as condições que dois recursos devem satisfazer para que um vínculo entre eles possa ser estabelecido. Finalmente é representado o *dataset* ***Mapping***, onde são armazenados o conjunto de links descobertos no processo.

[12] https://wiki.dbpedia.org/
[13] http://linkedgeodata.org/

Vale destacar que tanto no OM quanto no IM são gerados os novos mapeamentos M com diferentes relações Ψ. Assim, Ψ pode inclusive ser definida para representar ligações além das relações de similaridade, como o exemplo visto na etapa 2 da Figura 6. No entanto nem sempre é possível estabelecer um critério formal para isso. Por exemplo, para definir se uma cidade é próxima a outra, torna-se necessária definir critérios para que essa proximidade geográfica seja considerada. Por outro lado, para definir se há uma relação entre *Belas Artes* e *Pintura clássica*, pode ser preciso ir além dos recursos disponíveis nos *datasets*.

O processo tradicional de LD foca em realizar a análise estrutural e sintática dos *datasets* na busca por relações, a partir de um critério bem definido. Já em (ATHANASIOU et al., 2019) os autores definem o processo de enriquecimento de *dataset* como um conjunto de ações mais amplo, que vão além daquelas previstas pelo processo tradicional de LD, de modo a permitir estabelecer critérios mais complexos para a descoberta de relações Ψ. Através dessas novas ações, é possível criar propriedades extras relacionando os recursos dos *datasets*, aumentando a riqueza e completude dos dados.

Por exemplo, em (SHERIF; NGOMO; LEHMANN, 2015) os autores realizaram esse processo de enriquecimento de *dataset* utilizando técnicas de PLN para analisar os textos associados aos recursos dos *datasets* (e.g. comentários, descrições, abstracts, etc.). Ao descobrir entidades dentro desses textos, são identificadas novas relações com os recursos em questão, rotuladas com o predicado padrão *fox:relatedTo*. Outra iniciativa (OLIVEIRA et al., 2019) utiliza técnicas de mineração em grafo para descobrir novas relações do tipo "relacionado com". Porém, ainda assim, a semântica das relações encontradas é vaga.

2.5 Extração de Relações

Em (MAYNARD; BONTCHEVA; AUGENSTEIN, 2016), os autores explicam que o Processamento de Linguagem Natural (do inglês *Natural Language Processing* – NLP) é o processamento automático de linguagem humana. As técnicas de PLN podem ser usadas para aprimorar dados da Web com semântica. Por exemplo, uma forma de fazer isso é identificando nos textos das páginas Web, referências a entidades do mundo real e suas relações, e atribuindo uma *Uniform Resource Identifier* – URI a cada uma dessas entidades.

O processo de extrair informações de um texto e transformá-las em dados estruturados, conhecido como *Information extraction* (IE), é uma das tarefas de NLP. Além de identificar entidades do mundo real (tarefa de *Named Entity Recognition* – NER), chamadas Entidades Nomeadas (ou *Named Entities* – NE's, em inglês), a IE deve ser capaz também de extrair as relações entre elas. Em (HAN et al., 2019), os autores afirmam que *Relation Extraction* – RE, uma subtarefa da IE, tem a função de extrair *links* entre as

NE's.

Existem vários tipos de informações interessantes que podem ser extraídas de textos não estruturados. As NE's são geralmente consideradas as principais componentes do texto. Elas podem ser, por exemplo, pessoas, locais, organizações, nomes próprios ou expressões temporais.

Um exemplo de um texto não estruturado com entidades e relações é apresentado na Figura 9. Nela são apresentadas três NE's: Albert Einstein, Alemanha e 1879. Além disso, são marcadas também duas relações: entre o cientista e o país e entre o cientista e seu ano de nascimento.

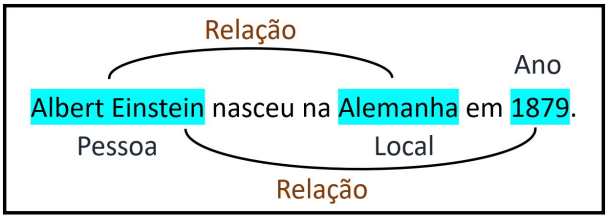


Figura 9 – Exemplo de entidades e relações.

Em (COLLOVINI et al., 2020), os autores afirmam que a tarefa de RE a partir de textos é um dos principais desafios na área de IE, considerando o conhecimento linguístico necessário e a sofisticação das técnicas de processamento de linguagem empregadas. O acesso a uma quantidade crescente de informações e a implantação de tecnologias da informação de ponta viabilizam a RE. Nas últimas décadas, esse tipo de trabalho foi realizado sem o apoio de ferramentas computacionais específicas. A atual tecnologia de NLP está atingindo a maturidade necessária para ajudar a organizar essas informações não estruturadas.

Embora sejam independentes, tratadas por trabalhos que focam especificamente em cada uma, é possível melhorar a eficiência das tarefas de NER e RE por meio da sua realização simultânea e, por exemplo, reduzir a propagação de erros que acontecem ao se realizar essas tarefas de forma sequencial (YU; LAM, 2010; LI; JI, 2014; MIWA; SASAKI, 2014). Pesquisas atuais utilizam aprendizado de máquina com esse objetivo (ZHOU et al., 2005; CHAN; ROTH, 2011; GUPTA; SCHÜTZE; ANDRASSY, 2016). Mais recentemente e com o mesmo propósito, um subconjunto dessa linha de pesquisa vem utilizando redes neurais artificiais, que são modelos computacionais inspirados pelo sistema nervoso central de um animal (MIWA; BANSAL, 2016; REN et al., 2017; WANG et al., 2018; FU; LI; MA, 2019).

2.6 Considerações Finais

A Web Semântica estende a Web clássica, provendo uma estrutura semântica para as páginas Web. Nela, os recursos são representados utilizando principalmente o formato RDF. A adoção desse formato favorece a interligação de dados, cuja representação mais conhecida é a LOD. Essa interligação de dados é favorecida com a adoção de vocabulários controlados e ontologias que estão disponíveis para reuso em repositórios como o LOV. Outro processo que também auxilia a incrementar o conhecimento sobre os dados é o de enriquecimento. Tanto no processo de interligação quanto no de enriquecimento podem surgir novas ligações com semântica pobre. O capítulo a seguir tratará de trabalhos relacionados a esse problema.

3 TRABALHOS RELACIONADOS

Muitos trabalhos têm realizado a interligação de *datasets* da *Web* de dados através de tarefas de LD, conforme apresentado na Seção 2.4. Nesse contexto, ferramentas como o Silk e o LIMES trazem a preocupação de realizar a interligação com predicados além do *owl:sameAs*. Outros trabalhos, como o MRAR+ e o DEER, estendem a tarefa de LD ao explorar mais recursos disponíveis nos *datasets* e, assim, encontrar relações de forma mais complexa. Uma outra linha de trabalhos relacionados vem contribuindo para a construção da WS através o uso de NLP. Nesta sessão serão apresentados trabalhos relacionados a essas linhas de pesquisa.

3.1 Silk

O Silk (BIZER et al., 2009) é uma das primeiras ferramentas de LD que procurou tratar o problema de interligação de modo mais genérico, permitindo que o usuário defina o critério e a propriedade que deve ser criada entre os recursos. O critério é definido através de regras especificadas manualmente tanto para links do *owl:sameAs* quanto para outros tipos de relacionamentos. Além disso, a ferramenta se apoia em aprendizado supervisionado. Ela utiliza métricas de similaridade de *strings* para realização da tarefa de interligação de *datasets*. A arquitetura da ferramenta está ilustrada na Figura 10. Nessa arquitetura, pode-se destacar a base ***Config***, de onde é lido o parâmetro em que o usuário especifica o link que será gerado na etapa ***Core Logic / Comparsion Loop***.

Por exemplo, utilizando a linguagem de especificação de link disponibilizada pela ferramenta, são definidos, entre outros, os parâmetros *LinkType*, que é o predicado que será utilizado para a ligação, e *LinkageRule*, que são as regras para que a ligação seja materializada. Essas definições ficam armazenadas na base *Config*. Uma demonstração de uso[1] com esses parâmetros é apresentada na Figura 11. Nela é possível observar que o predicado *rdrel:workManifested*[2] está configurado para ligar livros com título (*rdfs:label*) igual entre dois *datasets*. Assim, a semântica da ligação criada fica limitada ao que foi previamente especificado pelo usuário da ferramenta.

Com essa arquitetura, o Silk pode ser empregado em ambientes distribuídos sem que haja necessidade de se replicar localmente os *datasets* que serão interligados. Ele também pode ser usado em situações em que termos de vocabulários diferentes são misturados.

[1] https://bit.ly/3scznuc
[2] http://rdvocab.info/RDARelationshipsWEMI/workManifested

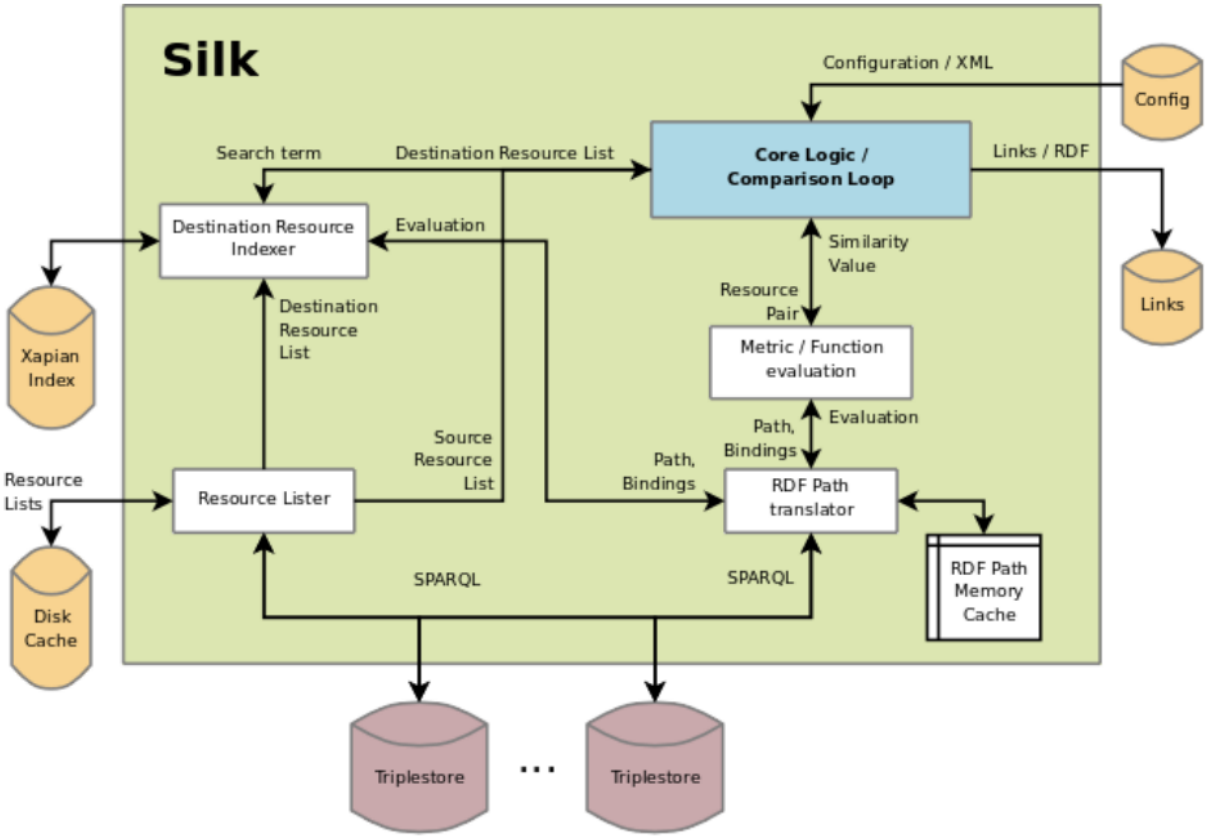


Figura 10 – Arquitetura do Silk (BIZER et al., 2009).

3.2 LIMES

Já o LIMES (NGOMO; AUER, 2011), assim como o Silk, é uma ferramenta que suporta tanto a configuração manual quanto as técnicas de aprendizado de máquina supervisionado (programação genética e aprendizado ativo), porém suporta também o aprendizado de máquina não supervisionado (também com programação genética). A ferramenta oferece diferentes técnicas de aproximação baseadas em espaços métricos para estimar as semelhanças entre instâncias. Semelhante ao Silk, ele pode produzir links *owl:sameAs* ou especificados pelo usuário, também limitando a semântica dos links gerados ao que foi previamente especificado pelo usuário.

O fluxo de trabalho geral implementado no LIMES compreende quatro etapas, conforme apresentado na Figura 12. No primeiro passo é reservado um conjunto de amostras do *dataset* alvo, no qual é realizado um cálculo de similaridade. Nas duas etapas seguintes, para cada item do *dataset* fonte é realizada uma análise com aproximação com os item reservados na primeira etapa. Na etapa final os dados são serializados no formato escolhido pelo usuário, possibilitando acrescentar os links gerados ao *dataset* fonte. Porém este tipo de abordagem é limitada pois não há como definir critérios para relações mais complexas como por exemplo, a relação de produção entre gene e proteína ("gene produz proteína") ou a relação de tratamento entre fármaco e doença ("fármaco trata doença"). Encontrar

```
<LinkType>rdrel:workManifested</LinkType>
<SourceDataset dataSource="lobid" var="b">
  <RestrictTo>
    ?b rdf:type bibo:Book
  </RestrictTo>
</SourceDataset>
<TargetDataset dataSource="dbpedia" var="a">
  <RestrictTo>
      ?a dcterms:subject category:Literarisches_Werk
  </RestrictTo>
</TargetDataset>
<LinkageRule>
  <Aggregate type="max">
    <Compare metric="equality">
      <TransformInput function="lowerCase">
        <TransformInput function="replace">
          <TransformInput function="regexReplace">
            <Input path="?a/rdfs:label"/>
          </TransformInput>
          <Param name="search" value="_"/>
          <Param name="replace" value=" "/>
        </TransformInput>
      </TransformInput>
      <TransformInput function="lowerCase">
        <Input path="?b/isbd:P1004"/>
      </TransformInput>
    </Compare>
  </Aggregate>
</LinkageRule>
```

Figura 11 – Trecho de um *script*[1] na linguagem de especificação do Silk. Os parâmetros *LinkType* e *LinkageRule* são utilizados para criação do link *rdrel:workManifested*.

essas relações mais complexas depende de buscar outras informações em fontes externas, talvez além dos dados do *dataset*.

3.3 DEER

No contexto de ferramentas que vão além do que o processo tradicional de LD propõe, temos a proposta do DEER (SHERIF; NGOMO; LEHMANN, 2015), que utiliza o LIMES mas chega a prover um enriquecimento maior do *dataset*. O algoritmo apresentado utiliza aprendizado de máquina supervisionado para realização do enriquecimento dos dados utilizando metadados que estão implícitos no *dataset* fonte do enriquecimento.

A ferramenta realiza tarefas que utilizam NLP para reconhecer todas as entidades incluídas no literal apontado pelo predicado *rdfs:comment*. No exemplo apresentado na Figura 13 (1), o resultado é um conjunto de entidades relacionadas, todas elas relativas

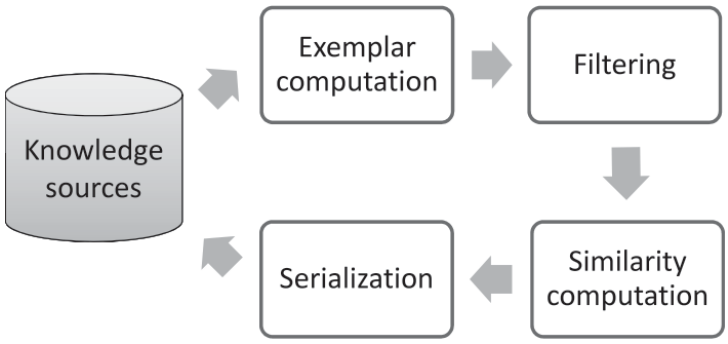


Figura 12 – Fluxo de trabalho geral do LIMES (NGOMO; AUER, 2011).

ao recurso *ex:Ibuprofen* pelo predicado padrão *fox:relatedTo*. Realizando uma análise no texto: *"Ibuprofen was extracted by the research arm of Boots Company during the 1960s ..."*, apontado pelo predicado *rdfs:comment*, a ferramenta consegue enriquecer semanticamente a relação para *ex:relatedCompany*, conforme Figura 13 (2). Porém a semântica da relação continua pobre, pois não esclarece, por exemplo, se *ex:Ibuprofen* é fabricado ou distribuído por *BootsCompany* e a ferramenta não consulta dados externos para realização da tarefa.

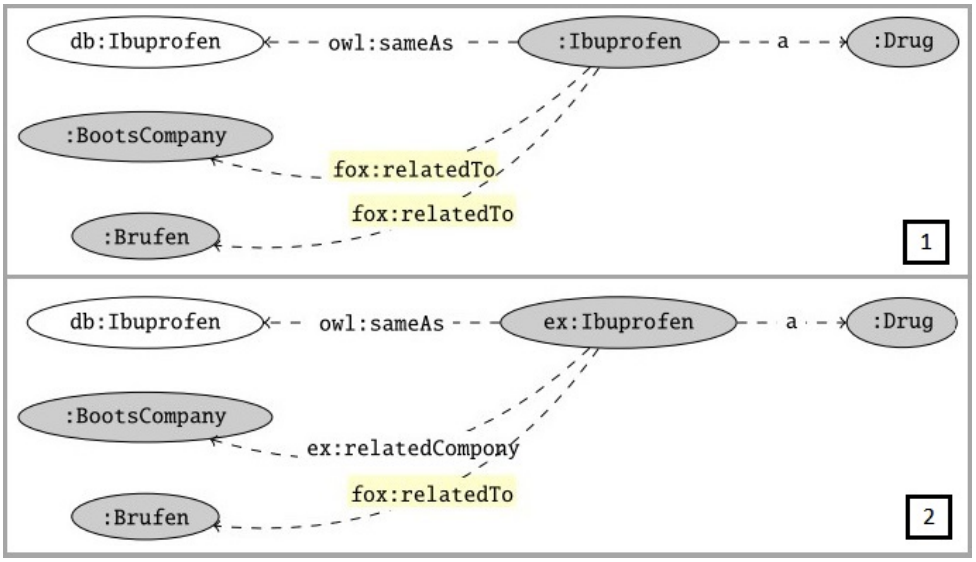


Figura 13 – Resultado do enriquecimento baseado na anotação do esquema. Adaptado de (SHERIF; NGOMO; LEHMANN, 2015).

3.4 LSVS

Em (AHMED; SHERIF; NGOMO, 2019), os autores apresentam uma abordagem, denominada *Link Specification Verbalization and Summarization - LSVS*, baseada em sumarização para descrever os links encontrados no processo de LD. A análise tem como resultado uma descrição sobre todo o conjunto de links gerados. O trabalho foi motivado

pela complexidade de se realizar a descoberta de links em *datasets* de quantidade e tamanho cada vez maiores.

Na abordagem, é realizada a comparação entre recursos de dois *datasets* distintos, por meio da verificação do atributo "*name*". Utilizando recursos de NLP foi possível descrever, em linguagem natural, os links previamente estabelecidos entre *datasets* e a condição em eles foram gerados, conforme exemplo extraído do artigo, de uma sumarização realizada a respeito dos links gerados entre dois *datasets*:

"The link will be generated if the title of the source and the target resources has a 66% of Cosine similarity"

Com essa abordagem é possível descrever, de forma ampla, o índice de similaridade conseguido na tarefa de LD, auxiliando no entendimento de como os *datasets* foram conectados, porém, não há incremento semântico nas relações descobertas.

3.5 MRAR+

Indo além das abordagens de LD, em (OLIVEIRA et al., 2019) os autores propõem um algoritmo para minerar regras de associação de multirrelação em mais de um *dataset*, chamado de MRAR+, desenvolvido como uma extensão do MRAR (RAMEZANI et al., 2014). Seu diferencial é se voltar para o enriquecimento de *datasets* pela busca por informações adicionais que não somente os atributos. Ele se apoia na estrutura topológica, isto é, nos caminhos entre os recursos, para descobrir potenciais associações entre eles.

No trabalho, a descoberta de regras de associação na Web de Dados foi viabilizada por meio da atribuição de uma máscara de busca durante o processo de mineração. Isso fez com que o algoritmo gerasse apenas as regras que estavam relacionadas aos recursos de maior frequência e que estivessem vinculados a recursos de *datasets* externos, reduzindo assim o custo computacional. A Figura 14 apresenta os passos seguidos pelo MRAR+.

O processo inicia-se com a mineração do *dataset* fonte, DtA, (passo 1), passando pela seleção de recursos externos (passo 2). Uma vez identificados os recursos externos, passa-se, então, para o processo de ampliação do conhecimento existente do DtA (passo 3), com as informações encontradas no *dataset* dos recursos externos (DtB), gerando o dataset $DtA+$. Em seguida (passo 4), minera-se esse novo conjunto de dados ($DtA+$) para encontrar novas regras. No passo 5, compara-se as regras geradas sobre o *dataset* DtA com as geradas pelo conjunto $DtA+$. Por fim, no passo 6, o usuário tem a opção de reajustar o valor de *MinSup* (suporte mínimo para que a regra seja considerada), para permitir que o valor utilizado seja suficiente tanto para a geração das regras novas quanto das antigas, encontradas na primeira execução do MRAR+.

O que se busca com esse processo são caminhos comuns, que partem do mesmo

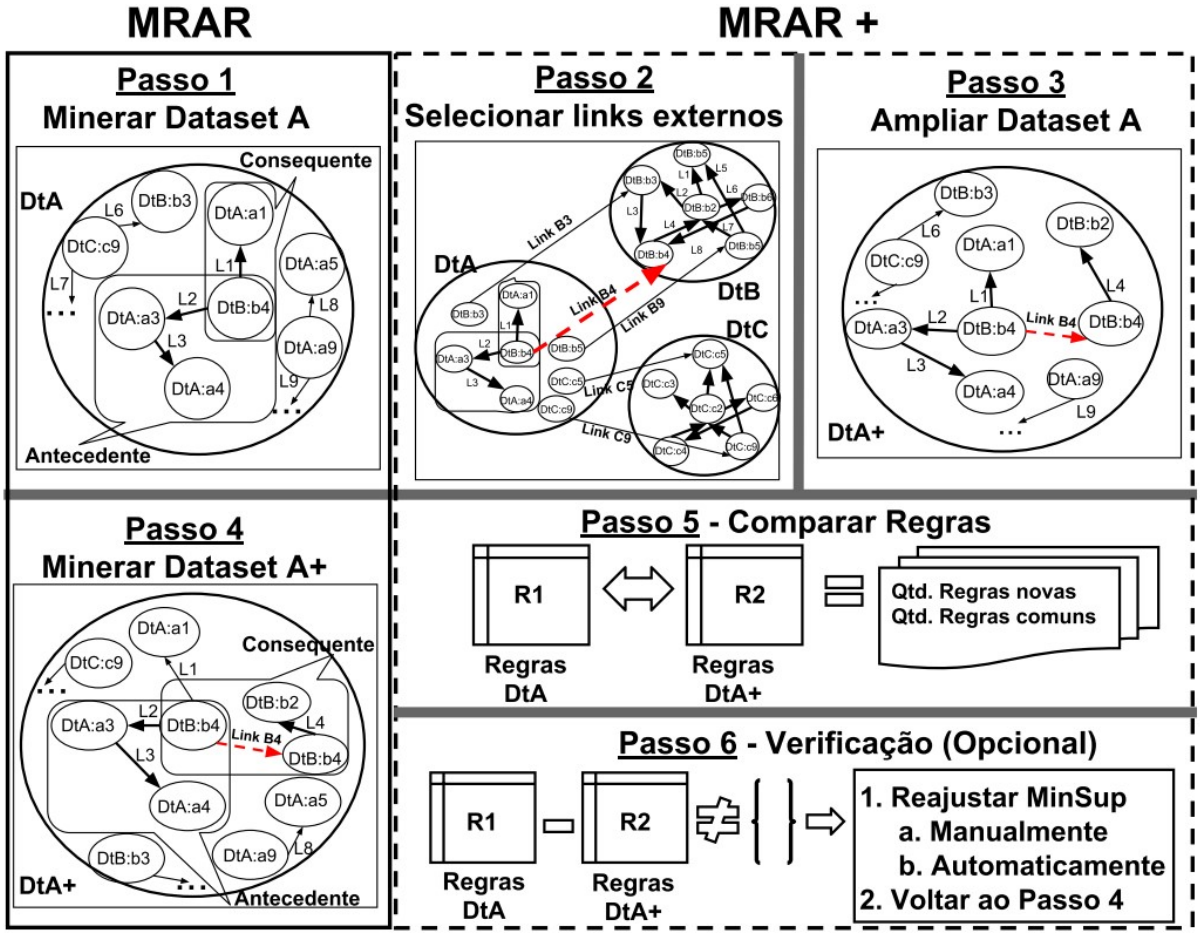


Figura 14 – Visão geral ilustrando cada um dos passos aplicados a proposta (OLIVEIRA et al., 2019).

tipo de recurso e que podem levar a associações inusitadas como a apresentada na Figura 6, entre Belas Artes e Pintura clássica. O resultado desse algoritmo é um conjunto de regras que associam recursos que possuem potencial para interligação. Observa-se que esse conjunto é passível de enriquecimento semântico em suas ligações, porém o algoritmo não chega a sugerir a semântica da relação entre os recursos. A observação dessa oportunidade de enriquecimento serviu de motivação para o desenvolvimento do presente trabalho.

3.6 BERT

Já no campo do NLP, observam-se pesquisas que podem apoiar a tarefa de RE e identificar links entre as NE's. Em (DEVLIN et al., 2018) os autores explicam que o desenvolvimento de um modelo de linguagem pré-treinado mostrou-se efetivo para essas finalidades. Nesse trabalho, os autores apresentam o modelo BERT (*Bidirectional Encoder Representations from Transformers*), cujos resultados estão de acordo com recentes avanços do NLP. Entre os índices de desempenho mais utilizados na área, o BERT superou

por exemplo, a pontuação GLUE[3] para 80,5% (7,7% de melhoria absoluta), a acurácia MultiNLI[4] para 86,7% (4,6% de melhoria absoluta), F1 do teste de *question answering* do SQuAD[5] v1.1 para 93,2 (1,5 ponto de melhoria absoluta) e o F1 do teste SQuAD v2.0 para 83,1 (5,1 pontos de melhoria absoluta).

O BERT é capaz de modelar uma linguagem e retornar uma *Word Embedding – WE*, que é uma representação em um vetor de números a partir de um texto de entrada, no qual palavras semelhantes têm valores semelhantes. Uma evolução importante na técnica utilizada pelo BERT é a aplicação do treinamento bidirecional do Transformer (VASWANI et al., 2017), que é um recurso capaz de aprender as relações contextuais entre as palavras de um texto. Isso é um diferencial em relação aos estudos anteriores que comumente analisavam uma sequência de texto apenas da esquerda para a direita ou da direita para a esquerda.

Em (SHI; LIN, 2019) é apresentada a arquitetura de um modelo para RE que utiliza o BERT na camada de WE. Para esse modelo devem ser dadas como entrada sentenças estruturadas no formato:

```
[CLS] sentença
[SEP] sujeito [SEP] objeto [SEP]
```

Nesse formato, o [CLS] é um símbolo adicionado na frente de cada exemplo de entrada e o [SEP] é um símbolo de separação. Por exemplo, pode ser dada como entrada para o modelo a sentença "*Obama was born in Honolulu*". Nessa sentença, "*Obama*" é a entidade sujeito, do tipo pessoa e marcada como [S-PER]. Já "*Honolulu*" é a entidade objeto, do tipo local e marcada como [O-LOC]. No exemplo, a estrutura fica:

```
[CLS] [S-PER] was born in [O-LOC]
[SEP] Obama [SEP] Honolulu [SEP]
```

Para a atividade de RE, acontece uma análise bidirecional realizada sobre as características de cada um dos termos da sentença dada como entrada. Essa análise é chamada bidirecional por ser realizada no contexto da esquerda para a direita e da direita para a esquerda. Então, com o apoio do modelo de linguagem pré-treinado, é possível para essa arquitetura de RE identificar a relação "*per:city_of_birth*" (cidade de nascimento) entre as entidades marcadas, tendo como resultado a relação: *Obama per:city_of_birth Honolulu*.

A arquitetura apresentada não foi utilizada no contexto de dados estruturados ou da Web de Dados, sendo aplicada somente para dados não estruturados.

[3] https://gluebenchmark.com/
[4] https://cims.nyu.edu/ sbowman/multinli/
[5] https://rajpurkar.github.io/SQuAD-explorer/

3.7 RelP++

Ainda na linha dos trabalhos voltados para RE, em (COLLOVINI et al., 2020) os autores apresentaram o RelP++, um *framework* que combina Reconhecimento de NEs e RE. Um dos diferenciais da ferramenta é seu foco em realizar tarefas de NLP em textos na língua portuguesa. Na Figura 15 é apresentada uma visão geral da sequência de funcionamento do processo com cinco etapas:

1. Pré-processamento, onde acontece a segmentação e tokenização das sentenças do texto;
2. Módulo NER, onde as entidades nomeadas recebidas nas sentenças do módulo anterior são identificadas;
3. Já o módulo seguinte faz a correspondência entre cada entidade identificada na etapa anterior com outra reconhecida e consolidada previamente. Assim gera pares de entidades nomeadas nas sentenças;
4. O módulo *Features* gera vetores de características, utilizando uma notação própria da ferramenta, a respeito de cada par de NE e das palavras entre essas entidades. Entre essas características estão a sua classe sintática e a sequência de outras palavras que podem ser adotadas em algumas posições anteriores e posteriores;
5. Finalmente, o módulo *CRF Model* se responsabiliza pela aplicação do modelo gerado a partir dos vetores de características. Para cada *Feature* recebida da etapa anterior é aplicado um peso que resulta em uma matriz de pesos que possibilita realizar a atividade de RE, conforme o exemplo apresentado no Quadro 1.

Quadro 1 – Exemplo do resultado da RE conseguida com o RelP++ (COLLOVINI et al., 2020)

Sentença	**Relação**
Durão Barroso discursava na sessão plenária do Parlamento Europeu.	discursar

Seguindo essas etapas, o *framework* é capaz de reconhecer as relações existentes entre as NEs, porém busca por relações contidas no próprio texto de entrada e fora do contexto da Web de Dados e do enriquecimento de *datasets*, como apresentado no Capítulo 2.

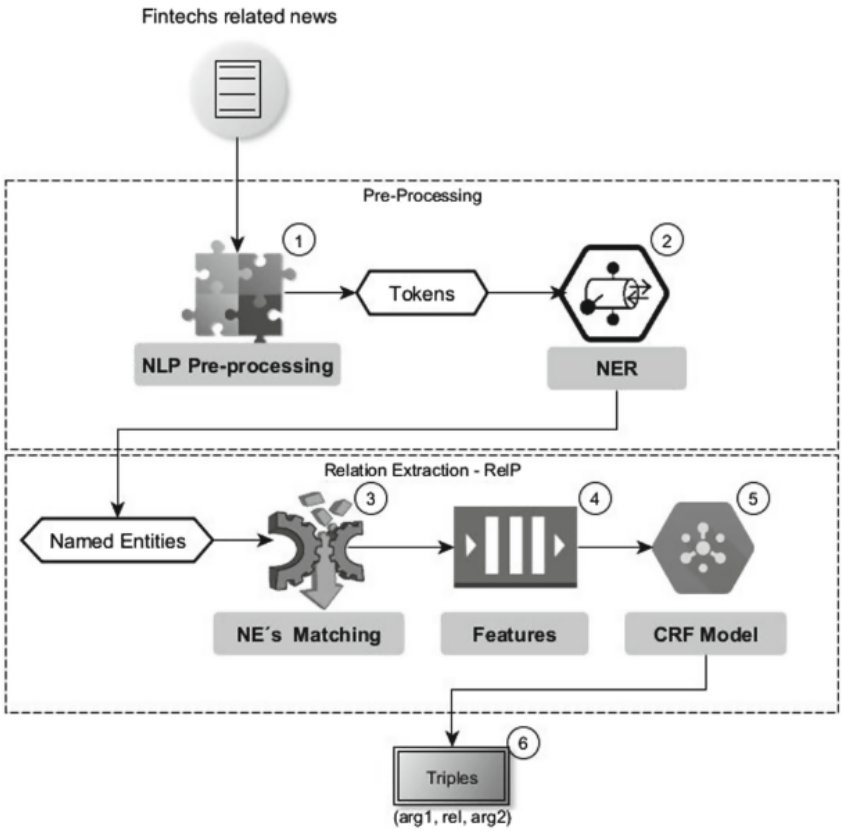


Figura 15 – Visão geral do *framework* RelP++ (COLLOVINI et al., 2020).

3.8 OpenNRE

Já em (HAN et al., 2019) apresenta-se o OpenNRE, um conjunto de ferramentas para desenvolver e aplicar modelos de RE, de forma a realizar a tarefa de RE entre NEs como o RelP++ e faz uso de modelos como o BERT para atingir seus objetivos.

Esse conjunto de ferramentas provê a possibilidade de se implementar modelos neurais para a tarefa de RE. Em um dos módulos é possível realizar essa tarefa a nível de sentença, considerando que esta foi previamente anotada, ou seja, que foram determinados os dois recursos sobre os quais se deseja saber a relação.

Por exemplo, pode-se oferecer para a ferramenta a seguinte sentença: "*Like Siegfried Sassoon, Schuster was of Jewish descent.*". Além da sentença, deve-se destacar as duas entidades as quais se deseja obter a relação, que neste caso são: "*Siegfried Sassoon*" e "*Jewish*". Utilizando o modelo BERT, apresentado na Seção 3.6, a ferramenta retorna a relação "*religion*" entre as entidades. Essa é uma das 80 relações disponíveis a partir do conjunto FewRel (HAN et al., 2018) que foi utilizado para treinar previamente o modelo.

A tela do terminal com a sequência de comandos utilizadas nesse exemplo pode ser observado na Figura 16. No modo interativo do Python[6] é executado o comando

[6] https://www.python.org/

import opennre que realiza a importação do pacote OpenNRE[7]. Na sequência, o comando *get_model* carrega o modelo pré-treinado. A seguir, com o comando *infer* realiza a extração de relação a nível de sentença. Finalmente a ferramenta retorna com a relação inferida e sua pontuação de confiança.

Nesse exemplo é inserida a sentença *"Like Siegried Sassoon, Schuter was of Jewish descent."*, e as posições dos termos cabeça (do inglês *head* - h) e cauda (do inglês *tail* - t) que, neste exemplo, são *Siegried Sassoon* e *Jewish*, respectivamente. A ferramenta retorna a relação de religião (do inglês *religion*), com uma pontuação de confiança de 0,9989.

```
C:\Windows\system32\cmd.exe - python
(OpenNRE2020) C:\OpenNRE-master>python
Python 3.7.9 (default, Aug 31 2020, 17:10:11) [MSC v.1916 64 bit (AMD64)] :: Anaconda, In
c. on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import opennre
>>> model = opennre.get_model('wiki80_bert_softmax')
>>> model.infer({'text': 'Like Siegfried Sassoon, Schuster was of Jewish descent.', 'h':
{'pos': (5, 22)}, 't': {'pos': (40, 46)}})
('religion', 0.9989545345306396)
>>>
```

Figura 16 – *Prompt* de comando com a utilização do OpenNRE para extração de relação a nível de sentença.

Conforme observado, o *framework* OpenNRE é uma solução que foca na aplicação das tarefas de NLP em textos e não visa enriquecer *datasets*.

3.9 Considerações Finais

Na Tabela 1 é apresentado um comparativo entre os trabalhos relacionados. As seguintes características foram utilizadas para classificar os trabalhos:

(C1) Realiza Análise sobre *Datasets* da Web de Dados;

- Trabalhos que focaram em tarefas com *datasets* da Web de Dados.

(C2) Enriquece Semanticamente as Relações com Dados do próprio *Dataset*;

- Trabalhos que utilizaram recursos dos próprios *datasets* fonte e alvo para auxiliar no enriquecimento das relações.

(C3) Enriquece Semanticamente as Relações com Dados de outros *Datasets*;

- Trabalhos que utilizaram *datasets*, além dos fonte e alvo, para auxiliar no enriquecimento das relações.

[7] https://github.com/thunlp/OpenNRE

(C4) Utiliza NLP para extrair relações;

- Trabalhos que utilizaram a tarefa de RE.

(C5) Utiliza Textos Externos além do treinamento;

- Quando foram utilizados textos além dos constantes nos recursos dos *datasets* para auxiliar na rotulação de relações.

(C6) Utiliza Catálogos de Vocabulários Controlados.

- Quando houve a utilização desse tipo de catálogo na busca pelo enriquecimento semântico.

Tabela 1 – Comparação entre os trabalhos relacionados.

Trabalho	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
Silk (BIZER et al., 2009)	X					
LIMES (NGOMO; AUER, 2011)	X					
MRAR (RAMEZANI et al., 2014)	X					
DEER (SHERIF; NGOMO; LEHMANN, 2015)	X	X		X		
BERT (DEVLIN et al., 2018)				X		
MRAR+ (OLIVEIRA et al., 2019)	X					
LSVS (AHMED; SHERIF; NGOMO, 2019)	X	X	X			
OpenNRE (HAN et al., 2019)			X	X		
RelP++ (COLLOVINI et al., 2020)			X	X		
Nossa proposta	**X**	**X**	**X**	**X**	**X**	**X**

Entre outros aspectos, buscou-se por pesquisas na área de realização de análise sobre *datasets* da Web de Dados. Esse foi o aspecto predominante observado entre os trabalhos investigados. Com essa característica foram levantados os trabalhos que apresentaram as ferramentas: Silk (BIZER et al., 2009), LIMES (NGOMO; AUER, 2011), MRAR (RAMEZANI et al., 2014) e MRAR+ (OLIVEIRA et al., 2019).

Outro aspecto investigado foi a respeito da realização do enriquecimento semântico das relações entre os recursos. Sob essa condição, observou-se que dois trabalhos, com o DEER (SHERIF; NGOMO; LEHMANN, 2015) e a LSVS (AHMED; SHERIF; NGOMO, 2019), utilizam dados contidos no próprio *dataset* para o enriquecimento, sendo que o segundo utiliza também dados de outros *datasets* com finalidade semelhante.

Na sequência, também como uma característica pertinente a esta pesquisa, foi levantada também a utilização do NLP para realizar extração de relações. No contexto da Web de Dados, no DEER (SHERIF; NGOMO; LEHMANN, 2015), acontece a extração utilizando como origem os textos dos comentários do próprio *dataset* alvo do enriquecimento.

Ainda sobre a utilização de NLP para extração de relações, foram investigados três trabalhos, com o modelo BERT (DEVLIN et al., 2018) e os *frameworks* OpenNRE (HAN et al., 2019) e RelP++ (COLLOVINI et al., 2020). Essas pesquisas atendem a todas as etapas necessárias para a extração de relações em conjuntos de textos e seus resultados estão de acordo com recentes avanços do NLP.

A partir da análise dos trabalhos relacionados, observou-se que nenhum deles utiliza recursos externos, como, por exemplo, vocabulários ou textos externos, para nomear as relações entre itens de *datasets* distintos. A utilização desses outros recursos deve favorecer a proposição de sugestões mais adequadas. Além disso, alguns trabalhos, mesmo visando a descoberta de links e o enriquecimento de *datasets* no contexto da Web de dados, não levam em conta o uso de técnicas de RE combinado à exploração de catálogos de recursos semânticos (vocabulários, ontologias, etc.). Essa combinação deve diversificar e complementar as sugestões oferecidas.

Vale ainda destacar que nas soluções tradicionais de LD, e mesmo nas soluções mais avançadas (de enriquecimento), nem sempre é possível definir um critério para criar e rotular relações semânticas entre recursos de um *dataset*. Por exemplo, é difícil "descobrir" que um dado fármaco é *produzido* por um certo laboratório, como visto nas Seções 3.3 (Deer) e 3.5 (MRAR+). Nessas soluções o máximo que se consegue é saber que o fármaco está *relacionado com* o laboratório. Assim, fica a lacuna, como encontrar o rótulo mais apropriado para descrever a semântica dessas associações?

Como foi visto, a técnica de RE é normalmente aplicada a conjuntos de textos. No entanto, o presente trabalho vislumbra a sua aplicabilidade no enriquecimento semântico das ligações entre *datasets* na Web de Dados. Assim, o método descrito a seguir procura aliar a técnica de RE com este objetivo, indo além da informação contida nos *datasets* e recursos semânticos, como ontologias e vocabulários.

Até onde foi possível investigar, nenhum dos trabalhos relacionados identifica claramente a semântica das novas relações descobertas (exceto relações do tipo *owl:sameAs*), e consequentemente, não é possível rotular tais relações com clareza. Sendo assim, o presente trabalho apresenta uma proposta para preencher esta lacuna, como descrito na próxima seção.

4 PROPOSTA

O problema que buscamos tratar é caracterizado pela ausência de semântica constatada nos links gerados durante o processo de interligação de *datasets*. Observando os trabalhos publicados é possível notar que, após a mineração de links, outras relações podem ser inferidas e devidamente rotuladas.

Os processos do estado da arte focam no reconhecimento das relações mas deixam de realizá-las ou o fazem de forma limitada e sem o devido enriquecimento semântico. Haveria uma alternativa para realizar esse enriquecimento? Diante desse problema, a hipótese levantada é de que o uso de ontologias e vocabulários controlados combinado a técnicas de RE, podem favorecer a identificação e concepção dessas relações.

4.1 Método Proposto

O método proposto, denominado Predicate Labeling (SILVEIRA; CAVALCANTI, 2020), parte de um conjunto de relações que precisam de um rótulo com mais semântica e busca apoio nos vocabulários e ontologias existentes para sugerir novos rótulos. Além disso, o método proposto inclui tarefas complementares usando RE.

O método Predicate Labeling toma como base as seguintes definições:

Def.1 Seja D um *dataset* definido pela tupla $<C,I,R,A>$, onde:

- $c_i \in C$, i.e., C é um conjunto formado por classes c_i;
- $e_i \in I$, i.e., I é um conjunto formado por instâncias e_i;
- $x_i \in X$, onde X é um conjunto de recursos x_i, tal que cada recurso x_i pode ser uma classe ($x_i \in C$) ou uma instância ($x_i \in I$), i.e., $X = C \cup I$;
- $r_k \in R$, onde R é um conjunto de relações r_k que interligam recursos (sujeitos) a outros recursos (objetos).
- Seja A um conjunto de triplas (x_i, r_k, x_j), onde a relação $r_k \in R$ liga o recurso x_i ao x_j, e $x_i, x_j \in X$.

Def.2 Seja $(e_i, \mathit{type}, c_j) \in A$ uma tripla especial que representa uma instanciação, onde e_i é uma instância da classe c_j.

Def.3 Seja M um *dataset* de triplas formado por elementos de S e T, i.e., $M = \{(x_i, r_k, x_j) \,|\, (x_i, x_j) \in S \times T \wedge M \subset S\}$ (Seção 2.4).

Def.4 A função $subClassOf()$ é uma função tal que, dadas as classes c_i, $c_j \in C$, se existir a tripla (c_i, *rdfs:subClassOf*, c_j), então $subClassOf(c_j) = \{c_i\}$.

Def.5 A função $superClassOf()$ é uma função tal que, dadas as classes c_i, $c_j \in C$, se existir a tripla (c_j, *rdfs:subClassOf*, c_i), então $superClassOf(c_i) = \{c_j\}$.

Def.6 A função $ancestorOf()$ é uma função tal que, dadas as classes c_i, $c_j \in C$, $ancestorOf(c_i)$ tem como resultado todas as classes superiores no ramo hierárquico em que se encontra c_i, ou mais formalmente,
$ancestorOf(c_i) = \{c_j \mid \exists c_j = superClassOf(c_i)\} \cup ancestorOf(c_j)$.

Def.7 A função $descendantOf()$ é uma função tal que, dadas as classes $c_i, c_j \in C$, $descendantOf(c_i)$ tem como resultado todas as classes inferiores no ramo hierárquico em que se encontra c_i, ou mais formalmente,
$descendantOf(c_i) = \{c_j \mid \exists c_j = subClassOf(c_i)\} \cup descendantOf(c_j)$.

Def.8 A função $equivalentClass()$ é uma função tal que, dada uma classe $c_i \in C$, $equivalentClass(c_i)$ obtém como resultado todas as classes $c_j \neq c_i$ que contenham o mesmo conjunto de instâncias I.

Def.9 O conjunto C_{x_i} é um conjunto de classes que resulta da aplicação das funções definidas anteriormente (**Def.6**, **Def.7** e **Def.8**), mais formalmente pode-se dizer que $C_{x_i} = \{c_i\} \cup ancestorOf(c_i) \cup descendantOf(c_i) \cup equivalent(c_i)$,
onde $x_i = \begin{cases} e_i \in I \mid \exists\, (e_i, type, c_j) \in A \text{ (conforme } \textbf{Def.2}\text{), ou} \\ c_i \in C \end{cases}$

Def.10 A função $domain^{-1}()$ é uma função tal que, dada uma classe $c_i \in C$, obtém-se como resultado todas as relações r_k que podem ter c_i como sujeito. Mais formalmente, $domain^{-1}(c_i) = \{r_k \in R \mid \exists (c_i, r_k, x_j),\ x_j \in X\}$.

Def.11 A função $range^{-1}()$ é uma função tal que, dada uma classe $c_j \in C$, obtém-se como resultado todas as relações r_k que podem ter c_j como objeto. Mais formalmente, $range^{-1}(c_j) = \{r_k \in R \mid \exists (x_i, r_k, c_j),\ x_i \in X\}$.

Def.12 O conjunto P_{ij} é um conjunto de predicados que resulta da aplicação das funções definidas anteriormente (**Def.10** e **Def.11**), mais formalmente pode-se dizer que $P_{ij} = domain^{-1}(c_i) \cup range^{-1}(c_j)$.

Def.13 Seja o conjunto de tuplas $L_{ij} = \{\ t =< x_i, p, x_j > \mid x_i \in C_{x_i},\ p \in P_{ij} \text{ e } x_j \in C_{x_j}\}$, formado após a exploração dos catálogos de vocabulários controlados.

4.1.1 Visão geral do processo

A Figura 17 apresenta uma visão geral do método *Predicate Labeling*, através de um diagrama que utiliza a notação BPMN[1] (*Business Process Model and Notation*). Inicia-se o processo com a atividade **Ler Base de Dados Conectados** que faz leitura do *dataset* M (**Def.3**), que é um subconjunto de S ($M \subset S$). Nesta atividade são obtidas as triplas ($(x_i, r_k, x_j) \in M$) cujas relações r_k precisam ser rotuladas, pois possuem semântica pobre. Para cada tripla de M, as atividades seguintes são realizadas. Inicialmente o fluxo segue para a atividade **Consultar Classes e Propriedades correlatas**, que realiza uma exploração em catálogos de vocabulários controlados e ontologias. Neste ponto, o primeiro conjunto de sugestões de rótulos é criado.

Então, é oferecida para o usuário a oportunidade de incluir mais sugestões de rótulos usando técnicas de RE (**Identificar potenciais rótulos utilizando RE**). Finalmente, na atividade **Interagir com o usuário**, é oferecida para o usuário uma lista de candidatos a rótulos e ele pode ou não escolher algum desses rótulos. Todos os rótulos conseguidos nas etapas anteriores ficam disponíveis na lista. De acordo com a opção do usuário, uma das atividades: **Atualizar Predicado** ou **Manter Predicado Original** é realizada. Com base em cada nova escolha de predicado, uma nova tripla é então criada e adicionada no *dataset* **Triplas Rotuladas** (M'). Após repetir o processo para cada tripla M, acontece a atividade **Reagrupar base**, que atualiza as triplas semanticamente pobres de S (triplas M), com triplas M', e, assim, gera o *dataset* S'.

Na Consulta 1, representada em SPARQL, é apresentado um exemplo de seleção de triplas cujo predicado está como *ex:RelatedTo*. Neste exemplo observa-se na linha 1 a definição do prefixo que será utilizado na consulta. Na linha 2 acontece o comando da seleção de triplas completas, contando com sujeito, predicado e objeto. Finalmente, nas linhas seguintes é definido um filtro, para que sejam retornadas apenas as triplas com uma relação de semântica pobre que, neste exemplo, é o predicado *ex:RelatedTo*.

Consulta 1: Ler Base de Dados Conectados

Entrada: Base de dados conectados
Saída: Triplas com semântica pobre

```
1 PREFIX ex: <http://example.com/>
2 SELECT ?subject ?predicate ?object
3 WHERE {
4         ?subject ?predicate ?object ;
5                 ex:RelatedTo ?object .
6 }
```

[1] http://www.bpmn.org/

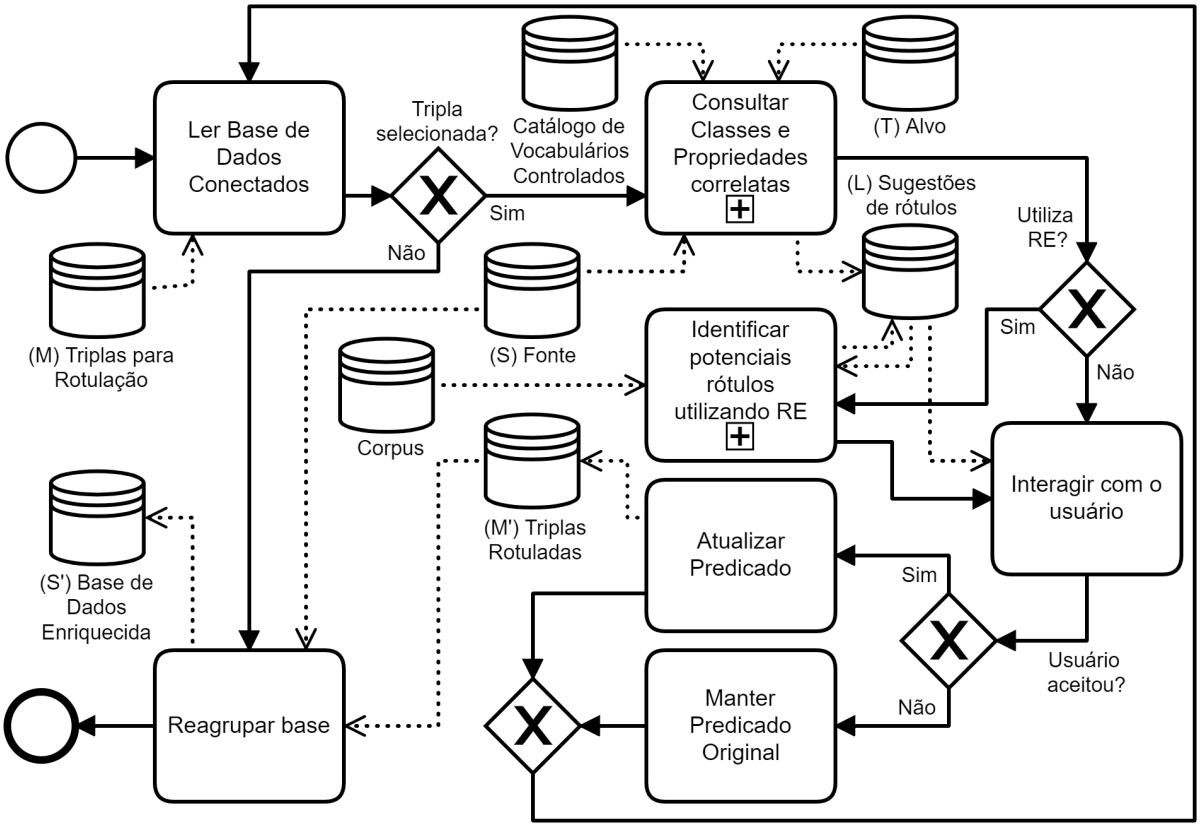


Figura 17 – Diagrama do método Predicate Labeling em BPMN.

4.1.2 Consultar Classes e Propriedades correlatas utilizando recursos semânticos

O fluxo principal então segue para a atividade **Consultar Classes e Propriedades correlatas** que é onde acontece a exploração de ontologias e **Catálogos de Vocabulários Controlados**. Todo o fluxo dessa atividade está detalhado no diagrama da Figura 18.

A sequência inicia-se com a atividade **Verificar se o recurso é Classe ou Instância**, que se desdobra em duas atividades paralelas exclusivas. Caso o recurso seja uma instância, ou seja, $x_i \in I_S$ e a informação sobre a sua classe não estiver formalizada em S, o fluxo paralelo exclusivo segue para a atividade **Buscar Classe em Catálogos de Vocabulários Controlados**, de forma a auxiliar nessa formalização.

Para cada recurso x_i em triplas $(x_i, r_k, x_j) \in A_S$, se $x_i \in I_S$ (onde S é dado por $<C_S,I_S,R_S,A_S>$), o fluxo segue para a atividade **Buscar Classe mapeada na Base de Dados de Origem** que realiza uma busca em S por sua classe correspondente c_i, ou seja, busca pela tripla $(x_i, type, c_i) \in A_S$ (**Def.2**). Caso c_i exista ou $x_i \in C_S$, ele será aproveitado nas consultas seguintes da atividade **Buscar Classes equivalentes**, com o uso das funções *ancestorOf()*, *descendantOf()* e *equivalentClass()* (**Def.6**, **Def.7** e **Def.8**, respectivamente). Com a aplicação dessas funções sobre c_i é construído o conjunto C_{x_i},

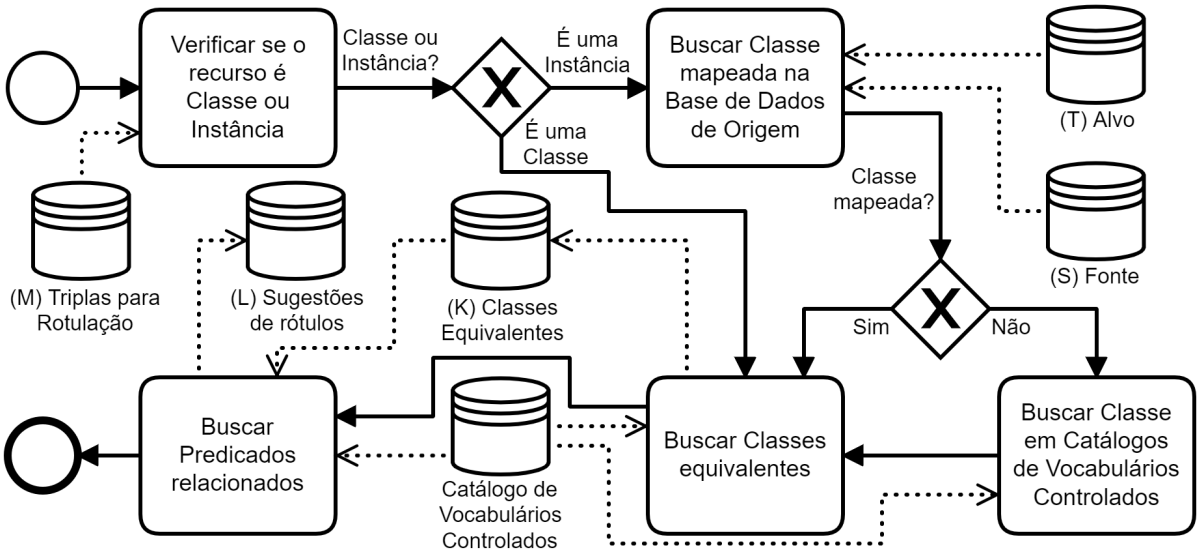


Figura 18 – Detalhamento da atividade Consultar Classes e Propriedades correlatas em BPMN.

conforme definido em **Def.9**, contendo, além da própria classe, todas as suas classes ancestrais, descendentes e equivalentes.

De forma a ilustrar essa busca, na Consulta 2 é apresentada uma consulta SPARQL a um catálogo de vocabulários. Tendo como entrada uma classe de interesse, a consulta retorna uma lista de classes equivalentes a ela. Por exemplo, ao se aplicar a Consulta 2 no catálogo do LOV, utilizando como entrada a classe *dbo:Plant*, tem-se como resultado o conjunto de classes apresentado no Quadro 2.

Analogamente à busca feita por c_i, que corresponde ao sujeito x_i da tripla em foco, faz-se também uma busca pelo objeto x_j, sendo que para x_j a busca por c_j é realizada em T. Com isso, o conjunto C_{x_j} é um conjunto obtido de forma semelhante ao C_{x_i}, porém com sua composição baseada em consultas a T.

Com os conjuntos C_{x_i} e C_{x_j} formados, o fluxo segue para a atividade **Buscar Predicados relacionados**, onde são utilizadas as funções $domain^{-1}()$ e $range^{-1}()$, definidas em **Def.10** e **Def.11**, quando é formado o conjunto P_{ij}, definido em **Def.12**.

Na Consulta 3 é ilustrada uma busca por predicados relacionados utilizando SPARQL. Dada uma classe de interesse, são retornados os predicados que têm essa classe como *domain* ou como *range*. Para exemplificar, executou-se essa consulta no catálogo do LOV[2], utilizando como entrada a classe *dbo:Plant*, tem-se como resultado o conjunto de predicados apresentados na Figura 19. Da forma como o resultado foi estruturado, observa-se com clareza quais são os predicados que são *domain* e *range* da classe.

Essa estruturação do resultado conta com o *label* do predicado com a informação

[2] https://lov.linkeddata.es/dataset/lov/sparql

Consulta 2: Consulta classes equivalentes à uma classe de interesse.

Entrada: Classe de interesse

Saída: Classes equivalentes à classe de interesse

```
1  PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
2  PREFIX owl: <http://www.w3.org/2002/07/owl#>
3  PREFIX prov: <http://www.w3.org/ns/prov#>
4  SELECT DISTINCT * WHERE {
5        {
6              SELECT ?type WHERE {
7                    ClasseDeInteresse owl:equivalentClass* ?type .
8              }
9        } UNION {
10             SELECT ?type WHERE {
11                   ?type owl:equivalentClass* ClasseDeInteresse .
12             }
13       } UNION {
14             SELECT ?type WHERE {
15                   ?type rdfs:subClassOf* ClasseDeInteresse .
16             }
17       } UNION {
18             SELECT?type WHERE {
19                   ClasseDeInteresse rdfs:subClassOf* ?type .
20             }
21       } UNION {
22             SELECT ?type WHERE {
23                   ?type prov:wasDerivedFrom* ClasseDeInteresse .
24             }
25       } UNION {
26             SELECT ?type WHERE {
27                   ClasseDeInteresse prov:wasDerivedFrom* ?type .
28             }
29       }
30 }
```

	predicate	predicateLabel	predicateURI	domainRange
1	CULTIVAR	"cultivar"@en	dbo:cultivatedVariety	D (->)
2	HYBRID	"hybrid"@en	dbo:hybrid	D (->)
3	HYBRID	"hybrid"@en	dbo:hybrid	R (<-)
4	PLANT	"plant"@en	dbo:plant	R (<-)

Showing 1 to 4 of 4 entries

Figura 19 – Resultado da consulta a predicados relacionados à classe *dbo:Plant* no LOV.

Quadro 2 – Resultado da realização da consulta apresentada na Consulta 2 ao LOV pela classe de interesse *dbo:Plant*.

Classes Equivalentes
dbo:Plant
http://www.wikidata.org/entity/Q756
dbo:CultivatedVariety
dbo:Conifer
dbo:Cycad
dbo:FloweringPlant
dbo:Grape
dbo:ClubMoss
dbo:Fern
dbo:Ginkgo
dbo:Gnetophytes
dbo:GreenAlga
dbo:Moss
dbo:Eukaryote
dbo:Species
owl:Thing
http://mappings.dbpedia.org/index.php/OntologyClass:Plant

da língua (inglês). Em seguida é apresentada a URI e, por fim, a informação se a classe de interesse está ligada a esse predicado como *domain* (D (->)) ou como *range* (R (<-)).

4.1.3 Identificar potenciais rótulos utilizando RE

A partir desta atividade, o método Predicate Labeling utiliza também as seguintes definições:

Def.14 Seja SE um *dataset* composto por um conjunto de elementos do tipo $<h, t, s>$, onde h é um termo que deve figurar como cabeça (do inglês *head*) da sentença, t é o termo cauda (do inglês *tail*) da sentença e s é a sentença em questão, que foi previamente selecionada em um **Subcorpus**.

Def.15 O conjunto Q é formado por triplas (h, q, t), cujas relações q são resultantes do processo de extração de relações entre h e t.

Com a realização da sequência de atividades detalhada pelo fluxo da Figura 18, é possível construir o conjunto de tuplas $t =< x_i, p, x_j >$, que vão integrar um *dataset* L_{ij} (**Def.13**). Esse *dataset* irá prover opções de rótulos para o usuário, que poderá utilizá-los nas ligações pobres com a ajuda das informações obtidas em catálogos de vocabulários controlados.

Consulta 3: Consulta predicados relacionados à uma classe de interesse.

Entrada: Classe de interesse
Saída: Predicados relacionados à classe de interesse

```
1  PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
2  PREFIX dbo: <http://dbpedia.org/ontology/>
3  SELECT DISTINCT * WHERE {
4        {
5              SELECT DISTINCT (UCASE (STR (?predicateLabel)) as ?predicate)
6                                    ?predicateLabel
7                                    ?predicateURI
8                                    ("D (->)" as ?domainRange) WHERE {
9                    ?predicateURI rdfs:domain ClasseDeInteresse .
10                   ?predicateURI rdfs:label ?predicateLabel .
11                   FILTER (LANG (?predicateLabel) = "en")
12             }
13       } UNION {
14             SELECT DISTINCT (UCASE (STR (?predicateLabel)) as ?predicate)
15                                   ?predicateLabel
16                                   ?predicateURI
17                                   ("R (<-)" as ?domainRange) WHERE {
18                   ?predicateURI rdfs:range ClasseDeInteresse .
19                   ?predicateURI rdfs:label ?predicateLabel .
20                   FILTER (LANG (?predicateLabel) = "en")
21             }
22       }
23 }
```

Na atividade seguinte do diagrama da Figura 17, **Identificar potenciais rótulos utilizando RE**, é realizado o processo de *Supervised Relation Extraction* (HAN et al., 2019). O fluxo detalhado da atividade é apresentado na Figura 20.

A atividade se inicia por **Buscar textos**, quando acontecem as consultas ao *dataset* L. É realizada então uma busca em um **Corpus**, no mesmo contexto do *dataset* S, por textos que façam referência a recursos disponíveis no *dataset* L_{ij}. Sendo assim, para cada par (x_i, x_j), contido nas triplas de L_{ij}, serão buscados textos que façam referência a x_i e x_j em um **Corpus** e todos esses textos serão armazenados na base **Subcorpus**.

Na sequência, a atividade **Buscar sentenças** realiza a leitura do **Subcorpus** e as sentenças s são identificadas. Em cada s é realizada uma busca para identificar a presença dos pares x_i, x_j. Então, a tupla $<h, t, s>$ que contenham esses pares de termos são então gravados na base SE (**Def.14**), sendo $h_i = x_i$ e $t_j = x_j$.

Já com a base SE construída, acontece a atividade **Realizar Extração de Relações** quando ocorre a extração supervisionada de relações. Fazendo uso de um **Modelo Pré-treinado** no contexto dos *datasets* S *e* T, é possível selecionar uma relação mais adequada entre as relações disponíveis na base de **Relações Predeterminadas**. O resultado

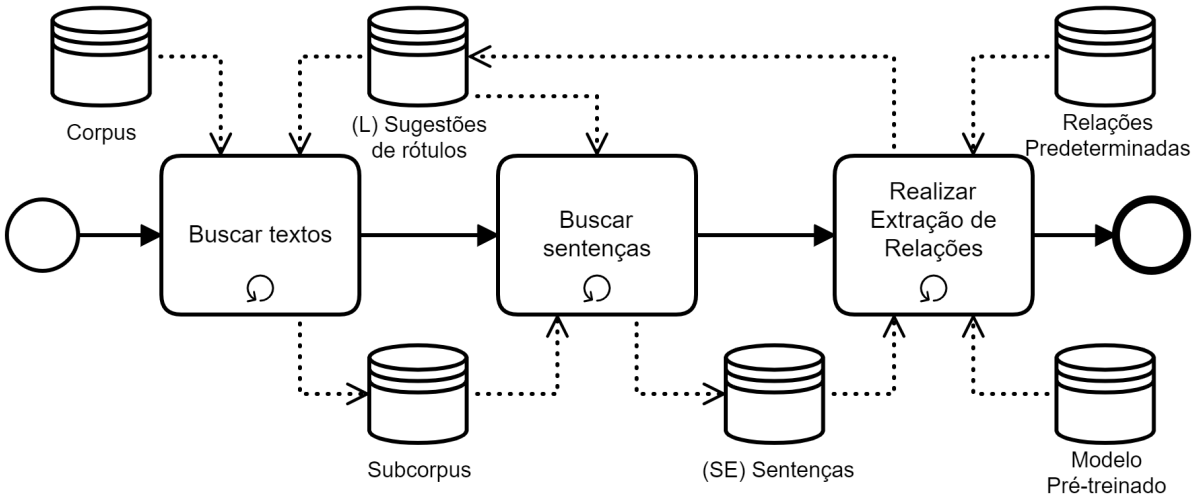


Figura 20 – Detalhamento da atividade Identificar potenciais rótulos utilizando RE em BPMN.

dessa atividade é o conjunto Q_{ij} (**Def.15**), que é formado por triplas (h_i, q_{ij}, t_j), onde cada q_{ij} é uma nova sugestão de rótulo. Finalmente, o *dataset* L é incrementado com Q_{ij}. Mais formalmente, $L = L \cup Q_{ij}$.

No Algoritmo 1 é ilustrada essa atividade. Na primeira iteração (linhas 1 a 3), acontece a constituição do **Subcorpus** com a concatenação de textos presentes em um **Corpus** que contenham os pares (x_i, x_j). Na segunda iteração (linhas 4 a 6), acontece a busca no **Subcorpus** pelas **Sentenças (SE)** em que estejam presentes os pares (x_i, x_j). Já na terceira iteração (linhas 7 a 9), é realizada a construção do conjunto **Q**, com as sugestões de rótulos com base em cada sentença conseguida após as iterações anteriores. Finalmente, na linha 10, O conjunto de rótulos original **L** é incrementado com os novos rótulos conseguidos nesta tarefa. Para exemplificar, foi representada uma iteração desse algoritmo, utilizando o OpenNRE, tendo como entrada uma sentença "*As plants transpire, the humidity around saturates leaves with water vapor.*" (s), onde buscou-se extrair a relação entre os termos *plants* (h) e *humidity* (t). Na Figura 21 é apresentado o resultado que teve como sugestão o predicado *main subject* (q).

4.1.4 Interagir com o usuário

Em seguida, o fluxo unifica-se na atividade **Interagir com o usuário**, onde são sugeridas ao usuário as opções para nomeação de cada ligação trabalhada. Estarão disponíveis todas as opções levantadas nas atividades anteriores (**Consultar Classes e Propriedades correlatas** e **Identificar potenciais rótulos utilizando RE**).

Algoritmo 1: Identificar Potenciais Rótulos Utilizando RE

Entrada: Corpus
Sugestões de Rótulos (L)
Subcorpus
Modelo Pré-treinado (PTM)
Relações Predeterminadas (PR)

Saída: Sugestões de Rótulos (L)

1 **para cada** $(x_i, x_j) \in L$ **faça**
2 | $Subcorpus \leftarrow Subcorpus +$ `buscarTextos(`(x_i, x_j), $Corpus$`)`;
3 **fim**
4 **para cada** $(x_i, x_j) \in L$ **faça**
5 | $SE \leftarrow$ `buscarSentencas(`(x_i, x_j), $Subcorpus$`)`;
6 **fim**
7 **para cada** $(h_i, t_i, s_i) \in SE$ **faça**
8 | $Q \leftarrow Q +$ `relationExtraction(`(h_i, t_i, s_i), PTM, PR`)`;
9 **fim**
10 $L \leftarrow L + Q$;
11 **retorna** L

```
C:\Windows\system32\cmd.exe - python
(OpenNRE2020) C:\OpenNRE-master>python
Python 3.7.9 (default, Aug 31 2020, 17:10:11) [MSC v.1916 64 bit (AMD64)] ::
Anaconda, Inc. on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import opennre
>>> model = opennre.get_model('wiki80_bert_softmax')
>>> model.infer({"text": "As plants transpire, the humidity around saturates
leaves with water vapor.", "h": {"pos": [3, 9]}, "t": {"pos": [25, 33]}})
('main subject', 0.7907416224479675)
>>>
```

Figura 21 – *Prompt* de comando com a utilização do OpenNRE para representar uma iteração da atividade Identificar potenciais rótulos utilizando RE.

4.1.5 Atualizar Predicado ou Manter Predicado Original

A sequência se desdobra em atividades paralelas exclusivas de **Atualizar Predicado** e **Manter Predicado Original**. Nelas, o usuário tem a oportunidade de escolher um rótulo entre os oferecidos ou manter a ligação sem rótulo. A cada escolha de novo predicado, uma nova tripla é então criada e adicionada à base de dados **Triplas Rotuladas**. Após repetir o processo para cada tripla de M o módulo **Reagrupar base** entra em ação, substituindo em S as triplas semanticamente enriquecidas, gerando com isso o dataset S'.

4.2 Implementação PLAIN

Como prova de conceito, foi implementada uma aplicação Web denominada PLAIN, acrônimo do método *Predicate LAbelINg* apresentado na Seção 4.1. Além de um módulo Web para consulta a um catálogo de vocabulários controlados, a PLAIN também conta com um módulo para extração de relações utilizando NLP.

4.2.1 Módulo de Consulta a Catálogos de Vocabulários Controlados

Para o presente trabalho, a aplicação implementa parcialmente o método. Por se tratar de uma linguagem do tipo *software* livre, foi escolhido o PHP[3] com sua biblioteca EasyRdf[4] para o desenvolvimento dessa aplicação. Em especial essa biblioteca foi usada na implementação das atividades **Buscar Classe em Catálogos de Vocabulários Controlados**, **Buscar Classes Equivalentes** e **Buscar Predicados relacionados**.

A PLAIN permite a realização de uma série de consultas ao SPARQL Endpoint do Catálogo LOV. O protótipo ainda não tem a opção de consultar diretamente uma base de dados de triplas para rotulação, e necessita que x_i e x_j sejam fornecidos pelo usuário, como sendo parte de uma tripla pobremente rotulada.

A representação dessa busca está organizada no Algoritmo 2. Nele buscou-se estruturar a sequência de tarefas de procura por predicados que a implementação realiza ao receber uma classe como entrada. Os predicados relacionados a essa classe são retornados na saída.

Algoritmo 2: Busca por predicados relacionados em Catálogo de Vocabulários

Entrada: Classe
Saída: Predicados Relacionados

1 $C \leftarrow$ `subClassOf(`*Classe*`)`;
2 $C \leftarrow C +$ `superClassOf(`*Classe*`)`;
3 $C \leftarrow C +$ `ancestorOf(`*Classe*`)`;
4 $C \leftarrow C +$ `descendantOf(`*Classe*`)`;
5 $C \leftarrow C +$ `equivalentClass(`*Classe*`)`;
6 **enquanto** *existir* $c \in C$ **faça**
7 | $PredRel \leftarrow$ `domain`$^{-1}$`(`c`)`;
8 | $PredRel \leftarrow PredRel +$ `range`$^{-1}$`(`c`)`;
9 **fim**
10 **retorna** $PredRel$

A Figura 22 mostra a interface da aplicação em três *frames*. O *frame* A possui dois campos para inserção das classes para consulta (x_i e x_j). O retorno, nos *frames* B e C, são relações de classes equivalentes (C_{x_i} e C_{x_j}, respectivamente), e na parte de baixo dos

[3] https://www.php.net/
[4] http://www.easyrdf.org/

frames estão organizados os predicados associados a essas classes após a aplicação das funções $domain^{-1}()$ e $range^{-1}()$, sobre as classes dos conjuntos C_{x_i} e C_{x_i}.

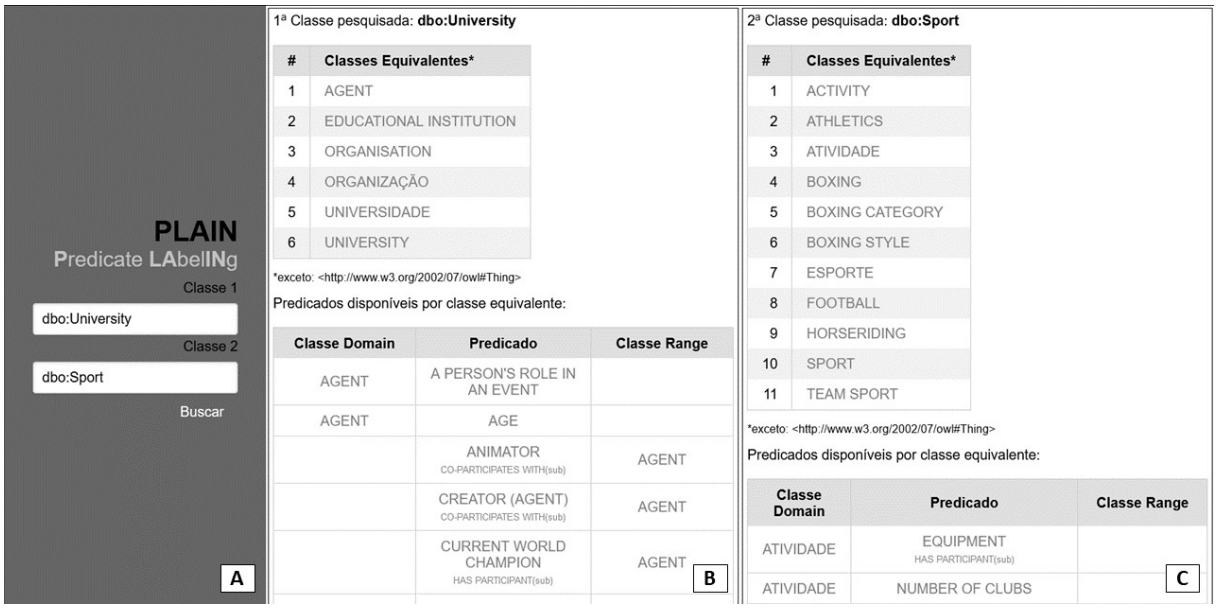


Figura 22 – Interface do Protótipo - PLAIN.

Na implementação da atividade **Ler Base de Dados Conectados**, o usuário da aplicação deve informar a Classe do item que está como sujeito e objeto da tripla a ser enriquecida. Com a ajuda do EasyRdf, as atividades **Buscar Classe em Catálogos** e **Buscar Classes Equivalentes** foram implementadas, permitindo que a aplicação faça consultas ao SPARQL Endpoint do catálogo LOV na busca por classes equivalentes ligadas a elas pelas relações: *owl:equivalentClass*, *rdfs:subClassOf* e *prov:wasDerivedFrom*. De forma a evitar uma generalização indesejada, a classe *owl:Thing* é desconsiderada.

As atividades **Ler Base de Dados Conectados**, **Consultar Classes e Propriedades correlatas**, **Oferecer para o usuário a oportunidade de rotular as ligações** e **Atualizar Predicado** do *Método Predicate Labeling* estão implementadas na PLAIN por meio da apresentação de uma página estruturada em HTML com o resultado do conjunto de consultas realizadas anteriormente, ou seja, C_{x_i} e C_{x_j}. O usuário tem à sua disposição todo o conjunto de classes e predicados para realizar análise e optar pelo predicado que entender ser o mais adequado para rotular a ligação semântica.

A partir do resultado estruturado dessas consultas, é formado o conjunto P_{ij} e o usuário tem a possibilidade de selecionar o predicado que melhor se adequar a cada ligação analisada.

4.2.2 Módulo para Extração de Relações

Já para atender à atividade de extração de relações, foi codificado um *script* em Python que realiza a leitura de um arquivo com um conjunto de sentenças. As sentenças

constantes no arquivo devem ser provenientes de um corpus no mesmo contexto do *dataset* S.

Essa implementação utiliza o OpenNRE para realizar a extração supervisionada de relações em cada sentença e registrar os resultados das extrações em um arquivo estruturado em JSON[5]. O código fonte completo do módulo é apresentado no Apêndice C e as tarefas mais relevantes executadas por ele são detalhadas a seguir.

A lógica do *script* inicia com a leitura da base **Sentenças**, implementada como um arquivo de texto (**entrada.json**) estruturado em JSON:

```
file = open("./plain_re/entrada.json", "r", encoding='UTF8')
```

Em seguida é informado o **Modelo Pré-treinado** com as **Relações Predeterminadas** que serão utilizados na extração:

```
opennre.get_model('wiki80_bert_softmax')
```

Na sequência, em um laço de repetição, é lida a sentença e as posições dos dois recursos (*head* e *tail*) sobre os quais se deseja saber a relação, e a inferência é realizada:

```
tupleModelInfer = model.infer({'text': texto,
        'h': {'pos': (h1, h2)},
        't': {'pos': (t1, t2)}})
```

O retorno então é incrementado em uma variável, convertido e registrado em L, implementado como um arquivo de texto (**saida.json**) estruturado em JSON:

```
with open("./plain_re/saida.json", 'w') as outfile:
     json.dump(sorted_sugestoes, outfile)
```

Uma atualização futura desse módulo deve incluir a captura das sentenças de forma automatizada, realizando uma busca mais abrangente e com diversos *head* e *tail* pertinentes. Dessa forma será possível construir o arquivo de entrada que contenha mais sentenças e, consequentemente, a saída trará um conjunto mais completo de sugestões.

4.2.3 Atualização da base de dados conectados

Finalmente, o fluxo proposto para atualização da base de dados conectados pode ser observado no Algoritmo 3. Caso o usuário opte por atualizar esse predicado na base de

[5] https://www.json.org/

dados, essa atualização deve passar pela inserção da nova ligação, seguida pela exclusão da relação pobre semanticamente, conforme recomendação da W3C SPARQL 1.1 *Update*[6].

Algoritmo 3: Atualiza predicado na base de dados conectados

Entrada: Tripla para atualização
Saída: Tripla com predicado enriquecido

1 `Apaga` $Tripla$ de S' onde:
2 $(Tripla = TriplaParaAtualizacao)$;
3 `Insere` $TriplaEnriquecida$ em S';

4.3 Considerações Finais

Nesta seção foi apresentado o método proposto, intitulado *Predicate Labeling*, desenvolvido com o objetivo de rotular ligações semânticas na Web de Dados. Foram apresentadas todas as definições em que o método se baseia. Além disso, foi exposta a implementação do método, denominada PLAIN, composta por dois módulos. O primeiro módulo é uma aplicação Web com a qual é possível, a partir de um par de classes oferecido, realizar uma série de consultas ao LOV para obter sugestões de predicados mapeados no catálogo. Já o segundo módulo é um *script* que realiza a atividade de RE e retorna sugestões de relações entre um par de entidades a partir de um conjunto de sentenças oferecidas em que essas entidades estejam presentes. Esses dois módulos implementam as principais funcionalidades do método proposto.

Fazendo uso dos dois módulos desenvolvidos de forma complementar, é possível explorar a riqueza de mapeamentos disponíveis nos vocabulários e ontologias. Esse resultado pode ser combinado com a exploração de textos que trazem relações que podem ser extraídas com tarefas de RE. Essa combinação tem o potencial de favorecer na rotulação das relações semânticas.

Apesar do método tratar as atividades de forma integrada, os dois módulos implementados ainda trabalham de forma independente e não contam com a leitura e gravação direta em *datasets* de triplas. Outra evolução necessária é com relação ao acesso ao catálogo de vocabulários controlados, cujas consultas estão atreladas ao LOV.

Na concepção e implementação do método, uma dificuldade significativa encontrada foi acerca da integração dos módulos. Sendo que as abordagens necessárias para lidar com o módulo de exploração de catálogos estão voltados para a Web de Dados, já os procedimentos relativos a RE convergem para o tratamento de dados não estruturados. Integrar essas duas abordagens ainda é uma melhoria prevista em trabalhos futuros.

[6] https://www.w3.org/TR/2013/REC-sparql11-update-20130321/

5 ESTUDOS DE CASO

De forma a validar o método proposto, foram realizados três estudos de caso utilizando os módulos da PLAIN. Inicialmente, um estudo de caso introdutório foi realizado, com um par de *datasets* de pequeno porte, de modo a verificar a funcionalidade da ferramenta gerada e a coerência dos primeiros resultados. Um segundo estudo de caso foi realizado em um *dataset* com dados reais da área de geociências. Por fim, um terceiro estudo de caso procurou verificar o impacto quantitativo da abordagem. Desta vez partiu-se da identificação de pares de recursos associados através de ligações com semântica pobre na DBpedia.

Vale ressaltar que os estudos de caso 2 e 3 foram realizados no escopo de um único *dataset*, devido à dificuldade de encontrar *datasets* já relacionados e dentro de um domínio conhecido. Assim, no caso do estudo de caso 2 buscou-se pela rotulação da relação encontrada entre químicos detectados na água extraída em poços de captação de águas subterrâneas. Já no estudo de caso 3 buscou-se a rotulação da relação entre fármaco e doença que atualmente está materializada com o predicado *rdfs:seeAlso*, de semântica vaga e muito utilizado para realizar, inclusive, ligações com outros *datasets*.

5.1 Estudo de caso introdutório

Para realização do estudo de caso introdutório, foram utilizadas regras geradas como resultados de estudos de caso de (OLIVEIRA et al., 2019). No referido trabalho, foi realizada a mineração de regras de associação de multirrelação em um *dataset* de origem chamado de DtIME (S), que continha informações sobre professores e orientandos da Instituição de Ensino IME. O *dataset* externo utilizado para ampliação foi o DtEsportes (T), que possuía informações sobre a preferência de prática de esportes dos alunos de instituições de ensino.

Os autores realizaram estudos de caso que tiveram como resultados regras de associação entre pares de recursos $(s, t) \in S \times T$. No entanto, a semântica das relações entre os recursos de cada par, não são claras. O objetivo do presente estudo de caso foi enriquecer semanticamente essas relações. Por exemplo, a regra de associação *Plays(Vôlei_de_Praia)* $\rightarrow$ *Supervised_By(Work_On(IME))* informa que alunos que jogam vôlei de praia são supervisionados por alguém que trabalha na instituição de ensino IME. O que indica que há uma possível relação entre os recursos IME e Vôlei de Praia. Mas qual seria a semântica dessa relação?

O estudo de caso realizado no trabalho supra citado retornou um conjunto de regras como saída. Essas saídas foram entradas para o estudo de caso do presente trabalho. Como

os *datasets* S e T são pobres de metadados, para o estudo de caso deste trabalho as classes dos recursos utilizados foram consideradas como definidas pelo usuário. Foi utilizada a regra apresentada no Quadro 3.

Quadro 3 – Regra selecionada a partir do conjunto de regras geradas em (OLIVEIRA et al., 2019).

Regra utilizada
Plays(Vôlei_de_Praia) $\rightarrow$ *Supervised_By(Work_On(IME))*

A partir da regra utilizada, pode-se observar que existe uma relação entre Vôlei de Praia e a instituição de ensino IME. Aplicando a PLAIN para realização da análise, podemos explorar a classe Esporte, de Vôlei de Praia (proveniente de DtEsportes) e a classe Universidade, do IME (proveniente de DtIME). Para realização deste estudo de caso, foram utilizados os prefixos dbo[1] e dul[2].

O passo seguinte teve como resultado todas as classes equivalentes à classe *dbo:Sport*. A Consulta ao LOV retornou um conjunto de onze classes, conforme observado no *frame* C da Figura 22. Uma delas é a *dbo:Activity*, cujo *label* é Atividade. Entre os predicados mapeados como *Domain* e *Range* dessa classe está o *dbo:equipment* que tem mapeada a sub-propriedade *dul:hasParticipant*, cujo *label* é *has participant*.

Já do ponto de vista da Universidade, classe *dbo:University*, a consulta retorna um conjunto de seis classes (*frame* B da Figura 22). Entre as mapeadas está a classe *dbo:Agent*. Fica explícito que o predicado *dbo:currentWorldChampion* tem essa classe como *Range*. O conjunto de consultas também retorna a informação de que essa é uma sub-propriedade de *has participant*. A sequência de busca do processo é apresentada na Figura 23.

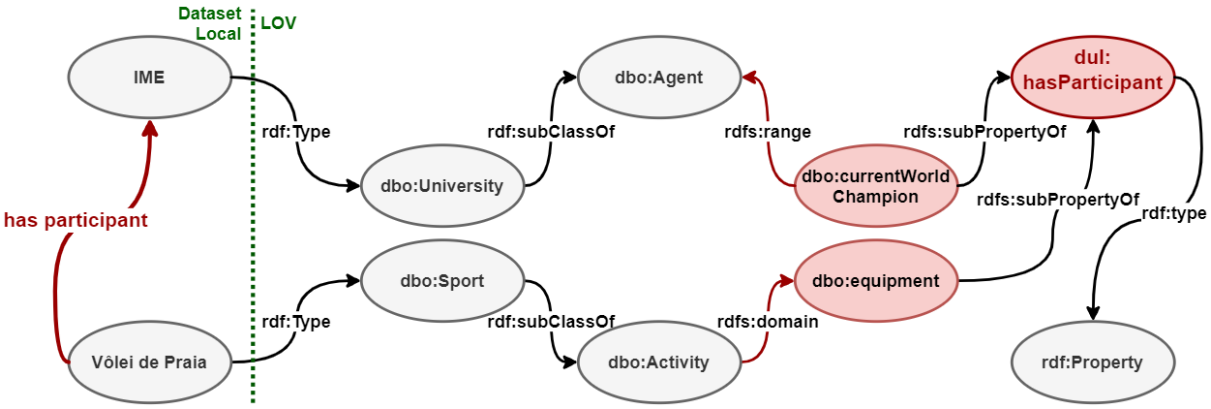


Figura 23 – Resultado da sugestão de predicado entre IME e Vôlei de Praia com uso da PLAIN.

Como resultado desse levantamento feito com o uso da PLAIN, é possível sugerir uma ligação semântica entre Vôlei de Praia e IME utilizando o predicado *dul:hasParticipant*.

[1] http://dbpedia.org/ontology/
[2] http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#

Além disso, o usuário pode navegar até o local de origem do predicado e explorar mais o seu significado e suas restrições de uso formalizadas.

Neste caso, é possível observar no *rdfs:comment* que o predicado *dul:hasParticipant* foi concebido para fazer a ligação entre um objeto e um processo. Dessa maneira, o usuário pode tomar a decisão de aceitar essa sugestão, entendendo que a instituição pode participar de algum processo relacionado com um esporte por um período de tempo, e assim sugerir a tripla apresentada no Quadro 4.

Quadro 4 – Triplas sugeridas com a utilização da PLAIN.

Módulo Utilizado	**Tripla Sugerida**
Cons. Cat. Voc. Controlados	*< :Volei_de_Praia dul:hasParticipant :IME >*
Extração de Relações	*< :Volei_de_Praia :partOf :IME >*

Outra iniciativa possível seria a de criar uma nova propriedade, como uma sub-propriedade ou propriedade equivalente a *dul:hasParticipant*, incluindo informações mais aderentes ao caso explorado.

Para complementar o estudo de caso, foi realizada a Extração Supervisionada de Relações utilizando o **Módulo para Extração de Relações** da PLAIN. A frase *"sport related to university"* foi dada como entrada na ferramenta que retornou a sugestão de nome *"part of"* para a relação entre *sport* e *university*, conforme complementado no Quadro 4.

Além do estudo de caso realizado a partir da regra do Quadro 3, foram realizados outros estudos de caso com a PLAIN a partir de regras geradas em (OLIVEIRA et al., 2019) e (OLIVEIRA et al., 2017). A Tabela 2 apresenta as regras selecionadas nesses trabalhos, o sujeito da tripla, o predicado encontrado a partir da navegação no catálogo de vocabulários (conforme o módulo **Consultar Classes e Propriedades correlatas utilizando recursos semânticos**), a sugestão de predicado e sua probabilidade pela aplicação do RE (conforme o módulo **Identificar potenciais rótulos utilizando RE**) e o objeto da tripla. Por exemplo, com base na regra 3 da Tabela 2, o IME está relacionado a Futebol. Para melhorar a semântica dessa relação, o **Módulo de Consulta a Catálogos de Vocabulários Controlados** sugeriu o predicado *dul:isParticipantIn*, formando a tripla *:IME dul:isParticipantIn :Futebol* . Já o **Módulo para Extração de Relações** sugeriu o predicado *field of work*, formando a tripla *< :IME :fieldOfWork :Futebol >*.

Em todos os estudos de caso é possível observar que o retorno dado pela RE é complementar àquele obtido com a navegação pelos vocabulários, demonstrando que a combinação das duas abordagens é mais rica. No exemplo da regra 3, as duas relações sugeridas são bem distintas e possíveis para os recursos envolvidos, i.e., IME e Futebol. Ou seja, tanto o IME pode participar de alguma forma do esporte Futebol, quanto o IME pode ter Futebol como área de trabalho.

Tabela 2 – Sugestões de rótulos utilizando o método Predicate Labeling.

	Regra Original	**Sujeito**	**Mód. Rec. Sem.**	**Mód. RE**	**Objeto**
1	*Expedition(Researcher) → Expedition(Researcher)*	*Researcher*	*rel:worksWith*	*said to be the same as*	*Researcher*
2	*Plays (Natacao) → Live_In (RJ)*	*Natacao*	*dul: hasParticipant*	*said to be the same as*	*RJ*
3	*Supervised_By(Maria_Claudia), Study_In(IME) → Plays (Futebol)*	*IME*	*dul: isParticipantIn*	*field of work*	*Futebol*
4	*Supervised_By (Cooperator (Work_On (Patronage(MIT)))) → Study_In (IUT)*	*MIT*	*dul: coparticipates With*	*said to be the same as*	*IUT*
5	*dbo:team(dbr:Brazil_national _under _20 _football_team) → Study_In (IUT)*	*dbr:Brazil_ nat_under_ 20_ft*	*frbr:realizer*	*main subject*	*IUT*

5.2 Estudo de caso com um sistema geocientífico

A fim de implementar o modelo proposto em um ambiente de maior complexidade, foi utilizado como base para estudo de caso um sistema geocientífico denominado SIAGAS[3] (Sistema de Informações de Águas Subterrâneas). Este sistema pertence à área fim da empresa pública CPRM[4] (Serviço Geológico do Brasil), que tem como missão gerar e disseminar conhecimento geocientífico com excelência, contribuindo para melhoria da qualidade de vida e desenvolvimento sustentável do Brasil.

5.2.1 Sistema de Informações de Águas Subterrâneas

O SIAGAS é um sistema desenvolvido pela CPRM composto por uma base de dados de poços de captação de águas subterrâneas. Um mapa gerado pelo sistema com a distribuição desses poços é apresentado na Figura 24. No mapa, cada ponto em azul representa um poço de captação de água subterrânea cadastrado até o mês de dezembro de 2020.

O sistema possui módulos capazes de realizar consultas, pesquisa, extração de dados, além de gerar relatórios. Ele se baseia no mapeamento e pesquisa hidrogeológica de poços no Brasil e contém informações sobre dados gerais e construtivos dos poços cadastrados, dados sobre aquíferos, dados geológicos, análise química das águas e dos solos, resultados de testes de bombeamento dos poços, dentre outras.

[3] http://siagas.cprm.gov.br/
[4] http://www.cprm.gov.br/

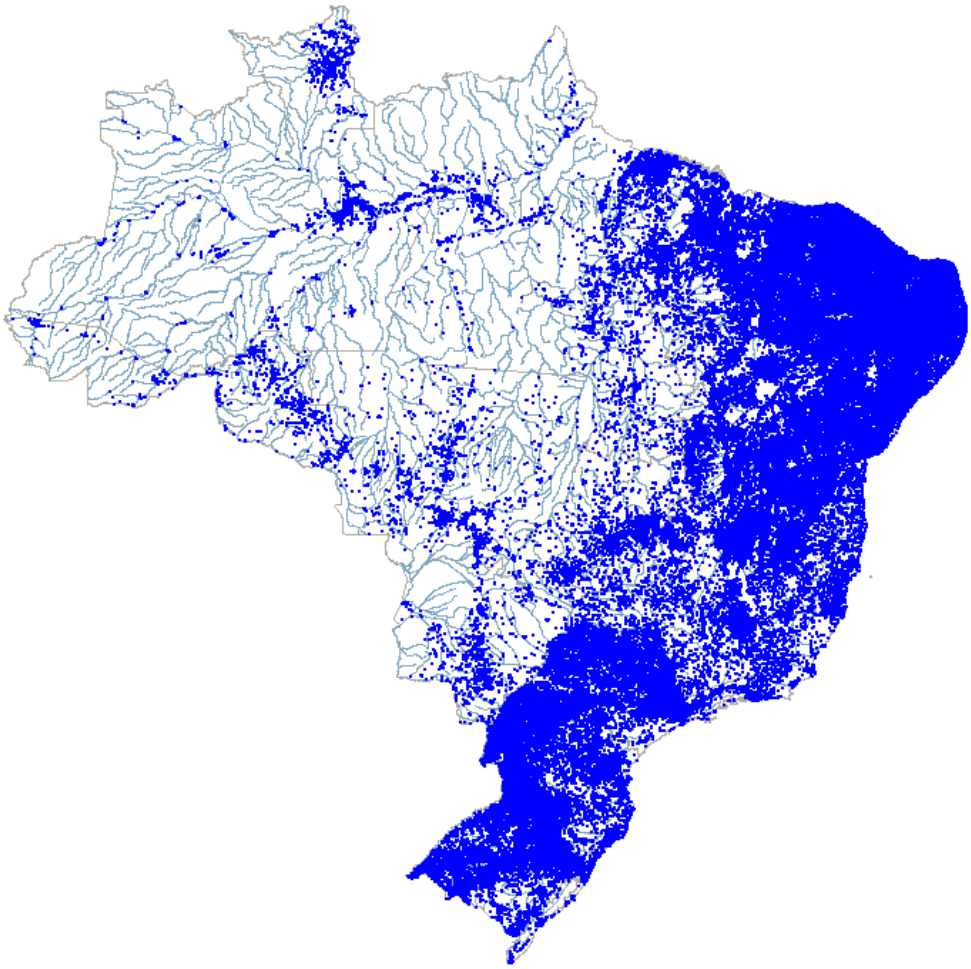


Figura 24 – Mapa da distribuição dos 333.421 poços cadastrados no SIAGAS.

5.2.2 Campos selecionados

Foram mapeadas as necessidades e estipuladas as etapas que serão seguidas para realização do estudo de caso. Em um primeiro momento foram selecionados campos com informações relevantes a respeito da presença de químicos entre os mais de 300 cadastrados. Na Tabela 3 é apresentada a relação dos campos que foram selecionados.

Em seguida, foi realizado um agrupamento pelas características, levando em consideração a localização e os elementos químicos identificados na amostra mais recente de cada poço. Assim foram descartadas as informações de amostras anteriores de um mesmo poço e informações de poços em que não foram realizadas amostras químicas da água.

5.2.3 Triplificação da base de dados

A etapa seguinte foi a triplificação dos dados selecionados. Para realizar essa tarefa foi utilizada a ferramenta Pentaho Data Integration[5], também conhecida como Kettle,

[5] https://www.hitachivantara.com/en-us/products/data-management-analytics/pentaho-data-integration.html

Tabela 3 – Campos selecionados para estudo de caso a partir do banco de dados do SIAGAS.

Campo Selecionado	**Descrição**
Idt_Poco	Identificador do ponto
Quimico	Nome do elemento químico
Cas_Number	Número do CAS (*Chemical Abstracts Service*)
Concentracao	Valor da concentração
Unidade	Unidade de concentração que está sendo usada
Município	Município a que o poço pertence
UF	Estado a que o poço pertence
Bacia_ANA	Nome da Bacia hidrografica
Bacia_Estadual	Nome da Bacia estadual

um artefato para *workflow* ETL da plataforma de Inteligência de Negócio (BI - *Business Intelligence*) Pentaho. O motivo da escolha do Kettle deveu-se às suas características *open source* e por possuir uma ampla comunidade de usuários, além de possibilitar extensões em suas funcionalidades.

No Kettle foi adicionada a extensão denominada ETL4LOD (SILVA, 2018), que executa atividades utilizando tecnologias relacionadas a Dados Conectados e, entre outros recursos, possibilita a geração de triplas RDF. Foram utilizados os seguintes componentes dessa extensão:

- ***Object Property Mapping*** - responsável pela transformação dos dados de entrada em triplas RDF;
- ***NTriple Generator*** - gera triplas RDF no formato N-Triples[6].

Na Figura 25 é apresentado o diagrama completo da Transformação ETL utilizado para triplificar dados do SIAGAS. O fluxo de trabalho parte da leitura dos dados planilhados que foram selecionados do banco de dados relacional do sistema. Em cada um dos três caminhos seguidos a partir dessa leitura, é realizado um recorte apenas das de interesse para cada caminho e ordenados para possibilitar o agrupamento. Na sequência de tarefas, é realizada a criação de URIs para os dados e a materialização das triplas. Após a realização dos três caminhos do processo, é gerado um arquivo com os dados em formato NT. Uma amostra desse arquivo pode ser encontrada no Apêndice A.

O modelo em grafo dos dados selecionados do SIAGAS, resultante do processo de triplificação que foi detalhado, é apresentado na Figura 26. Buscou-se por um modelo adequado para representação dos dados trabalhados, bem como para atender à extração de regras a ser realizada na sequência.

[6] https://www.w3.org/2001/sw/RDFCore/ntriples/

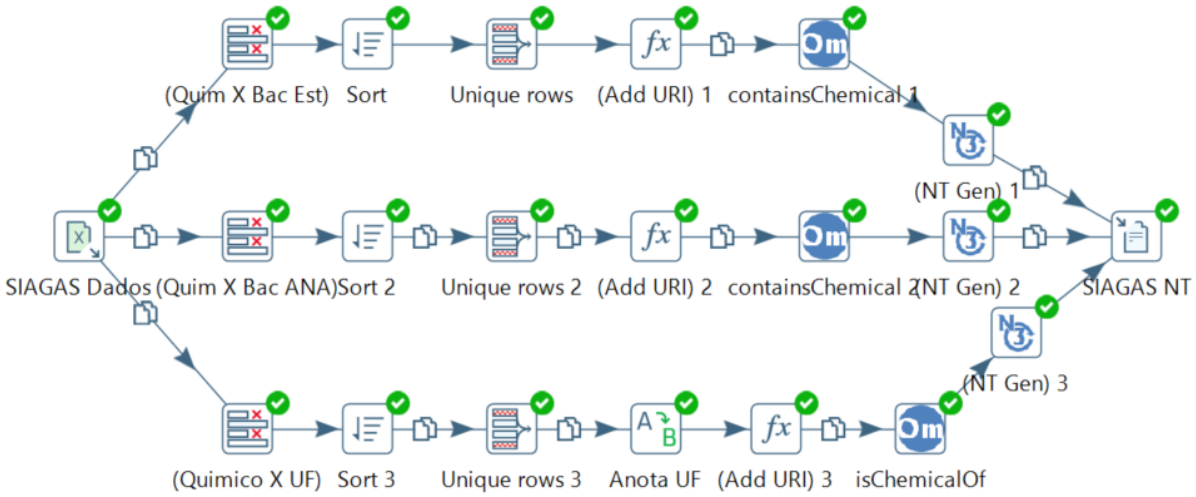


Figura 25 – Diagrama da Transformação ETL utilizada para triplificar dados do SIAGAS.

Observa-se nesse grafo que a classe *sgs:Chemical* está ligada à classe *dbpr:State* por meio do predicado *isChemicalOf*. Essa tripla traz a informação que um dado químico é encontrado em um Estado da Federação (Unidade Federativa - UF).

Outra tripla do grafo traz a ligação da classe *sgs:AnaBasin* com a classe *sgs:Chemical* por meio do predicado *sgs:containsChemical*. Essa tripla traz a informação que uma Bacia ANA contém um determinado químico.

Por último, a terceira tripla descreve a ligação entre as classes *sgs:StateBasin* e *sgs:Chemical*, fazendo uso do predicado *sgs:containsChemical*. Isso informa que uma Bacia Estadual contém um determinado químico.

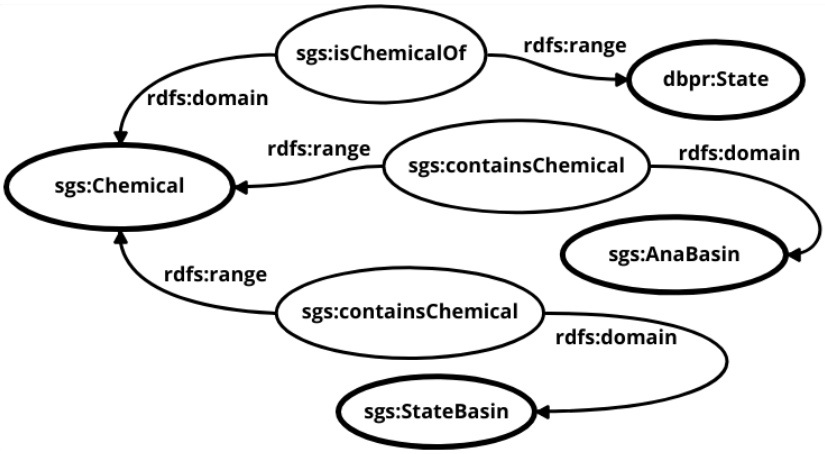


Figura 26 – Modelo em grafo do recorte do SIAGAS.

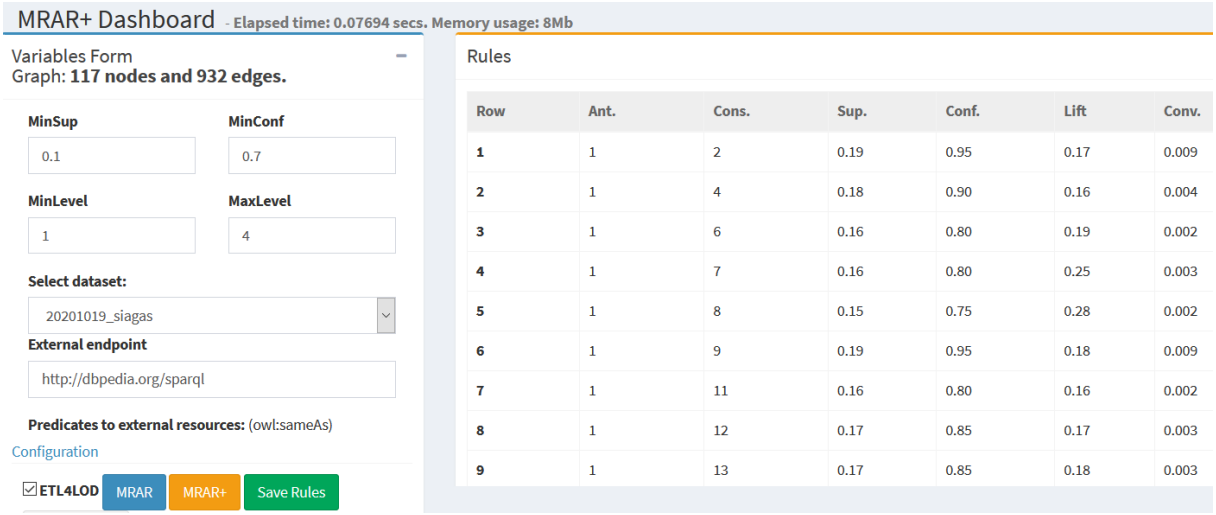
5.2.4 Mineração de regras de associação

Já com os dados triplificados, foi possível utilizá-los como entrada no MRAR+ para obtenção de regras de associação de multirrelação. Como já foi descrito na Seção 3.5,

a ferramenta MRAR+ percorre o grafo e identifica novas associações entre os recursos. No entanto, dependendo da forma como ocorre a triplificação, as regras encontradas podem não ser muito úteis, evidenciando uma deficiência da abordagem da ferramenta MRAR+. Assim, o esquema apresentado na Figura 26 não foi o primeiro a ser obtido. Foi necessário regerar as triplas ajustando-as a diferentes esquemas, até que se chegasse no esquema final.

Dessa forma, o ajuste das triplas *(segundo os novos esquemas)* geradas pelo Kettle, de forma a obter regras de associação no MRAR+, é uma tarefa que exige muitas repetições e ajustes até que se chegue a um resultado adequado. De modo a tornar esse processo mais eficiente, foi adicionada no MRAR+ a possibilidade de leitura direta dos arquivos gerados pelo Kettle. Esse aperfeiçoamento está disponível no GitHub[7].

Utilizando esse novo recurso do MRAR+, a funcionalidade de minerar as regras foi empregada. A tela da aplicação após o processo de mineração de regras pode ser vista na Figura 27. Nela estão presentes os parâmetros utilizados para mineração e parte dos resultados que o sistema retorna sobre cada uma das regras.



Row	Ant.	Cons.	Sup.	Conf.	Lift	Conv.
1	1	2	0.19	0.95	0.17	0.009
2	1	4	0.18	0.90	0.16	0.004
3	1	6	0.16	0.80	0.19	0.002
4	1	7	0.16	0.80	0.25	0.003
5	1	8	0.15	0.75	0.28	0.002
6	1	9	0.19	0.95	0.18	0.009
7	1	11	0.16	0.80	0.16	0.002
8	1	12	0.17	0.85	0.17	0.003
9	1	13	0.17	0.85	0.18	0.003

Figura 27 – Recorte da tela do MRAR+ após mineração de regras de associação de multirrelação do SIAGAS.

Ao trabalhar na base de dados que foi inserida na ferramenta, a mineração teve como resultado um conjunto de 1.216 regras que associaram a presença de um dado químico encontrado à possibilidade da presença de outros químicos na água. Fez-se uso do recurso de interface que permite ordenar as regras, dando destaque para as que contavam com mais suporte. Após a validação de um especialista na área podemos destacar as regras disponíveis no Quadro 5.

No site do Serviço Geológico Americano[8] é possível observar uma relação de elementos químicos que são contaminantes de águas subterrâneas bem como a sua procedência.

[7] https://github.com/rafans222/MRAR_plus
[8] https://on.doi.gov/3s9by6z

Quadro 5 – Regras destacadas extraídas das triplas do SIAGAS.

Regras Destacadas
containsChemical(Aluminum), containsChemical(Lead) $\rightarrow$ *containsChemical(Iron)*
containsChemical(Chromium) $\rightarrow$ *containsChemical(Lead)*
containsChemical(Fluoride), containsChemical(Nitrite) $\rightarrow$ *containsChemical(Nitrate)*

Com o apoio também dessas informações, pode-se observar na primeira regra a relação entre a presença de *alumínio* e *chumbo* com a presença de *ferro*. Sabe-se que o *alumínio* está presente naturalmente em algumas rochas, já *chumbo* e o *ferro* são possivelmente provenientes de atividades industriais.

Já na segunda regra, que associa a presença de *cromo* e a presença de *chumbo*, observa-se que ambos são possivelmente provenientes de atividades de mineração. Finalmente, a terceira regra destaca que a combinação da presença de *fluoreto* e *nitrito* favoreceu a presença também de *nitrato*. O que caracteriza uma possível contaminação por esgoto. Isso porque o *fluoreto* é amplamente utilizado como aditivo ao abastecimento de água municipal e o *nitrito* bem como o *nitrato* são possivelmente provenientes de resíduos humanos. Outra hipótese para justificar a presença de *nitrito* e *nitrato* é a contaminação da água subterrânea por uso de fertilizantes.

5.2.5 Aplicação do método *Predicate Labeling*

Ao se observar as regras geradas, é constatada a relação existente entre a causa da presença de químicos. Então utilizou-se a PLAIN com o objetivo de rotular essa ligação entre químicos. Também fez-se uso do módulo de extração de relações para examinar a possibilidade de encontrar outros rótulos interessantes.

5.2.5.1 Exploração de Recursos Semânticos

Representando uma classe adequada para um químico, foi fornecido como entrada para ferramenta a classe *dul:ChemicalObject*. Esta então retornou um conjunto de seis classes equivalentes e uma coleção de predicados que possuem essas classes como *rdfs:domain* e *rdfs:range*. Na Figura 28 é apresentada a tela da aplicação após a realização da busca. A interface está composta no primeiro *frame* pelas classes x_i e x_j. Neste caso, ambas são *dul:ChemicalObject*. Já no Quadro 6 é destacada a relação das classes C_{x_j} equivalentes retornada pela PLAIN.

Com o uso da ferramenta, observa-se, por exemplo, que a classe pesquisada *dul:ChemicalObject* é subclasse da sequência de superclasses: *dul:ChemicalObject*, *dul:PhysicalBody*, *dul:PhysicalObject*, *dul:Object* e *dul:Entity*. Ainda com a ajuda da ferramenta, pode-se observar que a propriedade *dul:follows* tem a classe *dul:Entity* como *Domain* e como *Range*. Essa sequência fica como apresentado na Figura 29. Nesse diagrama é destacado o fluxo

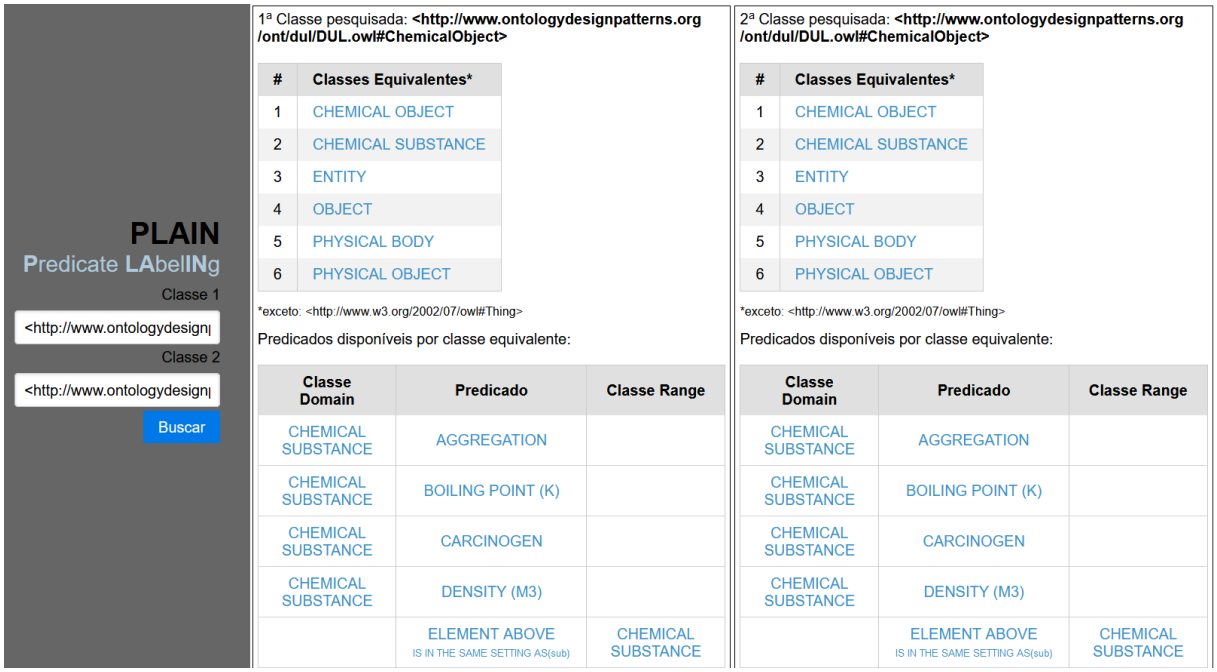


Figura 28 – Tela da PLAIN com o resultado da busca pelos predicados relacionados à classe *dul:ChemicalObject*.

Quadro 6 – Classes equivalentes a *dul:ChemicalObject*, pesquisadas com a PLAIN.

Classes Equivalentes a *dul:ChemicalObject*
CHEMICAL OBJECT
CHEMICAL SUBSTANCE
ENTITY
OBJECT
PHYSICAL BODY
PHYSICAL OBJECT

para encontrar um dos rótulos (p_{x_i}) interessantes. O conjunto completo de predicados P_{x_i} encontrados está disponível no Apêndice B.

Os resultados foram apresentados para um especialista no domínio, que reconheceu como opção interessante para rotulação, dentre as opções encontradas pela ferramenta, o predicado follows (*dul:follows*). O especialista esclareceu que o rótulo pode inspirar a determinação de que há afinidade química entre os elementos. Por exemplo, com base nesse predicado poderia ser gerada a tripla: *sgs:Nitrite dul:follows sgs:Nitrate*.

5.2.5.2 Utilização do Módulo de Extração de Relações

Seguindo a proposta do método *Predicate Labeling*, procurou-se por sentenças que contivessem os pares de entidades levantados na etapa de busca no catálogo de vocabulários para que dessa forma pudesse viabilizar a extração de relações utilizando o módulo de extração de relações da PLAIN.

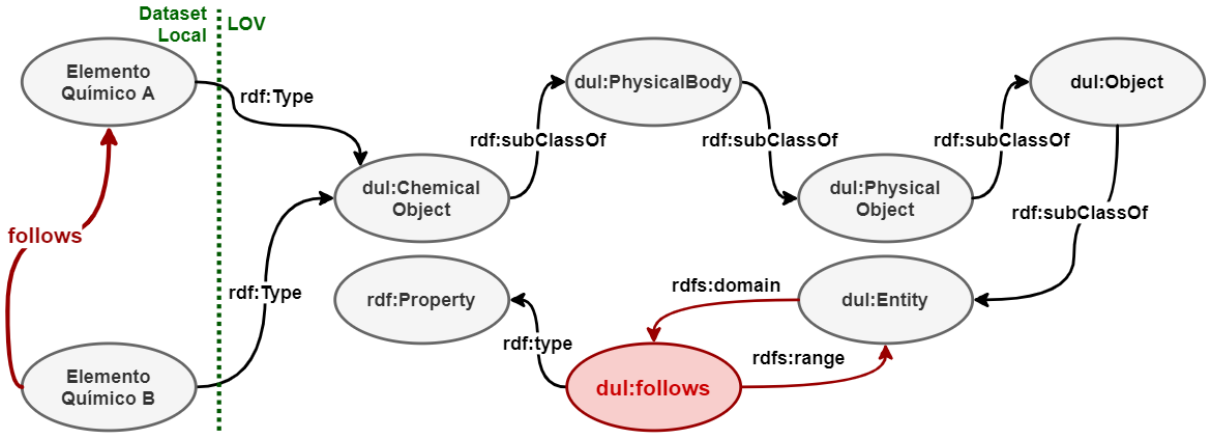


Figura 29 – Resultado da sugestão de predicado entre químicos com o uso da PLAIN.

Para construir essas sentenças de entrada foi utilizado o PubMed[9], que é um repositório que possui um motor de busca disponível na Web, desenvolvido e mantido pela Biblioteca Nacional de Medicina dos Estados Unidos[10]. Ele conta com uma base de citações e resumos de artigos na área de biomedicina. Até novembro de 2020, esse corpus contava com mais de 30 milhões de citações de literatura biomédica da MEDLINE[11], periódicos de ciências biológicas e livros online. A relação completa das orações selecionadas com o posicionamento das entidades está disponível no Apêndice D.

Um exemplo de sentença de entrada tendo como fonte o PubMed é:

{"text": "This kind of metals such as Chromium and Lead could affect health and the ecosystem.", "h": {"pos": [28, 36]}, "t": {"pos": [41, 45]}}

Essa entrada conta com o texto extraído e as posições de início e fim dos dois termos, chamados de *head* e *tail*, entre os quais se deseja extrair a relação. Neste exemplo, o *head* é o termo *Chromium* e o *tail* é o termo *Lead*.

Após a sequência de extrações para cada sentença, a ferramenta retorna como saída para o usuário um conjunto de sugestões de rótulos para o predicado. O conjunto de sugestões encontradas por esse módulo está sumarizado no Quadro 7.

Quadro 7 – Predicados entre químicos encontrados com o uso do módulo de Extração de Relações da PLAIN.

Predicados encontrados
followed by
has part
instance of
follows

[9] https://pubmed.ncbi.nlm.nih.gov/
[10] https://www.ncbi.nlm.nih.gov/
[11] https://www.nlm.nih.gov/bsd/medline.html

Observa-se que os predicados encontrados com o módulo de extração de relações são complementares aos que foram encontrados no módulo que utiliza recursos semânticos em catálogos de vocabulários e esse aspecto favorece a tarefa de rotular adequadamente as ligações.

Esse novo conjunto de resultados também foi apresentado para um especialista no domínio, que reconheceu como opção interessante para rotulação, dentre as opções encontradas pelo módulo de extração de relações, o predicado follows. Neste momento é interessante ressaltar a intercessão que ocorreu ao se encontrar a mesma sugestão de rótulo obtida pela navegação nos vocabulários controlados, além de esse predicado ser o selecionado pelo especialista. Isso pode sugerir que, quando ocorre a simultaneidade de sugestões entre os dois módulos, esse rótulo sugerido pode ser reforçado como uma possível adoção pelo usuário.

5.2.5.3 Conclusão sobre o estudo de caso com o SIAGAS

O SIAGAS é um sistema com informações de poços de captação de águas subterrâneas em todo Brasil. Para o presente estudo de caso foi realizada a escolha por campos relevantes com informações sobre a análise química da água. Após a sequência de tarefas que favoreceram a análise dessas informações em forma de Dados Conectados, foram destacadas três regras de associação interessantes entre químicos.

Sendo observada essa correlação entre químicos, foi possível aplicar o método proposto neste trabalho, por meio da utilização da implementação PLAIN. Assim, verificou-se que, com a aplicação do método, foram obtidas uma série de sugestões que possibilitam rotular a relação entre químicos, até então pobres semanticamente. Essa rotulação pode ser feita tanto pela investigação em um catálogo de vocabulários controlados quanto por meio de técnicas de NLP.

Ao fazer uso de um vocabulário controlado, pôde-se observar que a ligação entre dois químicos poderia ser adequadamente descrita, por exemplo, com predicados como *co-participates with*, *near to* ou *follows*. Com o auxílio de um especialista foi observado que essa última sugestão é uma opção válida porque remete à afinidade química. Já com a complementação do RE, apesar de também encontrar a relação *follows*, foram obtidos poucos predicados e que podem ser considerados inadequados para o contexto. O especialista destacou que o termo *follows*[12], em inglês, dá a conotação de que tais químicos são comumente encontrados juntos.

Essa inadequação se deve à limitação das relações disponíveis na base de **Relações Pré-determinadas**. Para o estudo de caso, foram oferecidas oitenta relações extraídas da Wikidata[13], sendo esse um conjunto não muito abrangente. Uma forma de melhorar

[12] https://dictionary.cambridge.org/
[13] https://www.wikidata.org/wiki/Wikidata:Main_Page

esse retorno dado pela PLN seria preparar essa base com relações dentro do contexto do *dataset* que se deseja enriquecer. Essa preparação contaria com o treinamento do modelo englobando sentenças como:

"However, the interaction *of iron with nitrite or nitrate present in the sludge has received little attention."*

Se sentenças como essa estivessem disponíveis para a construção da base de **Relações Pré-determinadas**, poderia-se ter o predicado interacts with como uma das sugestões do processo de RE. Isso favoreceria o enriquecimento da oferta de rótulos disponíveis para o usuário da aplicação.

5.3 Estudo de caso com um *dataset* interdisciplinar

Com o propósito de validar a utilização do método de forma mais ampla, buscou-se a estruturação de outro estudo de caso com um *dataset* proeminente da Web de Dados. A ideia desse estudo de caso era encontrar *datasets* onde houvesse relações semânticas pobres que poderiam ser substituídas por outras de semântica mais rica. Assim o estudo de caso foi realizado na DBpedia (LEHMANN et al., 2015), devido à sua relevância, grande volume de dados e por tratar de assuntos interdisciplinares.

5.3.1 Reconhecimento de Relações Semanticamente Vagas

Conforme apresentado em (SCHMACHTENBERG; BIZER; PAULHEIM, 2014), é frequente o uso da propriedade *rdfs:seeAlso* na formação de triplas. Realizando uma consulta de forma a verificar o contexto de utilização desse predicado, observa-se que ele é empregado para realizar vínculos entre recursos de forma genérica, assim, pode-se afirmar que haveria um incremento semântico com sua substituição por predicados mais adequados.

Um exemplo dessa utilização é apresentado no Quadro 8, que retrata uma tripla extraída da DBpedia, utilizando também prefixo *dbr*[14], cujos recursos estão relacionados por meio do predicado *rdfs:seeAlso*. É possível verificar que a relação poderia estar melhor expressa com a utilização de um predicado com semântica mais clara, já que está relacionando uma vacina a uma doença para qual a vacina foi desenvolvida.

Quadro 8 – Tripla da DBpedia com o predicado *rdfs:seeAlso*.

Tripla extraída da DBpedia
< dbr:Pertussis_vaccine rdfs:seeAlso dbr:Pertussis >

[14] http://dbpedia.org/resource/

Para chegar a essa relação e encontrar outras relações também com semântica pobre nesse *dataset*, foi estruturada a Consulta 4 para execução no SPARQL Endpoint da DBpedia[15]. O resultado da consulta totalizou 43 triplas, todas relacionando um medicamento a uma doença por meio do predicado *rdfs:seeAlso*. Um ajuste na consulta, buscando relacionar uma doença a um medicamento, com o mesmo predicado de semântica vaga, retornou um total de 14 triplas, totalizando assim 57 triplas entre essas duas classes.

Consulta 4: Consulta triplas formadas entre remédios e doenças com semântica pobre.

Entrada: Predicados com semântica pobre
Saída: Triplas relacionadas com semântica pobre

```
1 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
3 PREFIX dbo: <http://dbpedia.org/ontology/>
4 SELECT DISTINCT ?Drug ?Predicate ?Disease WHERE {
5         ?Drug ?Predicate ?Disease .
6         ?Drug rdf:type dbo:Drug .
7         ?Disease rdf:type dbo:Disease .
8         FILTER (REGEX (STR (?Predicate), "seeAlso"))
9 }
```

A Consulta 4 foi executada também sem a restrição dos predicados, para que fosse possível observar o total de relações entre medicamentos e doenças com qualquer predicado. Para isso fori retirada a linha 8 da Consulta 4. Foi retornado um total de 106 triplas que estão relacionadas no Apêndice E. Ajustando a consulta para relacionar doenças e medicamentos, o total retornado foi de 1015 triplas. Um gráfico que ilustra essa proporção é apresentado na Figura 30.

5.3.2 Levantamento dos Recursos Semânticos disponíveis

Dada a representatividade do conjunto de relações com semântica pobre mapeado entre instâncias, tem-se a constatação de que pode-se utilizar o módulo de exploração de recursos semânticos da PLAIN, tendo como entrada as classes *dbo:Drug* e *dbo:Disease*.

Na Figura 31 é apresentada a interface do módulo com o resultado estruturado das consultas realizadas ao catálodo do LOV. Observa-se o retorno com possibilidade de utilização de 25 predicados, a partir da investigação da classe *dbo:Drug* e 11 predicados, a partir da investigação da classe *dbo:Disease*. Diante do resultado apresentado, observa-se que os predicados disponíveis no catálogo não favorecem uma rotulação adequada entre as duas classes de recursos.

[15] https://dbpedia.org/sparql

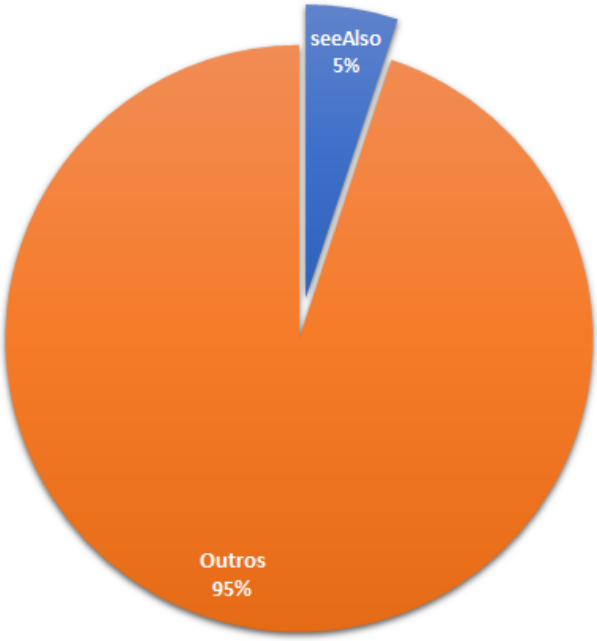


Figura 30 – Representatividade do predicado *rdfs:seeAlso* entre remédios e doenças na DBpedia.

5.3.3 Extração de Relações entre os Recursos

Em continuidade ao fluxo do método, buscou-se no PubMed por sentenças que tivessem em sua composição termos dentro do contexto de medicamentos e doenças, seguindo os resultados das consultas realizadas no *dataset* da DBpedia. A lista completa das sentenças selecionadas está disponível no Apêndice F. Um exemplo desse conjunto de sentenças de entrada é:

{"text": "Drugs are often seen as ancillary to the purpose of fighting diseases.", "h": {"pos": [0, 5]}, "t": {"pos": [61, 69]}}

Tendo como entrada esse conjunto de sentenças, o módulo de Extração de Relações da PLAIN retorna como sugestão os predicados relacionados no Quadro 9.

Quadro 9 – Predicados entre doenças e medicamentos encontrados com o uso do módulo de Extração de Relações da PLAIN.

Predicados encontrados
main subject
instance of
part of

5.3.4 Conclusão sobre o estudo de caso com o *dataset* interdisciplinar

Nesse estudo de caso com o *dataset* da DBpedia foi possível observar uma lacuna semântica em parte representativa das triplas que relacionam medicamentos a doenças

PLAIN
Predicate
LAbelINg

Classe 1
dbo:Drug

Classe 2
dbo:Disease

Buscar

1ª Classe pesquisada: **dbo:Drug**

#	Classes Equivalentes*
1	CHEMICAL SUBSTANCE
2	DRUG

*exceto: <http://www.w3.org/2002/07/owl#Thing>

Predicados disponíveis por classe equivalente:

Classe Domain	Predicado	Classe Range
CHEMICAL SUBSTANCE	AGGREGATION	
CHEMICAL SUBSTANCE	BOILING POINT (K)	
CHEMICAL SUBSTANCE	CARCINOGEN	
CHEMICAL SUBSTANCE	DENSITY (M3)	
	ELEMENT ABOVE IS IN THE SAME SETTING AS(sub)	CHEMICAL SUBSTANCE
CHEMICAL SUBSTANCE	FLASH POINT	
CHEMICAL SUBSTANCE	FORMULA	
CHEMICAL SUBSTANCE	LETHAL WHEN GIVEN TO CHICKENS	
CHEMICAL SUBSTANCE	LETHAL WHEN GIVEN TO MICE	
CHEMICAL SUBSTANCE	LETHAL WHEN GIVEN TO RABBITS	
CHEMICAL SUBSTANCE	LETHAL WHEN GIVEN TO RATS	
CHEMICAL SUBSTANCE	MELTING POINT (K)	
CHEMICAL SUBSTANCE	MOLECULAR WEIGHT	
	NOT SOLUBLE IN	CHEMICAL SUBSTANCE
CHEMICAL SUBSTANCE	PUBCHEM	
CHEMICAL SUBSTANCE	SOLUBILITY	
	SOLVENT WITH BAD SOLUBILITY	CHEMICAL SUBSTANCE
	SOLVENT WITH GOOD SOLUBILITY	CHEMICAL SUBSTANCE
	SOLVENT WITH MEDIOCRE SOLUBILITY	CHEMICAL SUBSTANCE
DRUG	BIOAVAILABILITY	
DRUG	BOILING POINT (K)	
DRUG	CHEBI	
DRUG	IUPAC NAME	
DRUG	MELTING POINT (K)	

** **Domain** do Predicado: Classe(s) que pode(m) ser Sujeito.
** **Range** do Predicado: Classe(s) que pode(m) ser Objeto.

2ª Classe pesquisada: **dbo:Disease**

#	Classes Equivalentes*
1	DISEASE

*exceto: <http://www.w3.org/2002/07/owl#Thing>

Predicados disponíveis por classe equivalente:

Classe Domain	Predicado	Classe Range
DISEASE	DISEASESDB	
DISEASE	EMEDICINE SUBJECT	
DISEASE	EMEDICINE TOPIC	
DISEASE	GENEREVIEWSID	
DISEASE	GENEREVIEWSNAME	
DISEASE	ICD1	
DISEASE	ICD10	
DISEASE	ICD9	
DISEASE	ICDO	
DISEASE	MEDLINEPLUS	
DISEASE	MESH ID	

** **Domain** do Predicado: Classe(s) que pode(m) ser Sujeito.
** **Range** do Predicado: Classe(s) que pode(m) ser Objeto.

Figura 31 – Interface da PLAIN com o resultado da consulta às classes *dbo:Drug* e *dbo:Disease*.

por meio do predicado *rdfs:seeAlso*. Conforme pode ser observado em uma consulta mais aberta, o *dataset* conta com outras triplas que representam relações semelhantes e estão com a semântica clara, utilizando predicados como *dbo:medicalCause* ou *dbo:treatment*, cujos exemplos são apresentados no Quadro 10. Tais predicados fazem parte da ontologia da DBpedia, e já estão em uso no *dataset*. Porém, a nova versão da ontologia da DBpedia não foi atualizada no catálogo do LOV, que é atualmente utilizado pela PLAIN. Essa falta de atualização inviabilizou uma sugestão mais adequada que poderia ser trazida por esse módulo da ferramenta PLAIN.

Quadro 10 – Exemplos de relações entre doenças e medicamentos adequadamente rotuladas na DBpedia.

Triplas
< bdr:Spasmodic_dysphonia dbo:treatment dbr:Botulinum_toxin >
< dbr:Pancreatitis dbo:medicalCause dbr:Alcohol_(drug) >

Já o módulo de extração de relações, que contou com a entrada de sentenças no contexto de doenças e medicamentos retirados do PubMed, entregou sugestões muito genéricas. Isso corrobora a necessidade que foi observada nos estudos de caso anteriores, de se treinar o algoritmo do OpenNRE, que extrai as relações com textos no contexto do assunto alvo, de modo a favorecer a disponibilização de sugestões mais pertinentes.

5.4 Considerações Finais

Nesta seção foram apresentados estudos de caso que seguiram a proposta do método *Predicate Labeling*. Eles foram realizados utilizando a implementação do método, denominada PLAIN, que possui um módulo com uma aplicação Web para exploração de um catálogo de vocabulários controlados e um segundo módulo que realiza a tarefa de RE. Os estudos de caso evidenciaram que a utilização do método tem potencial para trazer resultados consistentes no enriquecimento de *datasets* da Web de Dados.

Nos primeiros dois estudos de caso foi possível observar a sugestão de predicados úteis, validando a abordagem proposta. Em especial, o segundo estudo de caso utiliza um *dataset* construído a partir de dados reais do sistema SIAGAS da CPRM, que após ser analisado pelo conjunto de ferramentas MRAR+ e PLAIN, foi possível identificar novas relações entre os dados lá existentes. Já no terceiro estudo de caso, apesar das sugestões inadequadas, foi possível observar que se o catálogo de vocabulários (LOV) estivesse atualizado, seria possível enriquecer 5% das relações entre as classes "doença" e "fármaco", trazendo ganhos semânticos significativos para instâncias dessas classes na DBpedia. Outros pares de classes nesse mesmo *dataset* poderiam ser beneficiados pela abordagem proposta, enriquecendo semanticamente ainda mais a DBpedia.

Vale ressaltar ainda que, embora os dois últimos estudos de caso tenham sido realizados no contexto de um único *dataset*, seria possível identificar tais relações envolvendo mais de um *dataset*. A própria DBpedia relaciona-se com vários outros *datasets* através da relação *rdfs:seeAlso*, com 2,4% do uso desse predicado atualmente destinado à ligações com outros *datasets*, conforme quantificado na Tabela 4.

Tabela 4 – Relações que utilizam o predicado *rdfs:seeAlso* na DBpedia.

Tipo de Ligação	**Quantidade**
Interna	24.8448
Externa	5.899

6 CONCLUSÃO

Motivado pela demanda crescente da população por informação em tempo real, o aumento diário da quantidade de dados gerados pela sociedade é possibilitado por avanços tecnológicos. Nesse contexto, a Web destaca-se por sua amplitude e importância global, uma vez que possibilitou o amplo acesso da sociedade a todos os tipos de informações.

Como uma evolução da Web tradicional, a WS fornece uma estrutura comum que permite que os dados sejam compartilhados e reutilizados. A iniciativa denominada Dados Conectados, é um conjunto de boas práticas para a publicação de dados na Web. Nesse contexto, os Dados Conectados são dados publicados e ligados utilizando as tecnologias e padrões da WS.

Uma forma de aumentar o conhecimento sobre esses Dados Conectados é realizando uma ampliação dos *datasets* da WS. Atualmente, essas relações encontradas por esse processo de enriquecimento apontam para pares de recursos que podem ter alguma relação.

O desenvolvimento desse trabalho foi motivado pela necessidade de se resolver o problema caracterizado pela ausência de semântica observada nos links gerados no processo de enriquecimento de *datasets* da Web de Dados. Atualmente as relações encontradas após esse enriquecimento não são nomeadas. Além disso, conforme apurado em (SCHMACHTENBERG; BIZER; PAULHEIM, 2014), predicados com semântica pobre são amplamente utilizados tanto internamente quanto para realizar ligações entre *datasets* publicados de diferentes áreas de conhecimento.

6.1 Contribuições

Para solucionar esse problema, foi apresentado um método, intitulado como *Predicate Labeling*, que faz uso de ontologias e vocabulários controlados, bem como de técnicas de RE, para favorecer a identificação e rotulação dessas relações.

Como contribuições além do método, foi desenvolvida uma aplicação, com dois módulos, denominada PLAIN, que implementa parcialmente a funcionalidade do método proposto. O módulo de Reconhecimento Semântico é capaz de realizar uma série de consultas ao SPARQL Endpoint do LOV e organizar os resultados para análise pelo usuário. Já no âmbito do NLP, houve a codificação de um módulo que realiza um conjunto de extrações de relações com o objetivo de complementar ao resultado conseguido com a busca em vocabulários controlados. O código fonte está disponível no GitHub[1].

[1] https://github.com/rafans222/plain

Para favorecer a evolução da pesquisa, foi implementada no MRAR+ a possibilidade de receber como entrada um conjunto de triplas geradas com o Kettle. Isso dispensa a necessidade de se adaptar a formatação aumenta a produtividade dessa etapa presumivelmente custosa. O código fonte dessa versão do MRAR+ também está disponível no GitHub[2].

A PLAIN foi aplicada em estudos de caso em que foi observado que a rotulação utilizando como base as ligações disponíveis em um repositório de vocabulários, juntamente com extração de relações, é viável, e se mostra como uma solução consistente para o problema levantado por esta pesquisa.

6.2 Melhorias e Trabalhos Futuros

No decorrer desta pesquisa foi desenvolvida uma versão da implementação do método proposto, que realiza a consulta a um catálogo de vocabulários controlados. Essa versão realiza de forma independente a tarefa que consiste em utilizar NLP para maior enriquecimento dos resultados encontrados. Uma evolução na implementação pode mesclar as duas abordagens, integrando o resultado encontrado por elas. Isso pode ser feito, por exemplo, com a automatização do uso da saída de um módulo como entrada do outro.

Como trabalhos futuros sugere-se a realização de mais estudos de caso, incluindo a consulta a outros repositórios de *datasets* e utilização de *datasets* com triplas de áreas de conhecimento diversas. A PLAIN também pode incluir a funcionalidade de materializar as relações que foram encontradas e deve realizar o levantamento das sugestões de rótulos diretamente com a ajuda do NLP, incorporando esse módulo. A abordagem proposta poderia se beneficiar não só do catálogo de vocabulários/ontologias mas também de outras relações similares já existentes no próprio *dataset* ou em *datasets* do mesmo contexto (apesar de ser computacionalmente mais custoso).

De forma a enriquecer as sugestões de rótulos oferecidas pelo módulo de *Relation Extraction*, sugere-se a ordenação das sugestões de rótulos oferecidas, utilizando, por exemplo, o índice de confiança fornecido pelo modelo utilizado. Sugere-se também o treinamento do modelo utilizado com textos no contexto de cada *dataset* fonte do enriquecimento. Esse modelo otimizado deve favorecer a disponibilização de um conjunto de sugestões de rótulos mais apropriado.

Nesse sentido, um trabalho futuro é a extensão do método proposto de modo a considerar como parte do método o retreinamento da base usada pelo módulo de RE, a partir do *dataset* fonte a ser enriquecido.

Outra extensão possível do método poderia incorporar mais etapas de modo a

[2] https://github.com/rafans222/MRAR_plus

incluir maior interação com o especialista, facilitando a curadoria dos rótulos sugeridos para as novas relações.

Por fim, sugere-se também a continuidade da análise das informações sobre os químicos detectados nas amostras de águas subterrâneas, por meio do cruzamento de informações geolocalizadas sobre atividades de mineração, atividades industriais e tratamento de esgoto, e com isso dar suporte a ações de mitigação de contaminação dessas águas. Uma outra possibilidade de cruzamento dos dados do SIAGAS com bases de dados disponibilizadas na *Comparative Toxicogenomics Database*[3], permitiria a análise de correlações entre os químicos e as doenças causadas, e com isso dar suporte a ações no sentido de prevenção de doenças.

[3] http://ctdbase.org/

REFERÊNCIAS

AHMED, A. F.; SHERIF, M. A.; NGOMO, A.-C. N. Lsvs: Link specification verbalization and summarization. In: SPRINGER. *International Conference on Applications of Natural Language to Information Systems*. [S.l.], 2019. p. 66–78.

ATHANASIOU, S.; GIANNOPOULOS, G.; GRAUX, D.; KARAGIANNAKIS, N.; LEHMANN, J.; NGOMO, A.-C. N.; PATROUMPAS, K.; SHERIF, M. A.; SKOUTAS, D. Big poi data integration with linked data technologies. In: *22nd International Conference on Extending Database Technology (EDBT 2019)*. [S.l.: s.n.], 2019. p. 477–488.

BIZER, C.; VOLZ, J.; KOBILAROV, G.; GAEDKE, M. Silk - a link discovery framework for the web of data. In: *18th International World Wide Web Conference*. [S.l.: s.n.], 2009. v. 122.

CHAN, Y. S.; ROTH, D. Exploiting syntactico-semantic structures for relation extraction. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. [S.l.: s.n.], 2011. p. 551–560.

CHRISTEN, P. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. [S.l.]: Springer Science & Business Media, 2012.

COLLOVINI, S.; GONÇALVES, P. N.; CAVALHEIRO, G.; SANTOS, J.; VIEIRA, R. Relation extraction for competitive intelligence. In: SPRINGER. *International Conference on Computational Processing of the Portuguese Language*. [S.l.], 2020. p. 249–258.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

ELMAGARMID, A. K.; IPEIROTIS, P. G.; VERYKIOS, V. S. Duplicate record detection: A survey. *IEEE Transactions on knowledge and data engineering*, IEEE, v. 19, n. 1, p. 1–16, 2006.

FERRAZ, R. *Tendências da web*. [S.l.]: Senac, 2018.

FU, T.-J.; LI, P.-H.; MA, W.-Y. Graphrel: Modeling text as relational graphs for joint entity and relation extraction. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. [S.l.: s.n.], 2019. p. 1409–1418.

GRUBER, T. R. Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, Elsevier, v. 43, n. 5-6, p. 907–928, 1995.

GUPTA, P.; SCHÜTZE, H.; ANDRASSY, B. Table filling multi-task recurrent neural network for joint entity and relation extraction. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. [S.l.: s.n.], 2016. p. 2537–2547.

HALPIN, H.; HAYES, P. When owl: sameas isn't the same: An analysis of identity links on the semantic web. In: *LDOW*. [S.l.: s.n.], 2010.

HAN, X.; GAO, T.; YAO, Y.; YE, D.; LIU, Z.; SUN, M. Opennre: An open and extensible toolkit for neural relation extraction. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. [S.l.: s.n.], 2019. p. 169–174.

HAN, X.; ZHU, H.; YU, P.; WANG, Z.; YAO, Y.; LIU, Z.; SUN, M. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. *arXiv preprint arXiv:1810.10147*, 2018.

HERRERA, J. E. T.; CASANOVA, M. A.; NUNES, B. P.; LOPES, G. R.; LEME, L. Dbpedia profiler tool: profiling the connectivity of entity pairs in dbpedia. In: *Proceed-ings of the 5th International Workshop on Intelligent Exploration of Semantic Data (IESD 2016).[GS Search]*. [S.l.: s.n.], 2016.

HORROCKS, I. Ontologies and the semantic web. *Comm of the ACM*, ACM New York, NY, USA, v. 51, n. 12, p. 58–67, 2008.

ISOTANI, S.; BITTENCOURT, I. I. *Dados Abertos Conectados: Em busca da Web do Conhecimento*. [S.l.]: Novatec Editora, 2015.

KÖPCKE, H.; RAHM, E. Frameworks for entity matching: A comparison. *Data & Knowledge Engineering*, Elsevier, v. 69, n. 2, p. 197–210, 2010.

KÖPCKE, H.; THOR, A.; RAHM, E. Evaluation of entity resolution approaches on real-world match problems. *Proceedings of the VLDB Endowment*, VLDB Endowment, v. 3, n. 1-2, p. 484–493, 2010.

LAUFER, C. Guia de web semântica. *Gov. do Estado de SP e Gov. do Reino Unido*, 2015.

LEHMANN, J.; ISELE, R.; JAKOB, M.; JENTZSCH, A.; KONTOKOSTAS, D.; MENDES, P. N.; HELLMANN, S.; MORSEY, M.; KLEEF, P. V.; AUER, S. et al. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, IOS Press, v. 6, n. 2, p. 167–195, 2015.

LEHMANN, J.; SCHÜPPEL, J.; AUER, S. Discovering unknown connections–the dbpedia relationship finder. In: GESELLSCHAFT FÜR INFORMATIK E. V. *The Social Semantic Web 2007–Proceedings of the 1st Conference on Social Semantic Web (CSSW)*. [S.l.], 2007.

LI, Q.; JI, H. Incremental joint extraction of entity mentions and relations. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. [S.l.: s.n.], 2014. p. 402–412.

MANDREOLI, F.; MONTANGERO, M. Dealing with data heterogeneity in a data fusion perspective: Models, methodologies, and algorithms. In: *Data Handling in Science and Technology*. [S.l.]: Elsevier, 2019. v. 31, p. 235–270.

MAYNARD, D.; BONTCHEVA, K.; AUGENSTEIN, I. Natural language processing for the semantic web. *Synthesis Lectures on the SW: Theory and Technology*, Morgan & Claypool Publishers, v. 6, n. 2, p. 1–194, 2016.

MCGUINNESS, D. L. Question answering on the semantic web. *IEEE Intelligent Systems*, IEEE, v. 19, n. 1, p. 82–85, 2004.

MIWA, M.; BANSAL, M. End-to-end relation extraction using lstms on sequences and tree structures. *arXiv preprint arXiv:1601.00770*, 2016.

MIWA, M.; SASAKI, Y. Modeling joint entity and relation extraction with table representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. [S.l.: s.n.], 2014. p. 1858–1869.

NENTWIG, M.; HARTUNG, M.; NGOMO, A.-C. N.; RAHM, E. A survey of current link discovery frameworks. *Semantic Web*, IOS Press, v. 8, n. 3, p. 419–436, 2017.

NGOMO, A.-C. N.; AUER, S. Limes—a time-efficient approach for large-scale link discovery on the web of data. In: *Twenty-Second International Joint Conference on Artificial Intelligence*. [S.l.: s.n.], 2011.

OLIVEIRA, F. A. de; COSTA, R.; GOLDSCHMIDT, R.; CAVALCANTI, M. Multirelation association rule mining on datasets of the web of data. In: ACM. *Proceedings of the XV Brazilian Symposium on Information Systems*. [S.l.], 2019. p. 61.

OLIVEIRA, F. A. de; MARTINS, Y. C.; ROCHA, D. S.; SIQUEIRA, M. F. de; SILVA, L. A. E. da; COSTA, R. L.; GOLDSCHMIDT, R. R.; CAVALCANTI, M. C. Jabotg: Extending the herbarium dataset frontiers. In: *11th International Conference on Metadata and Semantics Research (MTSR'17)*. [S.l.: s.n.], 2017. p. 45–53.

PARIS, P.-H. Assessing the quality of owl: sameas links. In: SPRINGER. *European Semantic Web Conference*. [S.l.], 2018. p. 304–313.

RAMEZANI, R.; SARAEE, M.; NEMATBAKHSH, M. A. et al. Mrar: mining multi-relation association rules. *Journal of Computing and Security*, University of Isfahan and Iranian Society of Cryptology, v. 1, n. 2, p. 133–158, 2014.

REN, X.; WU, Z.; HE, W.; QU, M.; VOSS, C. R.; JI, H.; ABDELZAHER, T. F.; HAN, J. Cotype: Joint extraction of typed entities and relations with knowledge bases. In: *Proceedings of the 26th International Conference on World Wide Web*. [S.l.: s.n.], 2017. p. 1015–1024.

SCHMACHTENBERG, M.; BIZER, C.; PAULHEIM, H. Adoption of the linked data best practices in different topical domains. In: SPRINGER. *International Semantic Web Conference*. [S.l.], 2014. p. 245–260.

SHERIF, M. A.; NGOMO, A.-C. N.; LEHMANN, J. Automating rdf dataset transformation and enrichment. In: SPRINGER. *European Semantic Web Conference*. [S.l.], 2015. p. 371–387.

SHI, P.; LIN, J. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*, 2019.

SILVA, J. F. C. d. Etl4lod+: evolução do suporte ao ciclo de publicação de dados conectados. Universidade Federal do Rio de Janeiro, 2018.

SILVEIRA, R. N. da; CAVALCANTI, M. C. Método para rotular ligações semânticas na web de dados. In: *SBBD 2020 - Full Papers ()*. [s.n.], 2020. Disponível em: <https://sbbd.org.br/2020/wp-content/uploads/sites/13/2020/09/Rotular-ligacoes-ST3.pdf>.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2017. p. 5998–6008.

WANG, S.; ZHANG, Y.; CHE, W.; LIU, T. Joint extraction of entities and relations based on a novel graph scheme. In: *IJCAI*. [S.l.: s.n.], 2018. p. 4461–4467.

YU, L. *A developer's guide to the SW*. [S.l.]: Springer Science & Business Media, 2014.

YU, X.; LAM, W. Jointly identifying entities and extracting relations in encyclopedia text via a graphical model approach. In: *Coling 2010: Posters*. [S.l.: s.n.], 2010. p. 1399–1407.

ZHOU, G.; SU, J.; ZHANG, J.; ZHANG, M. Exploring various knowledge in relation extraction. In: *Proceedings of the 43rd annual meeting of the association for computational linguistics (acl'05)*. [S.l.: s.n.], 2005. p. 427–434.

APÊNDICE A – AMOSTRA DO RESULTADO DA TRANSFORMAÇÃO ETL REALIZADA NO ESTUDO DE CASO COM O SIAGAS

<http://lodsiagas/resource/basin/state/Bacia_metropolitana>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Manganese> .

<http://lodsiagas/resource/basin/state/Bacia_metropolitana>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Nitrate> .

<http://lodsiagas/resource/basin/state/Bacia_metropolitana>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Nitrite> .

<http://lodsiagas/resource/basin/state/Bacia_metropolitana>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Potassium> .

<http://lodsiagas/resource/basin/state/Bacia_metropolitana>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Silicon_Dioxide> .

<http://lodsiagas/resource/basin/state/Bacia_metropolitana>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Sodium> .

<http://lodsiagas/resource/basin/state/Bacia_metropolitana>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Sulfate> .

<http://lodsiagas/resource/basin/state/Bacias_da_lagoa_de_jacarepagua>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Potassium> .

<http://lodsiagas/resource/basin/state/Pequenas_bacias_litoraneas>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Calcium> .

<http://lodsiagas/resource/basin/state/Pequenas_bacias_litoraneas>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Chloride> .

<http://lodsiagas/resource/basin/state/Pequenas_bacias_litoraneas>

<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Iron> .

<http://lodsiagas/resource/basin/state/Pequenas_bacias_litoraneas>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Magnesium> .

<http://lodsiagas/resource/basin/state/Pequenas_bacias_litoraneas>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Nitrate> .

<http://lodsiagas/resource/basin/state/Pequenas_bacias_litoraneas>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Nitrite> .

<http://lodsiagas/resource/basin/state/Bacia_do_rio_vaza-barris>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Iron> .

<http://lodsiagas/resource/basin/state/Bacia_do_rio_vaza-barris>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Magnesium> .

<http://lodsiagas/resource/basin/state/Bacia_do_rio_vaza-barris>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Nitrate> .

<http://lodsiagas/resource/basin/state/Bacia_do_rio_vaza-barris>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Nitrite> .

<http://lodsiagas/resource/basin/state/Bacia_do_rio_vaza-barris>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Potassium> .

<http://lodsiagas/resource/basin/state/Bacia_do_rio_vaza-barris>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Silicon_Dioxide> .

<http://lodsiagas/resource/basin/state/Bacia_do_rio_vaza-barris>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Sodium> .

<http://lodsiagas/resource/basin/state/Bacia_do_rio_vaza-barris>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Sulfate> .

<http://lodsiagas/resource/basin/state/Bacia_do_rio_verde>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Aluminum> .

<http://lodsiagas/resource/basin/state/Bacia_do_rio_verde>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Calcium> .

<http://lodsiagas/resource/basin/state/Bacia_do_rio_verde>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Chloride> .

<http://lodsiagas/resource/basin/state/Bacia_do_rio_verde>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Chromium> .

<http://lodsiagas/resource/basin/state/Bacia_do_rio_verde>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Copper> .

<http://lodsiagas/resource/basin/state/Bacia_do_rio_tocantins>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Lead> .

<http://lodsiagas/resource/basin/state/Bacia_do_rio_tocantins>
<http://lodsiagas/property/containsChemical>
<http://lodsiagas/resource/chemical/Magnesium> .

<http://lodsiagas/resource/chemical/Calcium>
<http://lodsiagas/property/isChemicalOf>
<http://lodsiagas/resource/uf/Rio_Grande_do_Sul> .

<http://lodsiagas/resource/chemical/Carbon_Dioxide>
<http://lodsiagas/property/isChemicalOf>
<http://lodsiagas/resource/uf/Rio_Grande_do_Sul> .

<http://lodsiagas/resource/chemical/Chloride>
<http://lodsiagas/property/isChemicalOf>
<http://lodsiagas/resource/uf/Rio_Grande_do_Sul> .

<http://lodsiagas/resource/chemical/Chloride>
<http://lodsiagas/property/isChemicalOf>
<http://lodsiagas/resource/uf/Rio_de_Janeiro_(state)> .

<http://lodsiagas/resource/chemical/Hydroxide>
<http://lodsiagas/property/isChemicalOf>
<http://lodsiagas/resource/uf/Rio_de_Janeiro_(state)> .

<http://lodsiagas/resource/chemical/Iron>
<http://lodsiagas/property/isChemicalOf>
<http://lodsiagas/resource/uf/Rio_de_Janeiro_(state)> .

<http://lodsiagas/resource/chemical/Magnesium>
<http://lodsiagas/property/isChemicalOf>

<http://lodsiagas/resource/uf/Rio_de_Janeiro_(state)> .

<http://lodsiagas/resource/chemical/Nitrate>
<http://lodsiagas/property/isChemicalOf>
<http://lodsiagas/resource/uf/Rio_de_Janeiro_(state)> .

<http://lodsiagas/resource/chemical/Nitrate>
<http://lodsiagas/property/isChemicalOf>
<http://lodsiagas/resource/uf/Amazonas_State> .

<http://lodsiagas/resource/chemical/Potassium>
<http://lodsiagas/property/isChemicalOf>
<http://lodsiagas/resource/uf/Amazonas_State> .

<http://lodsiagas/resource/chemical/Selenium>
<http://lodsiagas/property/isChemicalOf>
<http://lodsiagas/resource/uf/Amazonas_State> .

<http://lodsiagas/resource/chemical/Silicon_Dioxide>
<http://lodsiagas/property/isChemicalOf>
<http://lodsiagas/resource/uf/Amazonas_State> .

<http://lodsiagas/resource/chemical/Sodium>
<http://lodsiagas/property/isChemicalOf>
<http://lodsiagas/resource/uf/Amazonas_State> .

<http://lodsiagas/resource/chemical/Sulfate>
<http://lodsiagas/property/isChemicalOf>
<http://lodsiagas/resource/uf/Amazonas_State> .

<http://lodsiagas/resource/chemical/Zinc>
<http://lodsiagas/property/isChemicalOf>
<http://lodsiagas/resource/uf/Amazonas_State> .

<http://lodsiagas/resource/chemical/Calcium>
<http://lodsiagas/property/isChemicalOf>
<http://lodsiagas/resource/uf/Bahia> .

<http://lodsiagas/resource/chemical/Chloride>
<http://lodsiagas/property/isChemicalOf>
<http://lodsiagas/resource/uf/Bahia> .

<http://lodsiagas/resource/chemical/Fluoride>
<http://lodsiagas/property/isChemicalOf>
<http://lodsiagas/resource/uf/Bahia> .

<http://lodsiagas/resource/chemical/Iron>
<http://lodsiagas/property/isChemicalOf>
<http://lodsiagas/resource/uf/Bahia> .

APÊNDICE B – RETORNO DA PLAIN NA BUSCA POR PREDICADOS RELACIONADOS A *DUL:CHEMICALOBJECT*

<CHEMICAL SUBSTANCE> <AGGREGATION> < >

<CHEMICAL SUBSTANCE> <BOILING POINT (K)> < >

<CHEMICAL SUBSTANCE> <CARCINOGEN> < >

< > <CLASSIFIES> <ENTITY>

< > <CO-PARTICIPATES WITH> <OBJECT>

<CHEMICAL SUBSTANCE> <DENSITY> < >

< > <DESCRIBES> <ENTITY>

< > <DIRECTLY FOLLOWS> <ENTITY>

< > <DIRECTLY PRECEDES> <ENTITY>

< > <ELEMENT ABOVE> <CHEMICAL SUBSTANCE>

< > <FAR FROM> <ENTITY>

<CHEMICAL SUBSTANCE> <FLASH POINT> < >

< > <FOLLOWS> <ENTITY>

< > <FORMALLY REPRESENTS> <ENTITY>

<CHEMICAL SUBSTANCE> <FORMULA> < >

< > <HAS COMMON BOUNDARY> <ENTITY>

< > <HAS COMPONENT> <ENTITY>

<ENTITY> <HAS CONSTRAINT> < >

<ENTITY> <HAS DATA VALUE> < >

< > <HAS GROUNDING> <ENTITY>

< > <HAS INTERPRETATION>

< > <HAS LOCATION> <ENTITY>

< > <HAS MEMBER> <ENTITY>

< > <HAS PART> <ENTITY>

< > <HAS PARTICIPANT> <OBJECT>

<ENTITY> <HAS QUALITY> < >

<ENTITY> <HAS REGION> < >

<OBJECT> <HAS ROLE> < >

<ENTITY> <HAS SETTING> < >

<ENTITY> <HAS TOPIC> < >

< > <INCLUDES OBJECT> <OBJECT>

< > <IS ABOUT> <ENTITY>

<ENTITY> <IS CLASSIFIED BY> < >

< > <IS COMPONENT OF> <ENTITY>

< > <IS CONSTITUENT OF> <ENTITY>

< > <IS CONSTRAINT FOR> <ENTITY>

<ENTITY> <IS DESCRIBED BY> < >

< > <IS FORMALLY INTERPRETED AS> <ENTITY>

<ENTITY> <IS FORMALLY REPRESENTED IN> < >

<ENTITY> <IS GROUNDING FOR> < >

< > <IS IN THE SAME SETTING AS> <ENTITY>

< > <IS INTERPRETATION OF> <ENTITY>

< > <IS LOCATION OF> <ENTITY>

<ENTITY> <IS MEMBER OF> < >

<OBJECT> <IS OBJECT INCLUDED IN> < >

<ENTITY> <IS OBSERVABLE AT>

< > <IS PART OF> <ENTITY>

OBJECT> <IS PARTICIPANT IN> < >

< > <IS QUALITY OF> <ENTITY>

<ENTITY> <IS REFERENCE OF> < >

<ENTITY> <IS REFERENCE OF INFORMATION REALIZED BY> < >

< > <IS REGION FOR> <ENTITY>

< > <IS ROLE OF> <OBJECT>

< > <IS SETTING FOR> <ENTITY>

< > <IS TIME OF OBSERVATION OF> <ENTITY>

<ENTITY> <IS TOPIC OF> < >

<CHEMICAL SUBSTANCE> <LETHAL WHEN GIVEN TO CHICKENS> < >

<CHEMICAL SUBSTANCE> <LETHAL WHEN GIVEN TO MICE> < >

<CHEMICAL SUBSTANCE> <LETHAL WHEN GIVEN TO RABBITS> < >

<CHEMICAL SUBSTANCE> <LETHAL WHEN GIVEN TO RATS> < >

<CHEMICAL SUBSTANCE> <MELTING POINT (K)> < >

<CHEMICAL SUBSTANCE> <MOLECULAR WEIGHT> < >

< > <NEAR TO> <ENTITY>

< > <NOT SOLUBLE IN> <CHEMICAL SUBSTANCE>

< > <OVERLAPS> <ENTITY>

< > <PRECEDES> <ENTITY>

<CHEMICAL SUBSTANCE> <PUBCHEM> < >

< > <REALIZES INFORMATION ABOUT> <ENTITY>

<CHEMICAL SUBSTANCE> <SOLUBILITY> < >

< > <SOLVENT WITH BAD SOLUBILITY> <CHEMICAL SUBSTANCE>

< > <SOLVENT WITH GOOD SOLUBILITY> <CHEMICAL SUBSTANCE>

< > <SOLVENT WITH MEDIOCRE SOLUBILITY> <CHEMICAL SUBSTANCE>

APÊNDICE C – SCRIPT EM PYTHON PARA EXTRAÇÃO SUPERVISIONADA DE RELAÇÕES EM VÁRIAS SENTENÇAS

```
import sys
import json

print('Iniciando...')
file = open("./plain_re/entrada.json", "r", encoding='UTF8')
sys.path.insert(1, './opennre/')
print('[OK]')

print('Carregando OpenNRE...')
import opennre
print('[OK]')

print('Carrengado Modelo BERT...')
model = opennre.get_model('wiki80_bert_softmax')
print('[OK]')

data = {}
data['sugestoes'] = []

print('Criando sugestões...')

with file as f:

    for line in f:

        jsonLinha = json.loads(line)
        #print(jsonLinha)

        texto = jsonLinha['text']
        h1 = jsonLinha['h']['pos'][0]
        h2 = jsonLinha['h']['pos'][1]
        t1 = jsonLinha['t']['pos'][0]
        t2 = jsonLinha['t']['pos'][1]

        tupleModelInfer = model.infer({'text': texto,
        'h': {'pos': (h1, h2)},
```

```
        't': {'pos': (t1, t2)}})

        data_set = {"predicado": tupleModelInfer[0],
        "indice": tupleModelInfer[1]}
        data['sugestoes'].append(data_set)

        json_dump = json.dumps(data)

        json_object = json.loads(json_dump)

        sorted_sugestoes = dict(json_object)
        sorted_sugestoes["sugestoes"] = sorted(json_object["sugestoes"],
        key=lambda x: x["indice"], reverse=True)

print('[OK]')

with open("./plain_re/saida.json", 'w') as outfile:
     json.dump(sorted_sugestoes, outfile)

print('Sugestões criadas com sucesso!')
```

APÊNDICE D – ORAÇÕES OFERECIDAS COMO ENTRADA PARA A PLAIN RE NO ESTUDO DE CASO COM QUÍMICOS

{"text": "The inorganic anions nitrate (NO3-) and nitrite (NO2-) were previously thought to be inert end products of endogenous nitric oxide (NO) metabolism.", "h": {"pos": [21, 28]}, "t": {"pos": [40, 47]}}

{"text": "Nitrate and nitrite ions are used as food additives to inhibit the growth of microorganisms in cured and processed meats.", "h": {"pos": [0, 7]}, "t": {"pos": [12, 19]}}

{"text": "Vegetables contain significant quantities of nitrate and nitrite.", "h": {"pos": [45, 52]}, "t": {"pos": [57, 64]}}

{"text": "Inorganic nitrate (NO3 -), nitrite (NO2 -) and NO are nitrogenous species with a diverse and interconnected chemical biology.", "h": {"pos": [10, 17]}, "t": {"pos": [28, 35]}}

{"text": "There was no significant difference between fluoride varnish and potassium nitrate in the reduction of pre-cementation sensitivity while one week after cementation, sensitivity was more relieved by potassium nitrate compared to fluoride varnish (p = 0.023).", "h": {"pos": [44, 52]}, "t": {"pos": [75, 82]}}

{"text": "In designing of the hydroxypyridinones (HPOs) as the therapeutic chelating agents for iron and aluminium overload pathologies, quantum mechanical (QM) calculations are necessary for predicting the binding energies and thermodynamic parameters of the metal-HPO complexes.", "h": {"pos": [86, 90]}, "t": {"pos": [95, 104]}}

{"text": "Despite such increases in iron content there were no significant changes in the activities of a wide range of cytoprotective enzymes apart from an increase in superoxide dismutase in the frontal cortex of the aluminium loaded rats.", "h": {"pos": [26, 30]}, "t": {"pos": [209, 218]}}

{"text": "With the increase of exposure time, lead exposure can changes in the contents of copper and iron in different brain tissues, body fluids and barriers in rats, among which, the contents of copper and iron in the amygdala, cerebrospinal fluid and brain microvessels increase significantly.", "h": {"pos": [36, 40]}, "t": {"pos": [92, 96]}}

{"text": "Certain five heavy metals viz. arsenic (As), cadmium (Cd), chromium (Cr)(VI), mercury (Hg), and lead (Pb) are non-threshold toxins and can exert toxic effects at very low concentrations.", "h": {"pos": [59, 67]}, "t": {"pos": [96, 100]}}

{"text": "This kind of metals such as Chromium and Lead could affect health and the ecosystem.", "h": {"pos": [28, 36]}, "t": {"pos": [41, 45]}}

{"text": "However, the interaction of iron with nitrite or nitrate present in the sludge has received little attention.", "h": {"pos": [28, 32]}, "t": {"pos": [38, 45]}}

{"text": "Among many harmful contaminants, nitrate and fluoride ions are more common.", "h": {"pos": [33, 40]}, "t": {"pos": [45, 53]}}

APÊNDICE E – TRIPLAS RELACIONANDO MEDICAMENTOS A DOENÇAS NA DBPEDIA

<http://dbpedia.org/resource/Nicotine>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>
<http://dbpedia.org/resource/Nicotine_withdrawal>

<http://dbpedia.org/resource/Pertussis_vaccine>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>
<http://dbpedia.org/resource/Pertussis>

<http://dbpedia.org/resource/Measles_vaccine>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Measles>

<http://dbpedia.org/resource/Dryvax>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Smallpox>

<http://dbpedia.org/resource/Polio_vaccine>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Poliomyelitis>

<http://dbpedia.org/resource/Benzylpenicillin>
<http://purl.org/linguistics/gold/hypernym>
<http://dbpedia.org/resource/Spectrum>

<http://dbpedia.org/resource/Dengue_vaccine>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Dengue_fever>

<http://dbpedia.org/resource/Rabies_vaccine>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Rabies>

<http://dbpedia.org/resource/ACAM2000>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Smallpox>

<http://dbpedia.org/resource/Matilda_Moldenhauer_Brooks>
<http://dbpedia.org/ontology/knownFor>
<http://dbpedia.org/resource/Cyanide_poisoning>

<http://dbpedia.org/resource/Rilotumumab>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Hepatocyte_growth_factor>

<http://dbpedia.org/resource/Memorial_Sloan_Kettering_Cancer_Center>
<http://dbpedia.org/property/speciality>
<http://dbpedia.org/resource/Cancer>

<http://dbpedia.org/resource/Amlodipine>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>
<http://dbpedia.org/resource/Calcium_channel_blocker_toxicity>

<http://dbpedia.org/resource/Midazolam>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>
<http://dbpedia.org/resource/Benzodiazepine_overdose>

<http://dbpedia.org/resource/Christian_J._Lambertsen>
<http://dbpedia.org/ontology/field>
<http://dbpedia.org/resource/Diving_medicine>

<http://dbpedia.org/resource/CYT006-AngQb>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Hypertension>

<http://dbpedia.org/resource/ACE_inhibitor>
<http://dbpedia.org/property/use>
<http://dbpedia.org/resource/Hypertension>

<http://dbpedia.org/resource/Flunitrazepam>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>
<http://dbpedia.org/resource/Benzodiazepine_overdose>

<http://dbpedia.org/resource/Epinephrine>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>
<http://dbpedia.org/resource/Novelty_seeking>

<http://dbpedia.org/resource/Amphetamine>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>
<http://dbpedia.org/resource/Stimulant_psychosis>

<http://dbpedia.org/resource/BCG_vaccine>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Tuberculosis>

<http://dbpedia.org/resource/Antipsychotic>
<http://dbpedia.org/property/use>
<http://dbpedia.org/resource/Schizophrenia>

<http://dbpedia.org/resource/Morphine>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>
<http://dbpedia.org/resource/Opioid_overdose>

<http://dbpedia.org/resource/Lysergic_acid_diethylamide>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>

<http://dbpedia.org/resource/Flashback_(psychology)>

<http://dbpedia.org/resource/Morphine>
<http://www.w3.org/2002/07/owl#differentFrom>
<http://dbpedia.org/resource/Morphea>

<http://dbpedia.org/resource/Triazolam>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>
<http://dbpedia.org/resource/Benzodiazepine_overdose>

<http://dbpedia.org/resource/Imatinib>
<http://dbpedia.org/property/use>
<http://dbpedia.org/resource/Chronic_myelogenous_leukemia>

<http://dbpedia.org/resource/Fluoxetine>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>
<http://dbpedia.org/resource/Serotonin_syndrome>

<http://dbpedia.org/resource/Cholera_vaccine>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Cholera>

<http://dbpedia.org/resource/Alvin_J._Siteman_Cancer_Center>
<http://dbpedia.org/property/speciality>
<http://dbpedia.org/resource/Cancer>

<http://dbpedia.org/resource/Sipuleucel-T>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Prostate_cancer>

<http://dbpedia.org/resource/Mumps_vaccine>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Mumps>

<http://dbpedia.org/resource/Chlordiazepoxide>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>
<http://dbpedia.org/resource/Benzodiazepine_overdose>

<http://dbpedia.org/resource/Anthrax_Vaccine_Adsorbed>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Anthrax>

<http://dbpedia.org/resource/Selective_serotonin_reuptake_inhibitor>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>
<http://dbpedia.org/resource/Serotonin_syndrome>

<http://dbpedia.org/resource/Fenethylline>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>
<http://dbpedia.org/resource/Substance_abuse>

<http://dbpedia.org/resource/Tetrazepam>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>
<http://dbpedia.org/resource/Benzodiazepine_withdrawal_syndrome>

<http://dbpedia.org/resource/Tetrazepam>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>
<http://dbpedia.org/resource/Benzodiazepine_overdose>

<http://dbpedia.org/resource/Ficlatuzumab>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Hepatocyte_growth_factor>

<http://dbpedia.org/resource/Flanvotumab>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Hepatocyte_growth_factor>

<http://dbpedia.org/resource/Rubella_vaccine>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Rubella>

<http://dbpedia.org/resource/Typhus_vaccine>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Typhus>

<http://dbpedia.org/resource/Tetanus_vaccine>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Tetanus>

<http://dbpedia.org/resource/Vancomycin>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>
<http://dbpedia.org/resource/Erythroderma>

<http://dbpedia.org/resource/Vemurafenib>
<http://dbpedia.org/property/use>
<http://dbpedia.org/resource/Melanoma>

<http://dbpedia.org/resource/Live_attenuated_influenza_vaccine>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Influenza>

<http://dbpedia.org/resource/Alprazolam>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>
<http://dbpedia.org/resource/Benzodiazepine_dependence>

<http://dbpedia.org/resource/Lorazepam>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>
<http://dbpedia.org/resource/Benzodiazepine_overdose>

<http://dbpedia.org/resource/Temazepam>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>

<http://dbpedia.org/resource/Benzodiazepine_withdrawal_syndrome>

<http://dbpedia.org/resource/Trovafloxacin>
<http://purl.org/linguistics/gold/hypernym>
<http://dbpedia.org/resource/Spectrum>

<http://dbpedia.org/resource/Prazepam>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>
<http://dbpedia.org/resource/Benzodiazepine_withdrawal_syndrome>

<http://dbpedia.org/resource/Prazepam>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>
<http://dbpedia.org/resource/Benzodiazepine_overdose>

<http://dbpedia.org/resource/Etizolam>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>
<http://dbpedia.org/resource/Benzodiazepine_overdose>

<http://dbpedia.org/resource/Fenbendazole>
<http://purl.org/linguistics/gold/hypernym>
<http://dbpedia.org/resource/Spectrum>

<http://dbpedia.org/resource/Sevirumab>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Cytomegalovirus>

<http://dbpedia.org/resource/Eprazinone>
<http://purl.org/linguistics/gold/hypernym>
<http://dbpedia.org/resource/Bronchospasm>

<http://dbpedia.org/resource/Imvanex>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Smallpox>

<http://dbpedia.org/resource/Prostvac>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Prostate_cancer>

<http://dbpedia.org/resource/Hormonal_breast_enhancement>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>
<http://dbpedia.org/resource/Cancer>

<http://dbpedia.org/resource/Nitrazepam>
<http://www.w3.org/2000/01/rdf-schema#seeAlso>
<http://dbpedia.org/resource/Benzodiazepine_withdrawal_syndrome>

<http://dbpedia.org/resource/Zoster_vaccine>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Chickenpox>

<http://dbpedia.org/resource/Zoster_vaccine>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Postherpetic_neuralgia>

<http://dbpedia.org/resource/Zoster_vaccine>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Ramsay_Hunt_syndrome_type_II>

<http://dbpedia.org/resource/Hepatitis_B_vaccine>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Hepatitis_B>

<http://dbpedia.org/resource/Hepatitis_A_vaccine>
<http://dbpedia.org/property/target>
<http://dbpedia.org/resource/Hepatitis_A>

<http://dbpedia.org/resource/Oxfendazole>
<http://purl.org/linguistics/gold/hypernym>
<http://dbpedia.org/resource/Spectrum>

<http://dbpedia.org/resource/Roland_MC-307>
<http://dbpedia.org/property/fx>
<http://dbpedia.org/resource/Delay_(audio_effect)>

<http://dbpedia.org/resource/Treatment_and_prognosis_of_renal_cell_carcinoma>
<http://dbpedia.org/ontology/wikiPageRedirects>
<http://dbpedia.org/resource/Renal_cell_carcinoma>

APÊNDICE F – SENTENÇAS OFERECIDAS COMO ENTRADA PARA A PLAIN RE NO ESTUDO DE CASO ENTRE MEDICAMENTOS E DOENÇAS

{"text": "Detailed phenotyping of disease presentation together with comprehensive representation of drug mechanism of action is considered as a path forward, and a big data spectrum has become available covering behavioral, clinical and molecular characteristics, the latter combining reductionist and explorative strategies.", "h": {"pos": [6, 25]}, "t": {"pos": [24, 31]}}

{"text": "It is suggested that apelin positively affects the treatment of non-cancerous diseases and may be considered as a therapeutic drug in many illnesses.", "h": {"pos": [126, 130]}, "t": {"pos": [78, 86]}}

{"text": "Inferring drug-disease associations is critical in unveiling disease mechanisms, as well as discovering novel functions of available drugs, or drug repositioning.", "h": {"pos": [133, 138]}, "t": {"pos": [61, 68]}}

{"text": "Drugs are often seen as ancillary to the purpose of fighting diseases.", "h": {"pos": [0, 5]}, "t": {"pos": [61, 69]}}

{"text": "Strong evidence exists to support the use of metoprolol, timolol, propranolol, divalproex sodium, sodium valproate, and topiramate for migraine prevention, according to the AAN.", "h": {"pos": [135, 143]}, "t": {"pos": [105, 114]}}

{"text": "Disulfiram has no proven effect on the long-term outcome of alcoholism.", "h": {"pos": [0, 10]}, "t": {"pos": [60, 70]}}

{"text": "In this population of critically ill youth, short-term use of quetiapine as treatment for delirium appears to be safe, without serious adverse events.", "h": {"pos": [90, 98]}, "t": {"pos": [62, 72]}}

{"text": "Quetiapine in an atypical antipsychotic drug that is frequently used for delirium and behavioral and psychological symptoms in dementia.", "h": {"pos": [73, 81]}, "t": {"pos": [0, 10]}}

{"text": "This case illustrates the problems associated with the diagnosis of erythroderma in intensive care patients and confirms that it is possible to prescribe vancomycin in cases with allergic reaction to teicoplanin.", "h": {"pos": [154, 164]}, "t": {"pos": [68, 80]}}