Universidade Federal do Rio de Janeiro
Curso de Bacharelado em Ciência da Computação

Amanda de Oliveira Medeiros
Julia Anne de Souza Alves

Triplificação de Dados Conectados: Uma Experimentação de Ferramentas no Contexto do Portal da Campanha Contra o Uso de Agrotóxicos

Rio de Janeiro
2018

Universidade Federal do Rio de Janeiro
Curso de Bacharelado em Ciência da Computação

Amanda de Oliveira Medeiros
Julia Anne de Souza Alves

Triplificação de Dados Conectados: Uma Experimentação de Ferramentas no Contexto do Portal da Campanha Contra o Uso de Agrotóxicos

Monografia apresentada para obtenção do
Grau de Bacharel em Ciência da Computação
pela Universidade Federal do Rio de Janeiro.

Orientadora:
Profa. Ph.D. Maria Luiza Machado Campos

Co-orientadora:
Karen Torres Teixeira

Rio de Janeiro
2018

Medeiros, Amanda de Oliveira
M488t Triplificação de dados conectados: uma experimentação de ferramentas no contexto do portal da campanha contra o uso de agrotóxicos / Amanda de Oliveira Medeiros, Julia Anne de Souza Alves. – 2018.

109 f.

Orientadora: Maria Luiza Machado Campos.

Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) - Universidade Federal do Rio de Janeiro, Instituto de Matemática, Bacharel em Ciência da Computação, 2018.

1. Web semântica. 2. Triplificação de dados. 3. Agrotóxicos. 4. Ferramentas de triplificação. I. Alves, Julia Anne de Souza. II. Campos, Maria Luiza Machado (Orient.). III. Universidade Federal do Rio de Janeiro, Instituto de Matemática. IV. Título.

Triplificação de Dados Conectados: Uma Experimentação de Ferramentas no Contexto do Portal da Campanha Contra o Uso de Agrotóxicos

Amanda de Oliveira Medeiros
Julia Anne de Souza Alves

Projeto Final de Curso submetido ao Departamento de Ciência da Computação do Instituto de Matemática da Universidade Federal do Rio de Janeiro como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Apresentado por:

Amanda de Oliveira Medeiros

Julia Anne de Souza Alves

Aprovado por:

Profa. Maria Luiza Machado Campos

Profa. Giseli Rabello Lopes

Karen Torres Teixeira

Rio de Janeiro, RJ - Brasil

30 de novembro de 2018

AGRADECIMENTOS

Meus mais profundos agradecimentos ao amor da minha vida, Elson, por ter me apoiado ao longo de todo o processo de realização deste trabalho, com conselhos e gestos sem os quais tudo teria sido muito mais difícil. Seu apoio foi essencial para que eu pudesse fazer as melhores escolhas possíveis e realizar este projeto.

Também agradeço à minha parceira por realizar não só este trabalho comigo, mas também pela parceria ao longo da maior parte da minha trajetória no curso de Ciência da Computação. Obrigada por me ouvir e compartilhar suas dores e conhecimentos durante os últimos anos.

Amanda de Oliveira Medeiros

Aos meus pais, Jerônimo e Ed, por todo amparo e apoio necessário para trilhar toda a graduação. Agradeço as palavras de amor nos momentos em que mais pensei em desistir ao longo desses anos.

À minha amiga e parceira deste trabalho e de toda a jornada nesse curso, Amanda. Obrigada por me entender como ninguém e por ser a minha sanidade em tantos momentos da vida pessoal e acadêmica.

Ao Rodrigo Carestiato Gonçalves de Souza que contribuiu com as soluções que foram aplicadas neste trabalho e com incentivo diário à finalização do mesmo.

Julia Anne de Souza Alves

RESUMO

O consumo de agrotóxicos no Brasil é um assunto preocupante, visto que o país é um dos maiores produtores agrícolas do mundo e figura entre os países com maior consumo de agrotóxicos. Em contrapartida, a disponibilidade de dados sobre o tema não possui tamanha proporção, sendo grande parte das bases de dados disponibilizadas em formatos que não são de fácil interligação com outras bases de dados existentes. Entretanto, existe uma tendência de representar dados em uma forma rica e mais flexível, visando facilitar a interligação de dados, de modo que o dado é representado em um grão menor e anotado com metadados para reduzir a ambiguidade. Esta forma de representar dados é feita a partir da construção de triplas, mas executar o processo de triplificação de dados é uma tarefa complexa.

Visando fazer uma experimentação de ferramentas de triplificação e colaborar na interligação de dados sobre agrotóxicos, esclarecendo assim ainda mais o processo de triplificação por meio de uma experimentação, este trabalho se propõe a triplificar duas bases de dados acerca de agrotóxicos, disponíveis no Portal da Campanha Contra o Uso de Agrotóxicos.

O resultado deste trabalho consiste na descrição detalhada do processo de triplificação realizado e na disponibilização das triplas de dados produzidas por este experimento, que representam cada uma das bases de dados selecionadas, possibilitando a interligação com outros dados gerados em trabalhos anteriores e que enriquecem as informações a respeito de agrotóxicos e seus efeitos na saúde humana.

Palavras-chave: Web Semântica. Triplificação de dados. Agrotóxicos. Ferramentas de triplificação.

ABSTRACT

Pesticide consumption in Brazil is a concerning subject, especially considering the country as both one of the world's most prominent farmers and a significant pesticide consumer of the world. In contrast, the availability of data about this topic has not such proportion, as most of the databases are in formats that are hard to link with existing databases. However, there's a trend to richly and flexibly represent data, aiming to make it easy to connect data, so that the data is served in a smaller grain and annotated with metadata to reduce ambiguity. This way of representing data is accomplished creating triples, but the process of converting data into triples is a complex task.

By experimenting data triplification tools and collaborating on the creation of linked data about pesticides, thus further clarifying the triplification process through experimentation, this paper proposes to triplicate two databases on pesticides, both available on the Portal of Campaign Against the Use of Pesticides.

The result of this work is the detailed description of the triplification process carried out and the availability of the triples produced by this experiment, which represent each of the selected databases, allowing the interconnection with other data obtained in previous works and that enrich the information regarding pesticides and their effects on human health.

Keywords: Semantic Web. Data triplification. Pesticides. Triplification tools.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

ABRASCO	Associação Brasileira de Saúde Coletiva
ANVISA	Agência Nacional de Vigilância Sanitária
API	Application Programming Interface
EMBRAPA	Empresa Brasileira de Pesquisa Agropecuária
FAO	Food and Agriculture Organization
HRW	Human Rights Watch
IBAMA	Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis
IDE	Integrated Development Environment
IUPAC	International Union of Pure and Applied Chemistry
JMPM	Joint Meeting in Pesticide Management
LMR	Limite Máximo de Resíduos em Alimento
LOD	Linked Open Data
NC-IUBMB	Nomenclature Committee of the International Union of Biochemistry and Molecular Biology
OMS	Organização Mundial da Saúde
ONU	Organização das Nações Unidas
OWL	Ontology Web Language
PAN	Pesticide Action Network
PARA	Programa de Análise de Resíduos de Agrotóxicos em Alimentos
PL	Projeto de Lei
R2RML	RDB to RDF mapping language
RDF	Resource Description Framework
SPARQL	SPARQL Protocol and RDF Query Language
SGBD	Sistema de Gerenciamento de Banco de Dados
XML	Extensible Markup Language

SUMÁRIO

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

Há mais de dez anos, o Brasil lidera o ranking mundial de consumo de agrotóxicos e, em consequência deste fato, efeitos negativos sobre a saúde dos brasileiros e sobre o meio ambiente vem sendo acarretados [ABRASCO, 2012]. O uso indiscriminado dos agrotóxicos não é exclusividade do Brasil, mas o país é um dos maiores produtores agrícolas do mundo, alimentando cerca de ¼ da população mundial [OMC, 2010]. Cerca de 80% dos agrotóxicos são usados nas mais diversas escalas e produções agrícolas e o Brasil ocupa o 7º lugar na relação de quantidade de produtos aplicados por hectare de terra cultivada, de acordo com dados da FAO - Organização das Nações Unidas para Agricultura e Alimentação. O Ministério do Meio Ambiente brasileiro informa que, no volume total, o Brasil ocupa o 1º lugar.

Nos últimos anos tem havido uma mobilização para o estudo das informações acerca deste tema no nosso país. Diversas informações vêm sendo divulgadas a partir de fontes de amplo espectro como áreas das ciências agrárias, biológicas, saúde coletiva, agroecologia, economia. No período de 2013 a 2015, foram estudadas amostras de alimentos produzidos em todo o Brasil e publicado no relatório das análises de amostras monitoradas pelo Programa de Análise de Resíduo de Agrotóxicos em Alimentos (PARA), onde mais de ⅓ apresentaram resíduos de agrotóxicos e 20% das amostras foram consideradas insatisfatórias [PROGRAMA DE ANÁLISE DE RESÍDUOS DE AGROTÓXICOS EM ALIMENTOS, 2016].

Apesar de todos estes fatos, o projeto de lei 6299/2202 aquece o tom dos debates que cercam o assunto. “O Projeto de Lei visa alterar a regulação dos agrotóxicos no Brasil e traz uma série de mudanças em relação a Lei dos Agrotóxicos atual que é de 1989. A proposta, apesar de ter passado por diversas

comissões na Câmara dos Deputados, só foi reprovada na Comissão de Seguridade Social e Família" [Câmara dos Deputados, 2002].

Diante disto, torna-se de extrema importância no contexto político econômico atual explorar os dados disponíveis sobre os agrotóxicos já comercializados e circulantes em escala mundial e seus possíveis impactos na saúde da população brasileira. Com os diversos questionamentos dos efeitos nocivos do uso de agrotóxicos que vêm sendo reunidos, existe hoje um grande número de bases de dados que fornecem informações relacionadas aos agrotóxicos e seus efeitos [TYGEL & CAMPOS, 2012].

Embora esses dados estejam disponíveis na Web, eles são fornecidos em formatos distintos e de forma não intuitiva de leitura e análise, agregando complexidade ao processamento destas informações de maneira simples e objetiva. No entanto, para que seja feito um estudo desta diversidade de informações com o intuito de se levantar questões relevantes acerca do uso de agrotóxicos, é necessário obter esses dados, tratá-los, processá-los e carregá-los posteriormente para serem analisados com base em diferente óticas.

Tendo em vista o grande volume de informações disponíveis e a falta de padronização dos dados que estão dispostos na internet, este problema agravado com o crescimento da mesma e aumentando exponencialmente, foi proposta uma solução para aplicar processamento automático à internet atual - a Web Semântica [BERNERS-LEE; HENDLER; LASSILA, 2001]. Nela cada informação disposta tem algum significado atribuído e possui uma associação com outros dados, trazendo contextualização àquelas informações nos conteúdos da Web. Utilizando conceitos de dados conectados, as bases de dados referentes a agrotóxicos serão tratadas seguindo conceitos de Web Semântica. Os conceitos apresentados acima serão explicados nos capítulos posteriores.

Como este assunto envolve claramente temas de natureza interdisciplinar, os dados disponíveis são oriundos das mais diversas fontes governamentais e científicas. Há, por esta razão, um esforço sendo realizado para que sejam extraídas

destes dados informações úteis e relevantes a respeito deste tema. Na UFRJ (Universidade Federal do Rio de Janeiro), vem sendo desenvolvido um trabalho para facilitar a exploração e enriquecimento dos dados disponíveis. O projeto "Observatório de Atenção Permanente ao Uso de Agrotóxicos – Portal de Informações Interligadas sobre Agrotóxicos e seus Efeitos sobre a Saúde e Meio Ambiente" criou e mantém um espaço virtual, que possibilita o armazenamento e recuperação de informações relacionadas ao uso de agrotóxicos no Brasil e aos danos que estes produtos vêm causando à saúde e ao meio ambiente [TYGEL; GONÇALVES; SANTOS; MARQUES; CAMPOS, 2015].

"É fundamental que os dados publicados no tema possam ser explorados de forma a apoiar a tomada de decisão sobre políticas públicas e ações concretas delas decorrentes" [TEIXEIRA, 2018, p.3]. Com o objetivo de experimentar as vantagens do uso de dados conectados que possuam conteúdo aberto, ou seja, LOD (Linked Open Data), no contexto de agrotóxicos e suas consequências para a saúde, neste trabalho, optou-se por analisar as bases de dados que contribuem para a análise desta questão e que se integrem com os trabalhos já desenvolvidos nas pesquisas da área. Ao analisar essas bases, poderá ser amplificado o conhecimento sobre os efeitos dos agrotóxicos em si e em quais países eles são usados, a fim de promover uma relação entre agrotóxicos e suas respectivas consequências na saúde dos habitantes dos países que usufruem dos mesmos. Com essa análise, também é possível obter mais uma evidência empírica de que o aparecimento de danos à saúde está diretamente relacionado ao consumo destas substâncias.

1.2 OBJETIVO

A transformação de dados a partir de qualquer formato para dados conectados consiste na geração de triplas, um formato de representação baseado em RDF (*Resource Description Framework*), que mantém o dado em um grão

pequeno, mais flexível para ser interligado com outros dados. Aproveitando-se dos dados abertos disponíveis relativos a agrotóxicos no Brasil e no mundo, este trabalho tem como objetivo realizar uma experimentação e avaliação das ferramentas existentes para triplificação de dados, isto é, transformação para dados conectados, com bases de dados reais. Estas ferramentas são de software livre e código aberto para permitir que o conhecimento aqui gerado possa ser difundido sem complicações.

Os dados aplicados neste trabalho, disponibilizados em diferentes bases de dados abertas para a sociedade, serão transformados em dados conectados e fornecidos para uma aplicação que permita visualização e consulta dos mesmos. Para isto serão realizadas as etapas de obtenção, limpeza e tratamento das bases, construção de uma modelagem com base em ontologias já existentes para facilitar o entendimento da semântica utilizada, triplificação e armazenamento em Linked Open Data. Após a finalização destas etapas, haverá a publicação destes dados para posteriores incorporações de mais informações.

O escopo inicial de dados trabalhados contemplados neste projeto é o da relação entre produtos químicos, com suas respectivas classes, com os sintomas crônicos e agudos causados, além do inventário de agrotóxicos banidos em cada país do mundo. Esse escopo foi escolhido para que as triplas geradas pudessem ser integradas com as fontes de dados resultantes do trabalho de análises das bulas dos agrotóxicos, foco de trabalho de mestrado associado a este [TEIXEIRA, 2018]. Como resultado desse enriquecimento, foi possível ser feita uma análise das propriedades dos agrotóxicos, em quais países estes são aplicados, quais bulas os referenciam e quais são os efeitos e sintomas causados pelos mesmos.

O resultado gerado a partir deste estudo propõe-se a abrir novos horizontes para a observação das demandas prejudiciais do uso de agrotóxicos e em quanto o Brasil pode estar posicionado de maneira retrógrada nas questões que tangem à saúde humana.

1.3 METODOLOGIA

A metodologia de desenvolvimento desta pesquisa foi constituída nos seguintes estágios:

1) Estudo dos conceitos de Web Semântica e das ferramentas que poderiam apoiar o desenvolvimento desta pesquisa.
2) Estudo dos conceitos a respeito dos agrotóxicos e das bases de dados relacionados acerca do tema.
3) Seleção das bases de dados relevantes para trabalhos anteriores e obtenção e limpeza dos dados disponíveis.
4) Modelagem da semântica a ser usada, com base nas ontologias existentes sobre o tema.
5) Triplificação dos dados experimentando ferramentas já existentes no mercado e avaliação das mesmas.
6) Armazenamento e exposição das triplas em RDF geradas a partir das bases escolhidas.
7) Validação da transformação dos dados com base em consultas e visualização das triplas.

1.4 ESTRUTURA DO TRABALHO

No Capítulo 1 é apresentada esta introdução incluindo motivação, objetivo, justificativa, metodologia e estrutura da pesquisa realizada.

No Capítulo 2 são abordados os conceitos de Web Semântica, *Resource Description Framework*, dados abertos conectados (*Linked Open Data*) e Vocabulário e Ontologias, com uma breve revisão de literatura sobre esses temas.

O Capítulo 3 contém a apresentação das ferramentas utilizadas, sendo expostas suas principais funcionalidades.

O Capítulo 4 é composto pela descrição do problema a ser abordado e a proposta de solução do mesmo, descrevendo as bases de dados a serem trabalhadas e suas características, além de apresentar as ontologias que serão utilizadas para a anotação de dados.

No Capítulo 5 é descrito todo o processo de triplificação e suas etapas para cada base de dados escolhida. Também são discutidas as limitações que foram descobertas na execução deste processo.

O Capítulo 6 engloba uso, visualização e exploração desses dados de maneira a responder questões de relevância, evidenciando a importância de seu tratamento como LOD e de seu uso conjunto.

O Capítulo 7 abrange as conclusões, dificuldades encontradas e sugestões de trabalhos futuro.

2 CONCEITOS DE WEB SEMÂNTICA

Inicialmente, a Web surgiu como uma maneira de compartilhar informações diversas em documentos dispostos na estrutura de hipertexto, de forma a facilitar a navegação entre estes.

Com o passar dos anos, a Web evoluiu de um meio de exibição de páginas de documentos estáticos para um espaço onde diversas aplicações podem requisitar informações que são a base para uma plataforma de execução de programas. Como diversas aplicações, cada uma com sua linguagem de programação específica, podem consumir e disponibilizar dados na Web, houve uma mistura de formato de dados oferecidos. Além disso, muitas vezes dados são dispostos em páginas da Web sem facilidades para conexão entre eles, especialmente por não serem publicados como dados "brutos", ou seja, em um grão menor, capaz de ser trabalhado com maior flexibilidade.

A partir disso, surgiu o problema dos conteúdos dessas páginas conseguirem ser vistos pelas máquinas além da forma sintática, ou seja, os dados precisariam ser interpretados e entendidos não só pelas pessoas mas também por programas. Com menos ambiguidade torna-se possível correlacionar dados, facilitando seu uso em conjunto. Com isso, será possível evoluir para um patamar onde a informação não esteja somente disponível e sem conexão alguma. Atualmente a Web é recheada de informações, mas poucas fontes são confiáveis e, quando são, exigem um grande processamento do intelecto humano.

2.1 WEB SEMÂNTICA

Em Linguística, semântica é o estudo do significado das palavras cuja finalidade é entender, por meio da língua falada, a expressão humana. Então, seria

necessário que houvesse uma semântica para computadores, o que quer dizer que seria necessária uma maneira que uma máquina entendesse, com base numa determinada informação disponível na Web, o significado pretendido e o significado entendido.

Em Maio de 2001, um artigo de Tim Berners-Lee, James Hendler e Ora Lassila foi publicado na Scientific American [BERNERS-LEE, 2001] onde foram definidas as bases da Web Semântica. De forma sintética, a Web Semântica pode ser definida como uma extensão da Web que visa facilitar o processo de entendimento e interoperabilidade das informações, dispostos de forma heterogênea através da incorporação de semântica aos dados.

De uma forma ideológica, o objetivo é criar através da Web semântica um "mapa mental" comum para facilitar o processo de desenvolvimento de aplicações que utilizem diversas fontes de dados para gerar informações conectadas e armazená-las.

A Web Semântica estabeleceu um conjunto de padrões para facilitar a identificação de recursos na Web. Este conjunto é formado por diversas camadas compondo a arquitetura da Web Semântica (Figura 1).

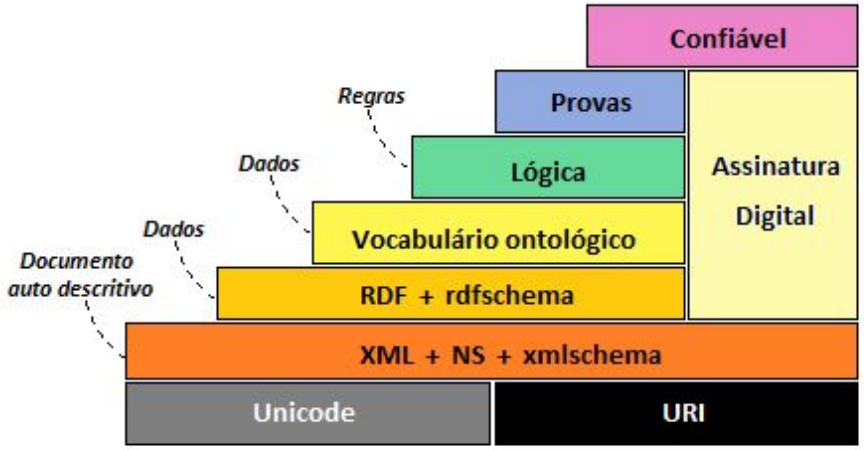


Figura 1 - As camadas da Web Semântica

Abaixo apresenta-se um breve resumo de cada camada:

- **Unicode/URI**: O unicode é um padrão de codificação universal que permite aos computadores representar e manipular texto de qualquer sistema de escrita existente. A URI é um padrão para identificar um recurso físico ou abstrato de maneira única. Tendo em vista isso, essa camada fornece a interoperabilidade em relação à codificação dos caracteres e ao endereçamento de um recurso na Web Semântica.
- **XML/Namespace/XML Schema**: XML (*Extensible Markup Language*) é uma linguagem de representação de uma infraestrutura única para diversas linguagens distintas. O XML Schema é uma linguagem para definição de regras de validação para uma classe de documentos XML. Ou seja, o XML Schema fornece elementos para descrever a estrutura e restringir o conteúdo XML. Assim sendo, esta camada fornece a interoperabilidade em relação a sintaxe de descrição de recursos da Web Semântica.
- **RDF/RDF Schema**: O *Resource Description Framework* (RDF) engloba um modelo de dados para expressar declarações sobre recursos em uma sintaxe baseada em XML e uma linguagem de definição de esquemas para vocabulário. O RDF Schema é uma linguagem para a definição de esquemas para os vocabulários utilizados para a declarações das triplas RDF. Desta forma, essa camada fornece um modelo para representar os metadados sobre os próprios recursos.
- **Ontologia**: É uma forma de ampliar o RDF, de modo que mais restrições sejam impostas às triplas definidas no RDF, limitando assim o número de interpretações possíveis para o modelo apresentado no RDF e, por consequência, reduzindo a ambiguidade, pois esta especifica os conceitos dentro de um domínio. Uma poderosa linguagem que descreve ontologias é a OWL (*Ontology Web Language*), que permite o uso de "dedução lógica para inferir informações adicionais dos fatos expostos explicitamente em uma ontologia", segundo o Guia da Web Semântica [NICBR, 2015]. A OWL facilita

o processamento do conteúdo da informação por meio de programas, uma vez que ela fornece vocabulário com uma semântica formal.

- **Lógica**: Esta camada fornece suporte para a descrição de regras para expressar relações sobre os conceitos de uma ontologia, as quais não podem ser expressas com a linguagem de ontologia utilizada.
- **Prova e Confiança**: Camadas que proporcionam a assistência para execução das regras, além de avaliar a correção e a confiabilidade dessa execução. Estas camadas ainda estão sendo desenvolvidas e dependem da maturidade das camadas anteriores.

2.2 RESOURCE DESCRIPTION FRAMEWORK

Tendo em vista o antigo padrão de exibição de informações na Web construído para entendimento e consumo humano, há outro problema: como as aplicações poderão processar os dados dispostos em forma de documentos para que sejam agregadas informações específicas garantindo a interoperabilidade entre os formatos?

Uma solução simples é o uso de metadados para descrever os dados já publicados, ou seja, o uso de dados sobre os dados. Com informações relacionadas aos dados, como descrições de estrutura, conteúdo e outras, seria possível apoiar o cruzamento dos diferentes metadados para um conjunto de vocabulários comuns.

Para representar as informações na Web de forma a fazer afirmações sobre recursos, foi criada uma estrutura, o *Resource Description Framework* ou RDF. Um recurso pode ser definido como qualquer coisa, tanto concreta como abstrata, por exemplo: uma pessoa, um objeto, um sentimento, uma cor, etc. Uma afirmação RDF, também conhecida como tripla, é constituída de três elementos com o seguinte formato: <sujeito> <predicado> <objeto> e tem como objetivo geral expressar uma relação entre dois recursos, o sujeito e o objeto. O predicado é a

natureza da relação entre os dois recursos que estão sendo relacionados na tripla RDF.

Um conjunto de triplas, ou declarações, é chamado de grafo RDF, que pode ser ilustrado como um diagrama de nós e arcos orientados, no qual cada tripla é representada como uma ligação nó-arco-nó.

Para elucidar, pode-se tomar como exemplo o caso em que se deseja descrever a instância de um Pesticida, como o Carbamato, que pertence à classe de Inseticidas. Em RDF, tal afirmação poderia ser modelada na seguinte tripla:

<http://purl.obolibrary.org/obo/CHEBI_38461>
is_a
<http://purl.obolibrary.org/obo/CHEBI_25944>.

Nesta, http://purl.obolibrary.org/obo/CHEBI_38461 é a URI identificadora do inseticida Carbamato e http://purl.obolibrary.org/obo/CHEBI_25944 é a URI identificadora de Pesticida. O predicado *is_a* é usado como definição, sendo assim aplicado para definir o Carbamato como um Pesticida.

Como um modelo de dados, o RDF possui algumas notações sintáticas, ou extensões, dentre as quais as mais utilizadas são: Turtle, RDF/XML e Ntriples. A primeira notação padronizada pelo W3C foi a RDF/XML, pois XML é uma notação conhecida e amplamente suportada pelas linguagens de programação. Entretanto, o uso da mesma traz como desvantagem a dificuldade de ser entendida por humanos, e por esta razão, Turtle passou a ser amplamente utilizada, uma vez que a mesma é de fácil entendimento humano. Por esta razão, neste trabalho foi escolhida a notação Turtle. A notação N-Triples, por sua vez, muito se assemelha à notação Turtle, exceto que a N-triples é um subconjunto da notação Turtle [N-TRIPLES].

“As triplas de N-triples são também triplas simples no formato Turtle, mas Turtle inclui outras representações de termos RDF e abreviações de triplas RDF. Por isso, quando analisada por um analisador Turtle, os dados no formato N-Triples

produzirão exatamente as mesmas triplas que um analisador para N-triples. O gráfico RDF representado por um documento N-Triples contém exatamente cada tripla correspondente à produção de uma tripla N-Triples", segundo a Recomendação da W3C [W3C, 2014] .

A seguir, o exemplo onde uma tripla gerada neste trabalho é exibida em cada um dos formatos citados. As triplas Turtle e N-triples são iguais pois uma é subconjunto da outra, entretanto, quando se deseja utilizar recursos mais avançados no arquivo RDF, o formato Turtle é que possui mais abrangência.

- Turtle:

```
<http://aims.fao.org/aos/agrovoc/c_31235>
<http://lodbr.ufrj.br/agrotoxicos/propriedade/proibidoEm>
<http://dbpedia.org/page/Austria>.
```

- N-Triples:

```
<http://aims.fao.org/aos/agrovoc/c_31235>
<http://lodbr.ufrj.br/agrotoxicos/propriedade/proibidoEm>
<http://dbpedia.org/page/Austria>.
```

- RDF/XML:

```
<rdf:Description rdf:about="http://aims.fao.org/aos/agrovoc/c_31235">
  <ex:proibidoEm>
    <rdf:Description rdf:about="http://dbpedia.org/page/Austria">
    </rdf:Description>
  </ex:proibidoEm>
</rdf:Description>
```

2.3 LINKED OPEN DATA

Com todas as informações conectadas seria possível gerar um imenso banco de dados global que possa ser consumido por diferentes aplicações. Para isso, é

indispensável que haja uma forma padrão que estabeleça a conexão destes dados de maneira alcançável e manipulável pelas ferramentas de Web Semântica. E não apenas isso, que também seja possível a adição de mais conexões entre diferentes dados. Essa coleção de conjuntos de dados interligados pode ser chamada de dados conectados, ou linked data.

Para melhor organização que favoreça a conexão dos dados, Tim Berners-Lee definiu 4 princípios para a conexão dos dados abertos [BERNERS-LEE, 2001]. São eles:

1. Utilize URI como nome dos recursos

URI - Uniform Resource Identifier, é um tipo de URL mais abrangente que permite identificar recursos, enquanto uma URL identifica a localização de uma informação. Um efeito causado por algum agente, como pesticidas, por exemplo, pode ser identificado por uma URI no seguinte formato: http://lodbr.ufrj.br/efeito/Hipertermia. É de extrema importância que essas URIs sejam definidas, garantindo que sejam persistentes ou perdurem pelo maior tempo possível, visto que é altamente custoso renomear uma URI caso ela se torne obsoleta.

2. Utilize HTTP URIs para que as pessoas procurem estes recursos.

É fundamental que as URIs em questão estejam disponíveis na Web, para que outros trabalhos possam referenciá-las.

3. Disponibilize informações relevantes nas URIs, utilizando os padrões (RDF, SPARQL)

Ao requisitar uma URI, é esperado que o servidor retorne informações relevantes acerca do recurso identificado pela URI. Por isso, ao criar uma URI, é imprescindível que haja de fato uma descrição sobre o recurso.

4. Inclua links para outras URIs para que se descubra mais informações

Para evitar redundância de informações, e em muitos casos, até mesmo ambiguidades, bem como não sobrecarregar a Web de informações desnecessárias

e poupar retrabalho, recomenda-se que sejam referenciadas outras URIs já existentes.

Além de todas as recomendações citadas anteriormente, com toda essa nova visão semântica da Web de Dados, os mais diversos tipos de informações são publicados em diferentes formatos e estruturas de armazenamento. Para orientar a publicação destes dados abertos na Web tendo como objetivo minimizar a confusão de dados, Tim Berners-Lee sugeriu um esquema de classificação de dados conectados de maneira evolutiva [BERNERS-LEE, 2001]. A escala dos cinco níveis de evolução dos dados é:

1. Sua informação na Web de qualquer formato deve ser publicada sob algum tipo de licença de dados abertos

Desta forma, será oficial e formalmente permitido o uso, redistribuição, modificação e propagação dos dados, o que gera uma contribuição concreta com a comunidade de dados abertos.

2. Esta informação deve ser publicada na forma de dados estruturados , como por exemplo, uma planilha excel no lugar de uma imagem png.

A vantagem do uso de dados estruturados é tê-los legíveis por máquinas, de modo que, por meio de um processo ou programa, é possível converter estes dados para outros formatos, de acordo com a conveniência e necessidade. Em particular, para o processo de triplificação de dados, de modo a contribuir para a Web Semântica, é fundamental que o formato original dos dados seja estruturado, pois só assim é possível utilizar alguma ferramenta para executar o processo. Caso os dados originais não sejam estruturados, se for desejado o uso de uma ferramenta de triplificação, é preciso, inicialmente, estruturá-los, para então triplificá-los.

3. Use formatos não proprietários - CSV ao invés de excel, por exemplo.

A manipulação dos dados de qualquer maneira que o consumidor queira é possível devido ao não confinamento das restrições do software específico.

4. Use URIs para identificar esses seus recursos de forma que as pessoas possam encontrar e apontar para sua informação.

Assim, tem-se dados disponíveis na Web através de uma URI e, desta forma, o consumidor pode apontar para estes dados de qualquer outro lugar e reusá-los, seja parte do dado ou como o todo. Uma representação nativa de dados é utilizar o RDF, porém outros formatos podem ser convertidos e mapeados caso seja necessário.

5. Os seus dados devem estar conectados com outros dados de maneira a oferecer um contexto das informações de forma interligada.

Um dado seguindo todas essas definições está disponível na Web e ligado a outros dados de maneira que se forma uma rede onde tanto consumidor quanto publicador podem se beneficiar. Estes cinco pontos visam otimizar a descrição semântica dos dados, assegurando simplicidade no procedimento, na conversão de dados e no descobrimento e consumo simultâneo dos mesmos.

2.4 VOCABULÁRIOS E ONTOLOGIAS

Apesar de ter sido definido até então o processo de construção de dados triplificados, o mesmo muito se assemelha a um novo processo de construção de dados estruturados, apenas seguindo uma nova estrutura. Sabendo que o grande diferencial da Web Semântica é justamente atribuir significado aos dados, qual seria então o método para tal? A utilização de vocabulários e ontologias.

No dicionário da língua portuguesa, um *vocabulário* é definido como "conjunto dos vocábulos de uma língua", sendo o termo *vocábulo* definido como "unidade pertencente, unidade mínima" [Dicionário Online de Português DICIO]. Traçando um paralelo entre a Web Semântica e a língua, é possível perceber o porquê da escolha da palavra *vocabulário* para o contexto de Linked Data, pois aqui, os vocabulários são justamente o conjunto de informações, ou unidades mínimas, responsáveis por descrever os termos referenciados pelas triplas RDF.

É importante ressaltar, entretanto, a diferença entre vocabulários e ontologias, pois vocabulários apenas descrevem termos, enquanto, conforme descrito na seção 2.1 deste capítulo, ontologias fornecem conhecimento adicional a ser inferido acerca da informação, fazendo uso de vocabulários e semântica formal e usando lógica para as inferências.

Para a manutenção da efetividade e facilidade de relacionar dados, é sempre melhor, caso já existam ontologias na Web que sejam úteis para a situação que se deseja modelar, utilizá-las, no lugar de construir uma nova. Visando também minimizar o retrabalho, essa é a forma mais eficiente, pois construir uma ontologia não é um trabalho trivial.

Os modelos de ontologias implementadas atualmente compartilham de componentes estruturais independentemente do contexto e da linguagem em que são expressas. Esses componentes são divididos em Classes, Instâncias, Axiomas, e Relacionamentos.

O componente classes são grupos abstratos, conjuntos ou coleções de objetos que podem conter indivíduos ou outras classes. Também pode ser interpretado como os elementos representando conceitos do domínio da ontologia. São alguns exemplos de classes: Pessoa, Molécula, Número, Carro, etc. Neste trabalho, pode-se citar uma classe como Agrotóxico.

As instâncias são os indivíduos descritos através da ontologia [NICBR, GUIA DA WEB SEMÂNTICA], ou seja, o próprio dado da ontologia. Retomando o exemplo anterior, pode-se citar, na classe Agrotóxico, uma de suas instâncias como sendo o Acefato.

O componente de relacionamentos é um atributo cujo valor é outro objeto dentro da ontologia, ou seja, um relacionamento descreve a relação entre objetos. O conjunto de todas as relações descreve a semântica do domínio. Para ilustrar estas relações pode-se usar o seguinte exemplo: o objeto cuja instância de Agrotóxico possui o nome de Acefato pode se relacionar de forma a causar um efeito colateral como a náusea. Neste caso, a náusea é simplesmente uma instância de outra

classe Efeito Colateral. Os objetos Acefato e náusea se ligam numa relação de causa e efeito, ou seja, pode-se admitir que o objeto Acefato causa náusea.

Por fim, os axiomas são utilizados para modelar sentenças consideradas sempre verdadeiras ou restrições e regras inerentes às instâncias.

Além dos componentes, as ontologias também se classificam quanto à sua função em quatro categorias: Topo, Domínio, Tarefas e Aplicação [Guizzardi, 2005].

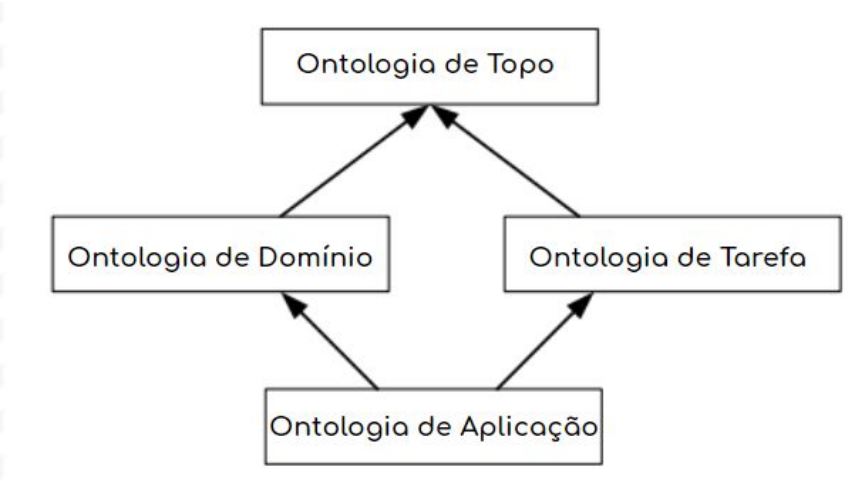


Figura 2 - Hierarquia das ontologias (Guizzardi, 2005, pg. 67)

Ontologias de topo são consideradas ontologias gerais que descrevem conceitos amplos, como elementos da natureza, espaço, tempo, coisas, estados, eventos, processos ou ações. As ontologias de domínio descrevem conceitos e vocabulários relacionados a domínios particulares como medicina e computação e ontologias de tarefas descrevem tarefas ou atividades genéricas que podem vir a contribuir na solução de problemas seja num processo de vendas ou em algum diagnóstico.

Por fim, ontologias de aplicação descrevem conceitos que dependem de um domínio particular e de uma tarefa específica, como mostra a Figura 2.

3 CICLO DE VIDA DOS DADOS CONECTADOS

Triplificar dados está longe de ser uma tarefa trivial como a maioria das conversões de dados existentes. Diferente de dados estruturados como SQL ou mesmo dados não estruturados como os de formato de extensão JSON, triplas RDF existem para agregar conhecimento aos dados, por isso, é um processo mais delicado construí-las.

Criar dados conectados pode ser descrito, de maneira geral, por uma sequência de ações constituída por 5 etapas, conforme descrito em [EQUIPE GT, *LinkedDataBR - Manual do Usuário,* 2011]: (i) identificação e seleção de dados, (ii) limpeza, anotação e transformação, (iii) mapeamento, (iv) interligação e (v) armazenamento e publicação. Embora essas etapas integrem o fluxo de triplificação de dados, algumas partes não são obrigatórias e cada etapa pode ser realizada de maneira bem particular e distinta dependendo do conjunto de dados a serem triplificados e da natureza das informações coletadas.

Estes passos para criação de LOD serão detalhados nas seções que seguem.

3.1 IDENTIFICAÇÃO E SELEÇÃO DE DADOS

É necessário identificar as fontes de dados que sejam relevantes ao tema a ser abordado no processo de criação dos dados conectados, considerando as perguntas que se pretende responder com os mesmos. Geralmente, as bases de dados são disponibilizadas de forma a facilitar a leitura e compreensão humana gerando, por consequência, uma maior dedicação na fase posterior (Limpeza, anotação e transformação).

Os dados são dispersos, usualmente, em formato XML, XLSX ou CSV, sendo indispensável no estudo dos dados selecionar quais são os atributos que farão parte

da triplificação. Além disso, é importante identificar as fontes de dados e sua confiabilidade para que as questões respondidas sejam fidedignas à realidade.

Esta etapa do ciclo de vida é descrita em detalhes na seção 4.2.1 do Capítulo 4 para as bases de dados deste trabalho.

3.2 Limpeza, anotação e transformação

Em muitas circunstâncias, a base de dados original não se encontra com uma estrutura razoável para passar por quaisquer processos automatizados. Esta etapa, portanto, existe para garantir que os dados sejam coerentes e sigam uma mesma estrutura que possibilite uma correlação automatizada com as URIs e predicados da modelagem construída.

Se, por exemplo, existir mais de uma label em uma mesma célula do dado CSV, isto potencialmente dificulta a criação de triplas, uma vez que cada dado deve ser associado a uma única URI. Logo, é de extrema importância que cada célula descreva um único dado. Por consequência, estes devem ser separados em linhas diferentes para passar por um processo automático de conversão de dados, seja qual for o processo - os possíveis processos de conversão a serem usados serão discutidos mais adiante neste mesmo capítulo.

Outro caso muito comum de limpeza de dados em se tratando de língua portuguesa, é a necessidade de remoção de acentuação e caracteres especiais.

Existem diversas opções para fazer a limpeza dos dados, desde manualmente utilizando algum programa que leia a extensão de dados alvo, como é o caso do Airtable ou utilizar algum programa que auxilie a limpeza de dados, como é o caso do Open Refine, ambas ferramentas descritas a seguir neste capítulo. Outra opção seria escrever um programa que realize a limpeza desejada.

3.3 MAPEAMENTO

Como triplas RDF não são simplesmente uma forma diferente de estruturar dados, não basta escrever um programa que mapeie e converta de uma estrutura para a outra. É necessário construir uma modelagem prévia bem clara e buscar nas ontologias termos que representam os dados presentes na base de dados que se deseja triplificar.

A modelagem dos dados consiste em entender que tipos de informações a base de dados em questão contém e como associá-las por meio de predicados, e, a partir disso, decidir que triplas deseja-se gerar. Sendo assim, é necessário, para cada grupo de dados, buscar ontologias que classifiquem os dados que existem na base original e que contenham predicados que descrevem a relação entre estes dados.

Em geral, esses grupos de dados consistem em uma coluna, para o caso de dados CSV, mas muitas vezes, para o caso de bases de dados que não misturam assuntos e de porte menor, uma única ontologia modela quase todas ou todas as colunas de uma mesma base de dados. Já no caso de bases de dados de volume maior e que integram mais de um assunto, é provável que a modelagem demande mais de uma ontologia.

Os predicados, ou seja, relações entre os dados, em geral, também se encontram na ontologia que possui as classificações para estes dados. Caso os predicados desejados, ou mesmo as classificações, se encontrem em mais de uma ontologia, a triplificação deverá então usar URIs e predicados de todas as ontologias necessárias, criando uma relação entre elas. Não é problema algum fazer tal uso, pelo contrário, isso vai ao encontro ao objetivo da Web Semântica, criando cada vez mais conexões e significados aos dados.

De maneira semelhante, para o caso em que os dados não encontrem correspondência em uma ontologia pré-existente, é possível criar URIs que não

existem e construir triplas com estas, mesmo sem construir uma ontologia ou disponibilizar informações na Web sobre as URIs criadas. Não é a circunstância ideal, pois sem a ontologia associada ou dados relacionados disponíveis na Web não existem dados agregados a essa URI, reduzindo assim a riqueza de informações nas triplas RDF geradas já que estas URIs não disponibilizam maiores informações do dado que elas descrevem, ferindo assim a quarta estrela de classificação de Tim Berners-Lee [BERNERS-LEE, 2001].

Feito isso, recomenda-se criar uma modelagem com um exemplo ou molde de cada tipo de tripla que se deseja gerar, isto é, para cada coluna, criar uma associação dela com outra coluna, para entender todas as triplas a serem geradas e também o que cada coluna modela dentro das ontologias escolhidas.

A seguir são descritas ferramentas que auxiliam - também - nesta etapa, criando um modelo que descreve a base de dados associada às URIs e propriedades finais para, em seguida, gerar triplas RDF.

3.4 INTERLIGAÇÃO

Após a escolha das fontes de dados, refinamento e limpeza das mesmas e mapeamento das ontologias a serem usadas na modelagem prevista para disponibilização dos dados, deve ser feita a etapa de interligação. Com os dados devidamente tratados para serem triplificados, é necessário usar URIs, preferencialmente disponíveis na Web, conforme explicado na seção 3.3 deste capítulo, para cada atributo dos dados escolhidos, com o intuito de disponibilizá-los online.

Cada dado da base de dados deve se associar a uma URI de forma a garantir que a base de dados possa ser conectada com outras já disponíveis na Web. Há diversas ontologias que oferecem URIs já mapeadas com informações acerca do dado em questão, então, é recomendável o uso dessas informações já existentes.

Esta etapa é uma etapa chave do processo de triplificação, pois aqui será aplicada a modelagem prevista em etapa anterior. Após o mapeamento de todos os dados contidos na base de dados para URIs, as mesmas devem ser conectadas por meio de uma propriedade, de modo a formar uma tripla, e é justamente o conjunto dessas triplas que irá formar o resultado final da base de dados convertida para RDF.

Existem algumas ferramentas disponíveis que facilitam o processo de interligação dos dados, como é o caso do Karma, já descrito na seção 3.2.2 deste capítulo, que além de auxiliar no processo de modelagem, possui um algoritmo capaz de sugerir URIs a serem associadas com células da base de dados nele importada. Também é possível gerar a interligação por meio de uma extensão RDF disponível para o Open Refine, descrito na seção 3.3.1 deste capítulo. Uma terceira opção, utilizada neste trabalho, consiste em combinar o Airtable, descrito na seção 3.2.1 deste capítulo, que fica responsável por mapear as células para URIs por meio de fórmulas, com o SQL Developer, descrito a seguir neste capítulo, que é responsável por concatenar URIs em pares, obtidas através de resultados de consultas SQL na base de dados, por meio de propriedades, formando assim, as triplas desejadas.

3.5 ARMAZENAMENTO E PUBLICAÇÃO

Esta é a etapa final do processo de triplificação de dados. Uma vez que os dados foram convertidos, é necessário publicá-los na Web, de modo que sejam disponibilizados para serem utilizados e integrados com outros dados já presentes na Web. O ideal a ser alcançado nesta fase é que os dados estejam disponíveis para download em páginas Web. Além disso, se for o caso de URIs terem sido criadas para alguns destes dados, como foi o caso deste trabalho, é desejável que estas URIs estejam hospedadas em páginas Web que contenham informações

sobre os mesmos, como é o caso, por exemplo, desta URI utilizada em triplas geradas neste trabalho:

http://purl.bioontology.org/ontology/MEDDRA/10000424

Esta URI, quando acessada por um navegador, disponibiliza informações acerca do dado nela mapeado, como sua label, seu id e URIs às quais ele é associado, por exemplo. Deve ser de modo semelhante para cada URI criada.

A etapa de publicação dos dados das URIs criadas está fora do objetivo deste trabalho, podendo ser realizada posteriormente por outros trabalhos. Entretanto, a etapa de publicação dos dados para download foi feita, e os dados estão disponíveis nesta URL: https://github.com/amandamedeiros94/tcc-agrotoxicos. Além disso, como forma de validar e emular a exibição dos dados, foi usado o GraphDB após a triplificação dos dados deste trabalho. A utilidade desta ferramenta será descrita na próxima seção.

3.6 DESCRIÇÃO DAS FERRAMENTAS UTILIZADAS NAS ETAPAS DO CICLO

Em cada etapa do ciclo de vida dos dados conectados podem ser usadas ferramentas para auxiliar no processo. Neste capítulo serão descritas as ferramentas que foram utilizadas neste trabalho para experimentar e de fato executar o processo de triplificação. As ferramentas aqui descritas podem ser utilizadas em apenas uma etapa ou em mais de uma etapa do ciclo a fim de atingir o objetivo deste trabalho, muito embora o propósito do desenvolvimento das mesmas não tenha sido especificamente ou exclusivamente feito para ser utilizada em nenhuma etapa do ciclo.

3.6.1 Airtable

O Airtable é um serviço colaborativo em nuvem com recursos de um banco de dados aplicados a uma planilha. A proposta dos fundadores do software foi democratizar a ferramenta, permitindo que qualquer pessoa possa utilizá-la sem burocracias de forma que a plataforma atenda às suas necessidades.

A proposta é que os usuários possam criar um banco de dados, configurar tipos de colunas, adicionar registros, vincular mais de uma tabela entre si, classificar registros e publicar as visualizações em sites externos, tudo isso de forma colaborativa (Figura 3).

O aplicativo está disponível para computadores e dispositivos móveis, onde as alterações são sincronizadas de forma instantânea nos dispositivos de todos os colaboradores da base de dados em questão. Para permitir uma maior produtividade do time, pode-se adicionar novos colaboradores às tabelas com diferentes níveis de permissão como *Creator*, *Edit Only* e *Read Only*. Além disso, há também a opção para download caso seja requerido um trabalho offline.

As bases podem ser facilmente compartilhadas com o público, através de uma funcionalidade chamada *Airtable Views*. Essas views podem ser adicionadas em um Website e permitem que os usuários, ao acessarem esse site, possam ter acesso às informações da planilha em tempo real. Um possível uso é um formulário para ser realizado o levantamento de feedbacks de um produto.

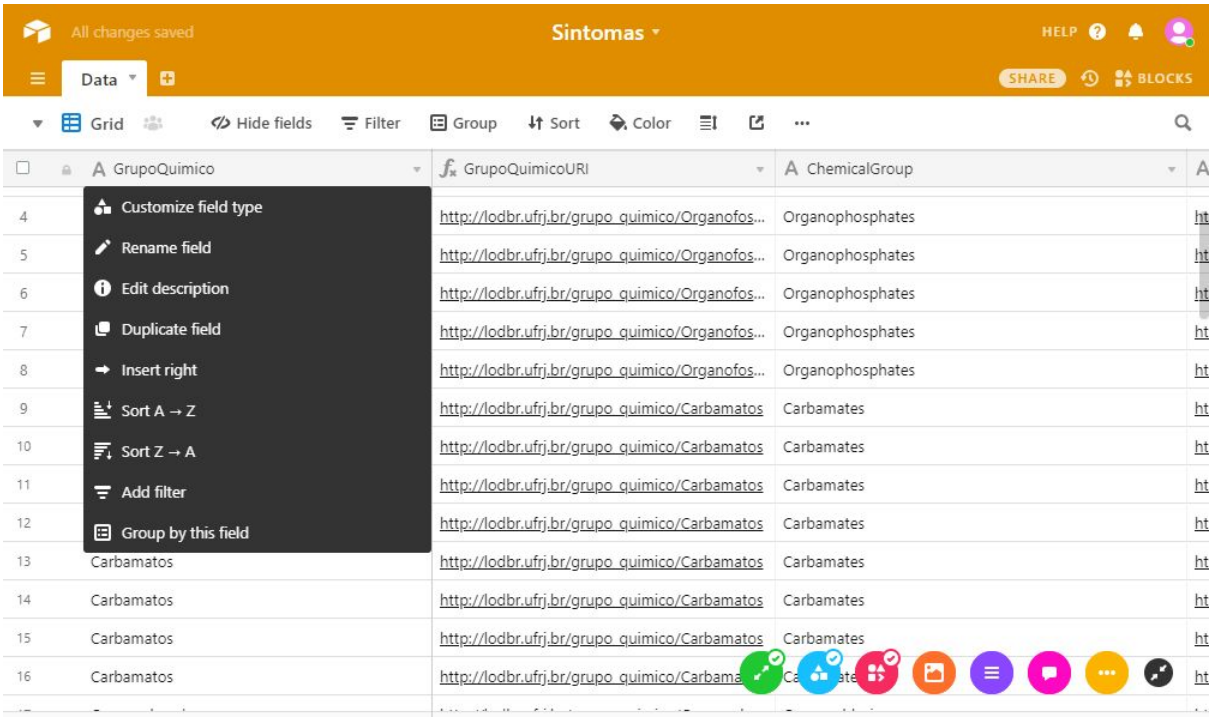


Figura 3 - Base de dados importada na plataforma Airtable

Os campos de uma tabela Airtable são semelhantes a células de uma planilha comum, porém existem outros tipos de opções de visualização e exposição do dado como *checkbox*, *phone number* e *drop-down list*. Os campos também podem manipular qualquer conteúdo que seja atribuído a eles, como anexos, texto longo, links para outras tabelas e até mesmo código de barras.

Outro recurso bastante útil disponível no Airtable são os *formula fields* (Figura 4). Trata-se de um tipo de campo atribuído à coluna, que permite a criação de colunas dinâmicas, populadas baseando-se no valor de células de outras colunas. Existe uma série de recursos associados aos *formula fields*, capazes de transformar, substituir e concatenar strings, por exemplo.

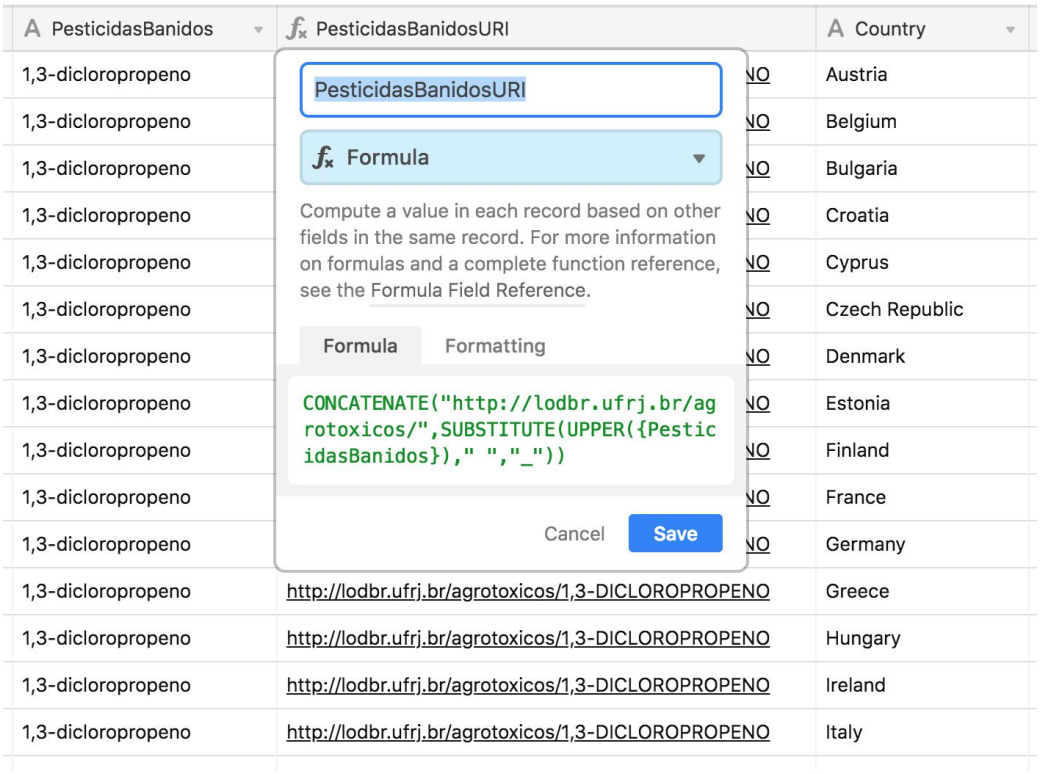


Figura 4 - Exemplo de uso do recurso formula field para criação de URIs

O Airtable também oferece uma poderosa filtragem, classificação e agrupamento de informações que fornecem maior liberdade de organização do trabalho ao gosto do usuário. Além de todas as funcionalidades citadas, ainda há a opção de conectar com outros aplicativos e serviços por meio da API Airtable que conecta-se a outros serviços Web tornando possível, assim, a troca de informações entre aplicativos externos e a planilha.

O Airtable é disponibilizado na versão gratuita e na versão paga para o uso de funcionalidades avançadas. Para fins de disseminação do conhecimento aqui descritos, neste trabalho utilizou-se sua versão livre.

3.6.2 Open Refine

O Open Refine, inicialmente conhecido como Google Refine, é uma ferramenta poderosa para trabalhar com dados, permitindo tratá-los, transformá-los de um formato para outro e estendendo-os com serviços da Web. Disponível em inglês, português, espanhol, francês, russo e em mais 9 línguas, o Open Refine é suportado pelo Google News Initiative [OPEN REFINE].

Esta ferramenta promete fornecer suporte para explorar grandes conjuntos de dados com facilidade, limpar e transformar dados e reconciliar e corresponder dados vinculando e estendendo o conjunto de dados através de requisições a serviços externos. É executada localmente, não possuindo uma versão na nuvem.

Para se iniciar um projeto com uma base de dados é necessário que esta base seja importada para o programa (Figura 5). O Open Refine entende uma variedade de formatos e arquivos de dados, usualmente detectando a extensão dos arquivos de um projeto. Também há a opção de apontar o Open Refine para uma URL de um arquivo de dados ou uma planilha do Google Sheets.

Para limpeza e lapidação de dados, o Open Refine possui uma navegação facetada de forma a prover um maior panorama dos seus dados e ao mesmo tempo filtrar apenas o subconjunto necessário de alterações.

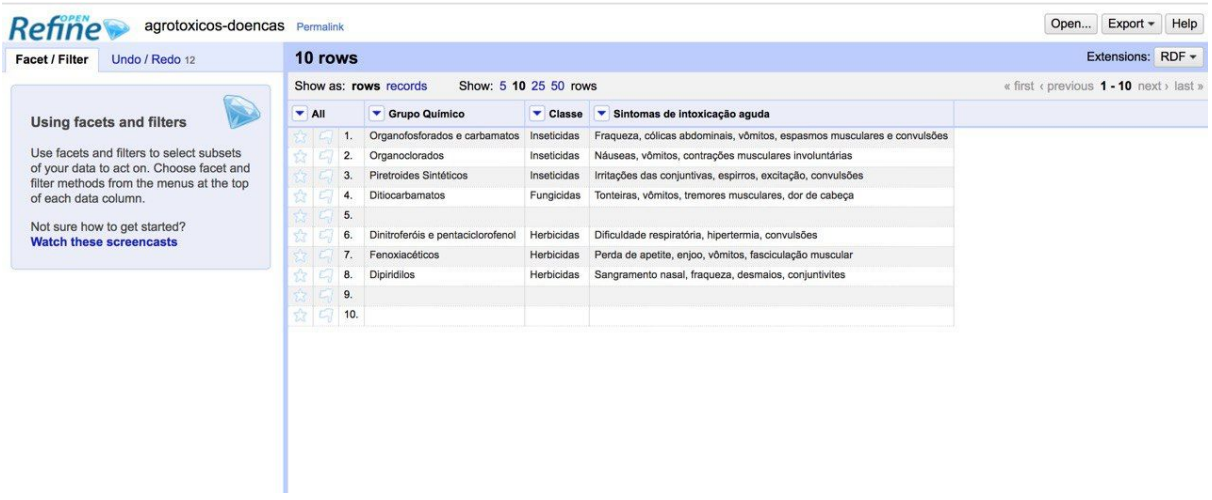


Figura 5 - Open Refine com uma base de dados simples importada

Além disso, pode-se editar células através de um mecanismo de clustering, agrupando somente células desejadas e podendo formatá-las como texto, número, data, booleano, entre outros; Editar colunas criando novos dados, através do menu drop-down ao lado do nome de cada coluna, permitindo mover colunas, concatená-las e aplicar expressões regulares; Editar linhas inteiras, removendo-as ou marcando-as como importantes.

É importante ressaltar que essa ferramenta não foi desenvolvida originalmente para triplificar dados, apenas para ser um software robusto que fornecia ricas funcionalidades para limpeza e tratamento de bases. Posteriormente foi desenvolvido uma extensão RDF com a finalidade de adicionar ao Open Refine a função de gerar triplas RDF (Figura 6).

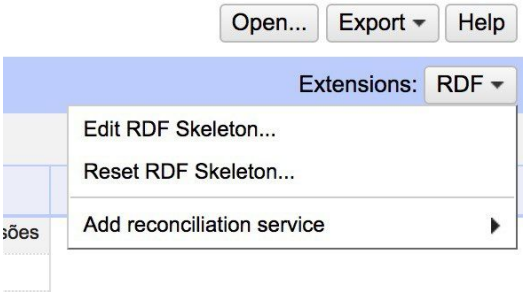


Figura 6 - Extensão RDF desenvolvida para o Open Refine

Basicamente, esta extensão permite transformar dados tabulares em RDF através da associação de cada coluna a uma propriedade ou classe de uma ontologia previamente importada para o Open Refine. Também é possível atribuir uma URI a cada coluna. É permitido caracterizar cada nó RDF de acordo com o seu conteúdo, como texto, número, data ou booleano (Figura 7).

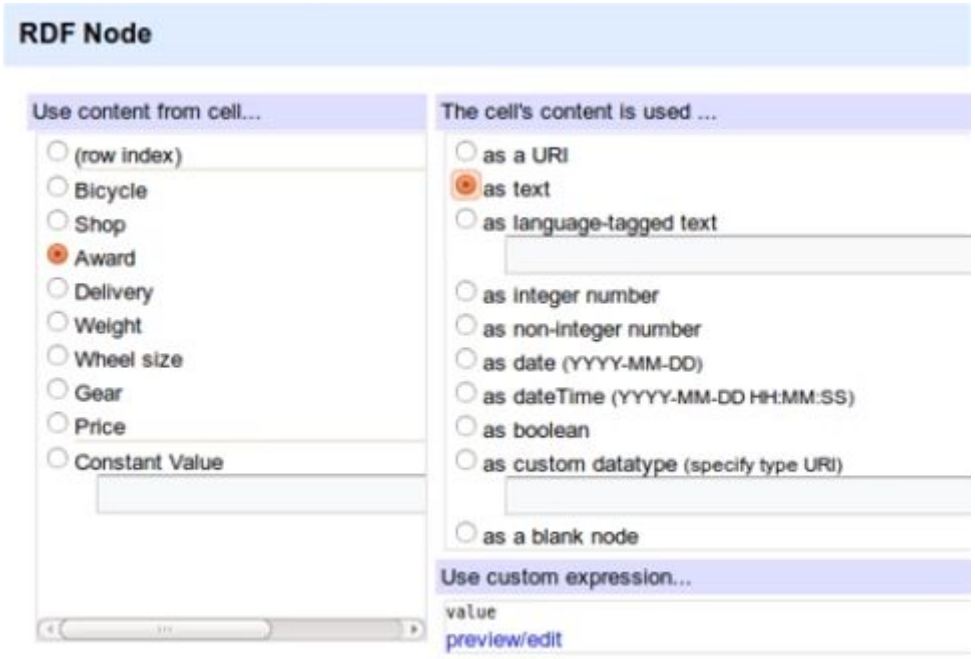


Figura 7 - Caracterização do conteúdo do nó RDF

Por fim, além de todas essas funcionalidades descritas, o Open Refine também pode exportar seus projetos nos formatos TSV, CSV, Excel e HTML Table.

3.6.3 Karma

O Karma (Figura 8) é uma ferramenta de integração de dados, resultado de um trabalho acadêmico desenvolvido na University of Southern California, com a intenção de tornar semi-automático o processo de triplificação de dados a partir de diversos formatos, entre eles CSV, JSON e XML. Diferente do Open Refine, o Karma foi desenvolvido com o intuito de triplificar bases de dados. A ferramenta oferece sugestões de modelagem, fazendo uso de modelos probabilísticos, em um ambiente visual que permite interligar dados tabulares a nós de grafos construídos visualmente na tela, com o intuito de representar a modelagem que será usada pelo programa para gerar as triplas RDF [KARMA].

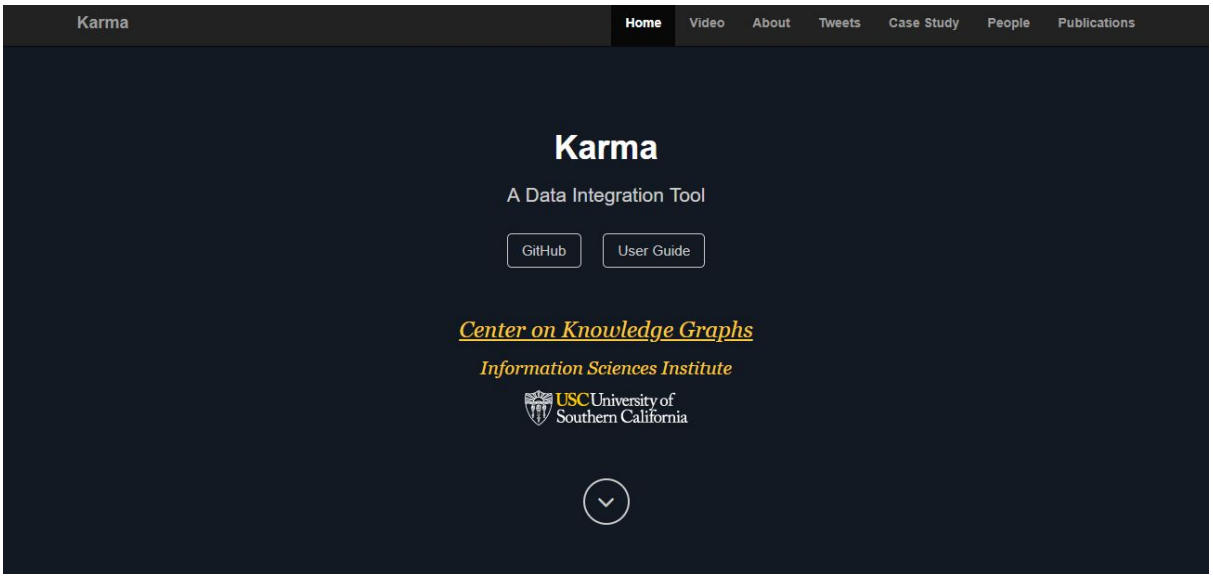


Figura 8 - Página inicial do Karma disposta no github.io

A ferramenta é descrita como fácil de usar, por fornecer técnicas de aprendizado de máquina e algoritmos de otimização de árvores para facilitar o processo de mapeamento dos dados com base em uma ontologia. Além de mapear fontes tabulares para ontologias e fontes de dados hierárquicos, como XML, JSON e KML. Ademais, o Karma oferece uma interface de programação que permite a definição de scripts para transformações de dados e permite a combinação de duas ou mais ontologias para o mapeamento destes dados para vocabulários padrão.

Além de disponibilizar o código fonte para o desenvolvimento de módulos adicionais e para o esclarecimento de dúvidas acerca do funcionamento do programa, existem, no mesmo repositório em que o código do Karma está hospedado, tutoriais completos ilustrando e explicando seu uso para a transformação de dados em triplas RDF, fazendo uso dos seus mais diversos recursos, que vão desde a transformação da base de dados original até a modelagem da mesma. Há também uma descrição para cada funcionalidade disponível nele.

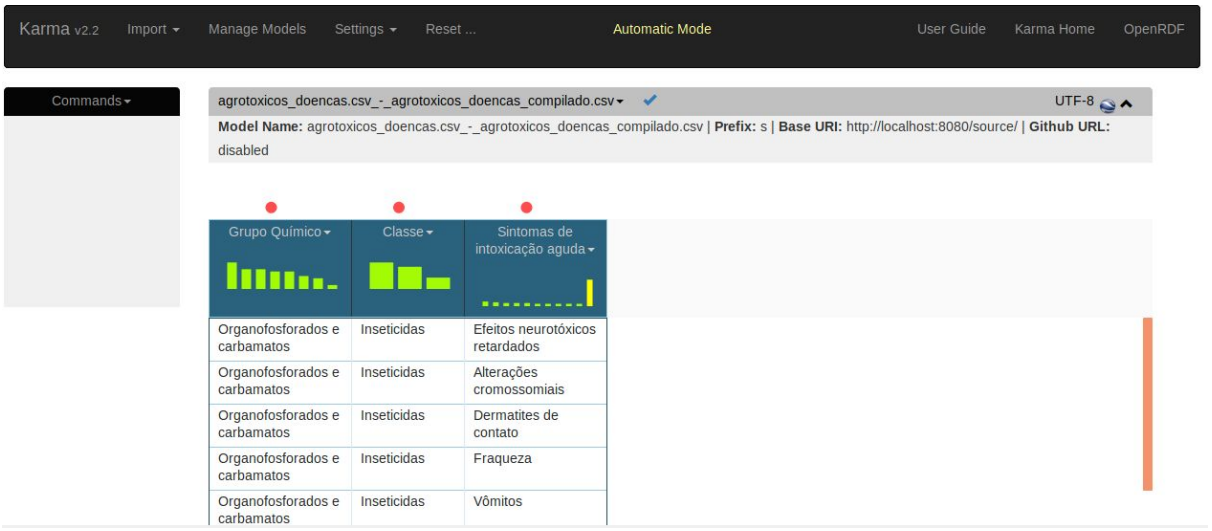


Figura 9 - Página inicial da ferramenta com uma base importada

Após a importação da base de dados (Figura 9) o Karma oferece, para limpeza de dados, funcionalidades básicas, como adicionar linhas e colunas, renomear colunas, criar novas colunas a partir das células de uma determinada coluna, dividindo os valores através de algum caracter, o que pode ser útil para o caso em que a base de dados original não está normalizada, contendo informações de entidades diferentes em uma mesma coluna. Ele também oferece algumas opções que permitem reorganizar a base de dados, como o "unfold" e o "fold", que transformam linhas em colunas e colunas em linhas, respectivamente, além da possibilidade de combinar múltiplas tabelas em uma única e de combinar um grupo de colunas em uma coluna maior, o que auxilia em casos de dados que pertencem, na modelagem, a uma mesma entidade como mostra a coluna *relatedArtworks* na Figura 10.

Além das funcionalidades básicas de transformação e agrupamento de dados, há também opções mais avançadas, que requerem algum conhecimento da linguagem Python, em especial o *PyTransform*, que permite todo tipo de manipulação de Strings suportados pela linguagem Python. Esta funcionalidade pode ser usada para gerar novas colunas com URIs compostas pelos dados existentes na tabela.

Após a importação e a limpeza dos dados, o Karma fornece o suporte para a modelagem de dados, ou seja, o mapeamento de seus dados para uma ontologia para que o Karma possa integrá-los com dados de outras fontes e publicá-los em um novo formato. O processo de modelagem do Karma consiste nas seguintes etapas: Especificação de tipos semânticos e especificação de modelos entre as classes. Na especificação de tipos semânticos é quando se define o relacionamento entre uma coluna de dados, contida na sua tabela importada, e uma propriedade ou uma classe da ontologia utilizada, como pode ser visto na Figura 10.

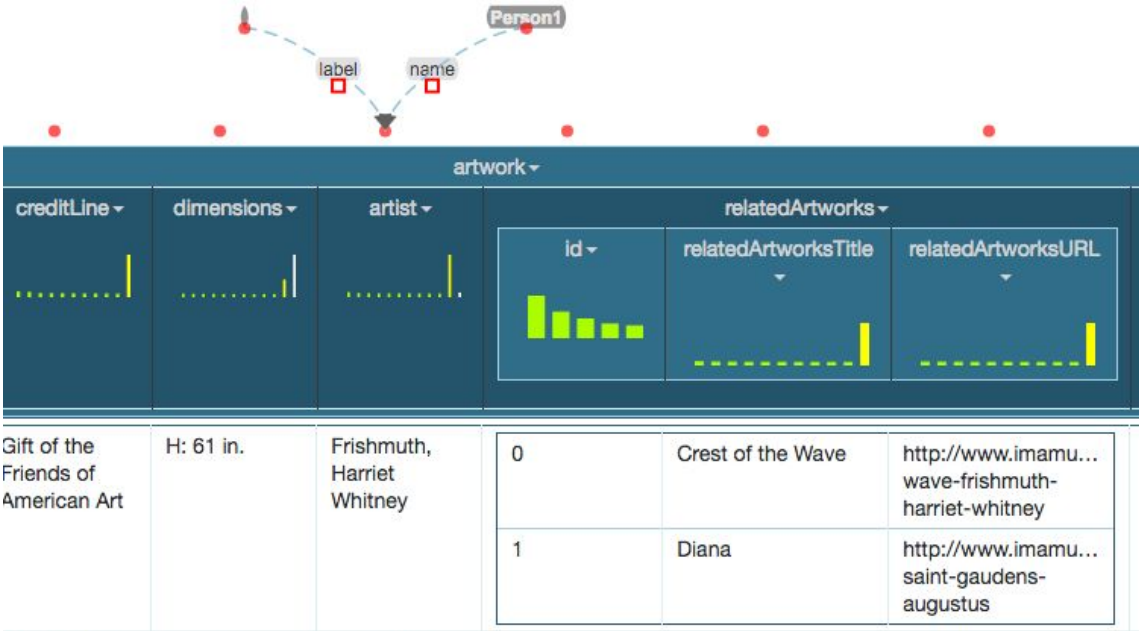


Figura 10 - Especificação de tipo semântico da coluna artista como classe Pessoa

Na especificação de relacionamento entre classes é necessário caracterizar a relação entre duas classes, ou seja, duas colunas distintas. No exemplo da Figura 11, é preciso definir o relacionamento entre Pessoa1, referente à coluna *artist*, e Pessoa2, referente à coluna sitters.

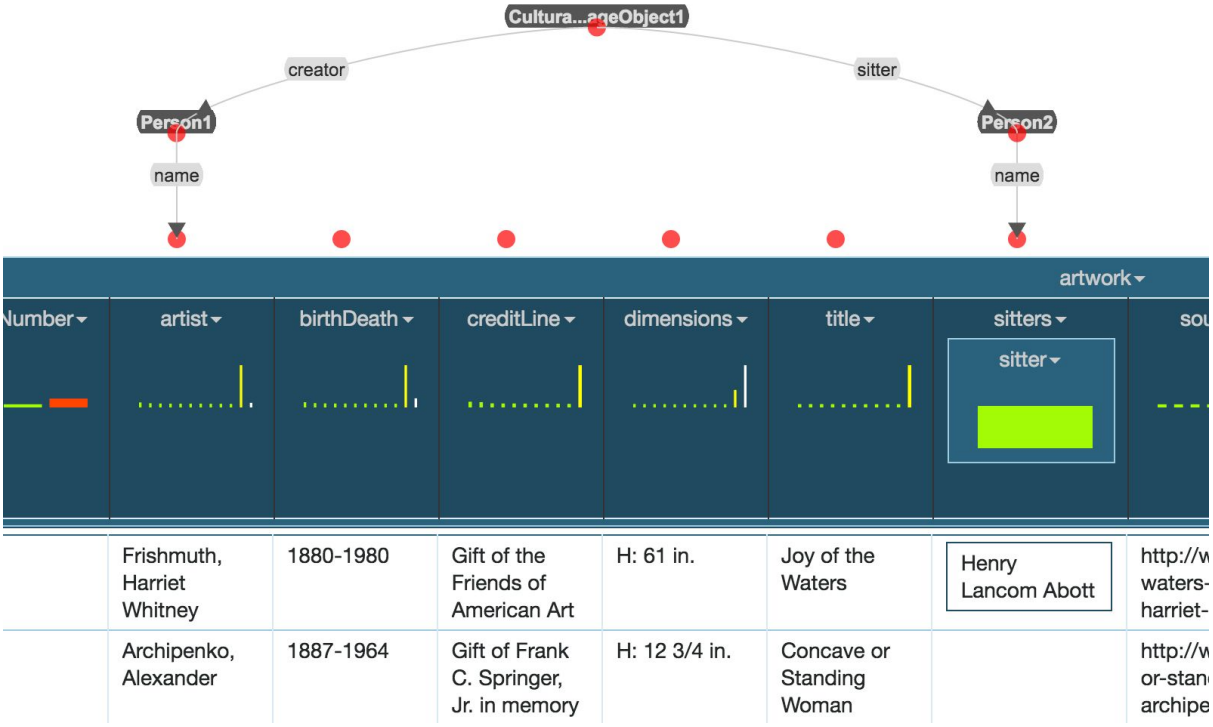


Figura 11 - Especificação de relacionamento entre as classes Pessoa

Após todas as etapas acima descritas, o Karma também fornece a opção de publicar seu modelo na linguagem R2RML, uma linguagem para expressão de mapeamentos personalizados de bases de dados relacionais em datasets RDF. Também pode-se usar o mesmo modelo gerado no modo em lotes ou no Spark para executar em um conjunto de dados maior.

3.6.4 SQL Developer

O Oracle SQL Developer é um ambiente de desenvolvimento integrado (IDE) para trabalhar com SQL em banco de dados Oracle, incorporando vários utilitários e funções diferentes em um único ambiente de desenvolvimento. As funcionalidades fornecidas por esta ferramenta incluem o gerenciamento e administração de vários banco de dados, a criação de tabelas, importação e exportação de dados.

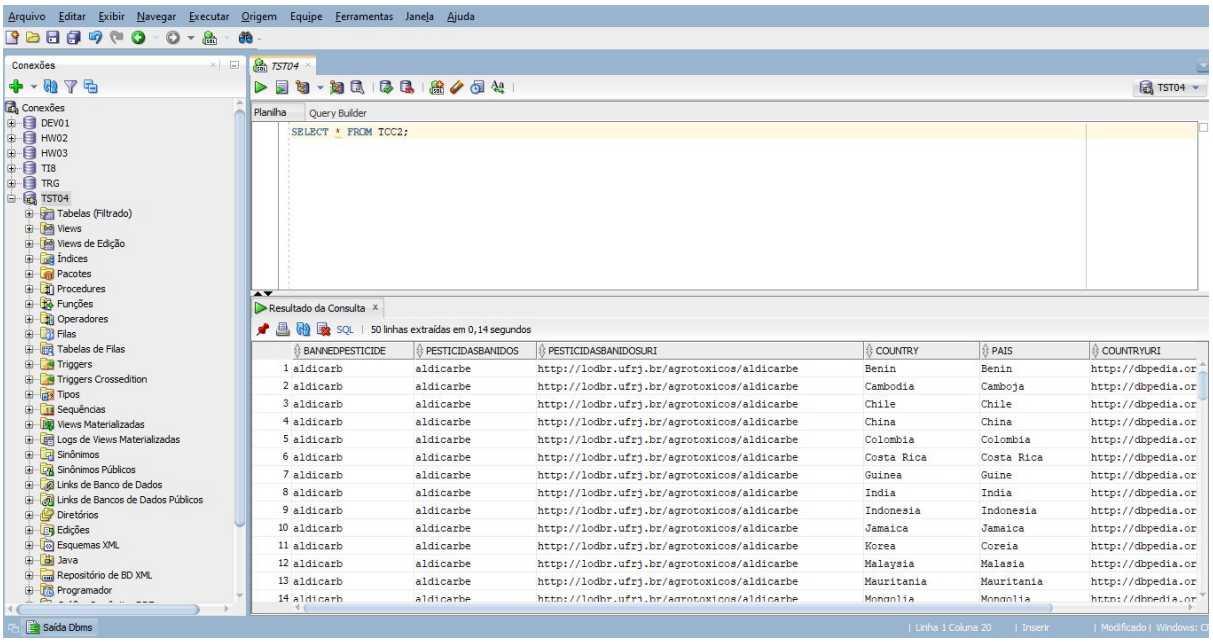


Figura 12 - SQL Developer após a execução da consulta de seleção

Uma vez que a conexão seja estabelecida com um banco de dados, no menu à esquerda é exibido todos os recursos que aquele banco oferece agrupado por pastas. Na parte central da janela localiza-se a aba de query builder (Figura 12), onde é permitida a manipulação do banco de dados através de consultas em SQL.

O SQL Developer também permite a manipulação de DBMS, Database Management System, que é um sistema que fornece aos usuários e programadores uma forma sistemática de criar, recuperar, atualizar e gerenciar dados de um banco de dados. O DBMS serve primordialmente como uma interface entre os usuários finais ou aplicativos e o banco de dados propriamente ditos, garantindo assim que os dados estejam organizados de maneira consistentes e permaneçam acessíveis facilmente (Figura 13).

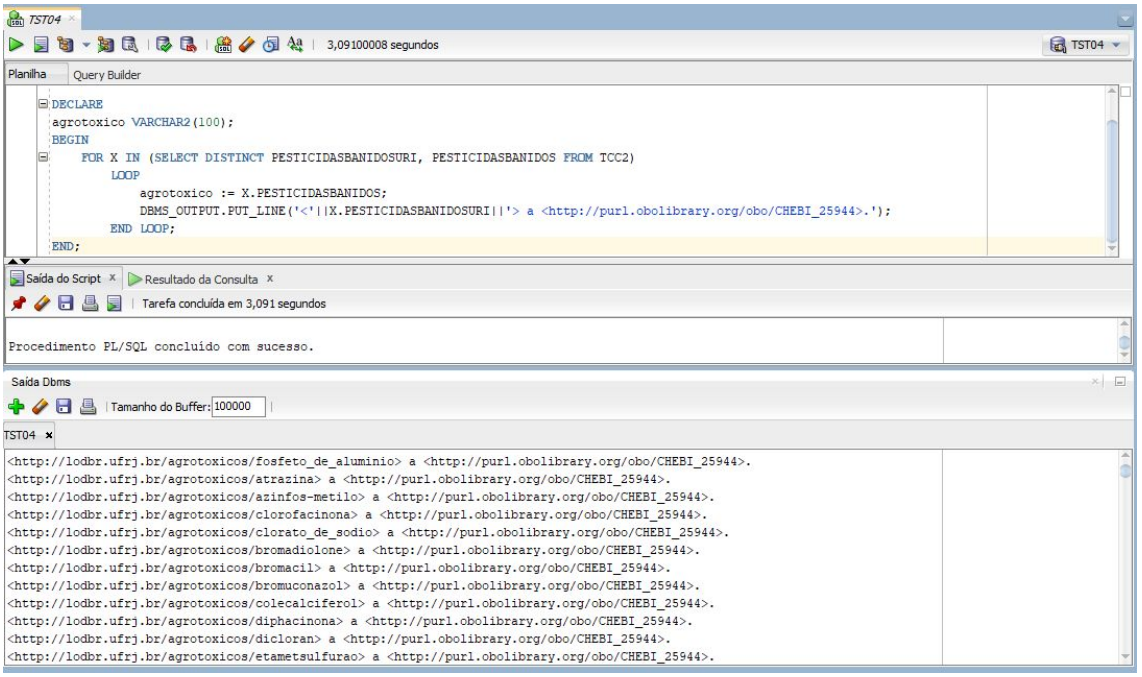


Figura 13 - Uso de DBMS no SQL Developer

3.6.5 GraphDB

GraphDB é um sistema de gerenciamento de dados baseado em estrutura de grafo que visa facilitar o armazenamento e a exploração e visualização de dados em RDF [*Ontotext GraphDB's Documentation*]. Ele implementa uma interface do RDF4J, uma framework open source desenvolvida em Java para o processamento de dados RDF, seguindo todas as recomendações da W3C acerca do SPARQL, linguagem semântica de buscas capaz de recuperar e manipular dados armazenados no formato RDF. O GraphDB é capaz de ler todas as extensões de dados RDF e oferece uma interface para explorar e lidar com buscas e inferências a grandes volumes de dados em tempo real, dividindo as bases importadas em contextos criados pelo próprio usuário, de modo a organizá-las e agrupá-las. A ferramenta está disponível em três versões: Free, Standard e Enterprise. Para fins de disseminação do conhecimento aqui descrito, neste trabalho utilizou-se sua versão livre.

Uma importante funcionalidade oferecida por esta ferramenta é a validação de arquivos RDF durante sua importação. Frequentemente, arquivos em RDF possuem diversas linhas e são de difícil visualização em editores de texto, de modo que se torna bem útil a validação automática do arquivo em questão, indicando se há algum erro sintático no arquivo que se deseja explorar, assim como a linha na qual o erro ocorre (Figura 14). Isso se torna especialmente útil se o mesmo foi construído ou convertido antes de ser utilizado.

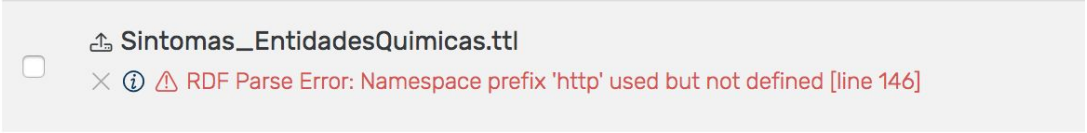


Figura 14 - Erro de análise no arquivo RDF importado no GraphDB

O GraphDB também oferece uma interface simples para realizar consultas à base de dados em RDF, através da linguagem SPARQL (Figura 15), com suporte a tudo que a linguagem oferece. O resultado das consultas realizadas fica disponível em diversos formatos, entre eles, principalmente o formato de tabela (Figura 16) e o formato bruto, ou seja, a resposta retornada pelo algoritmo do programa, em Json. É possível, ainda, salvar queries para serem executadas posteriormente, caso sua utilidade seja recorrente.

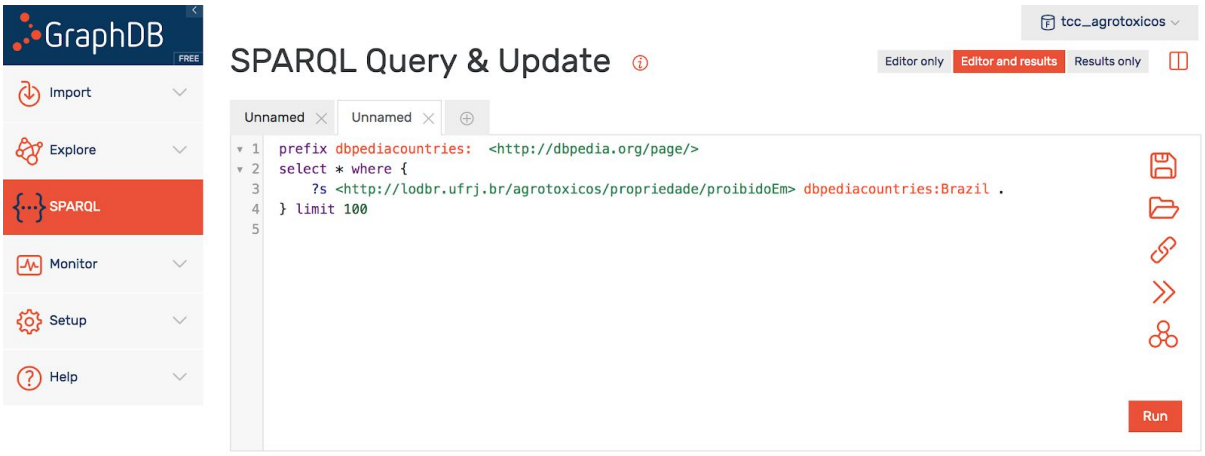


Figura 15 - Consulta em SPARQL realizada no GraphDB

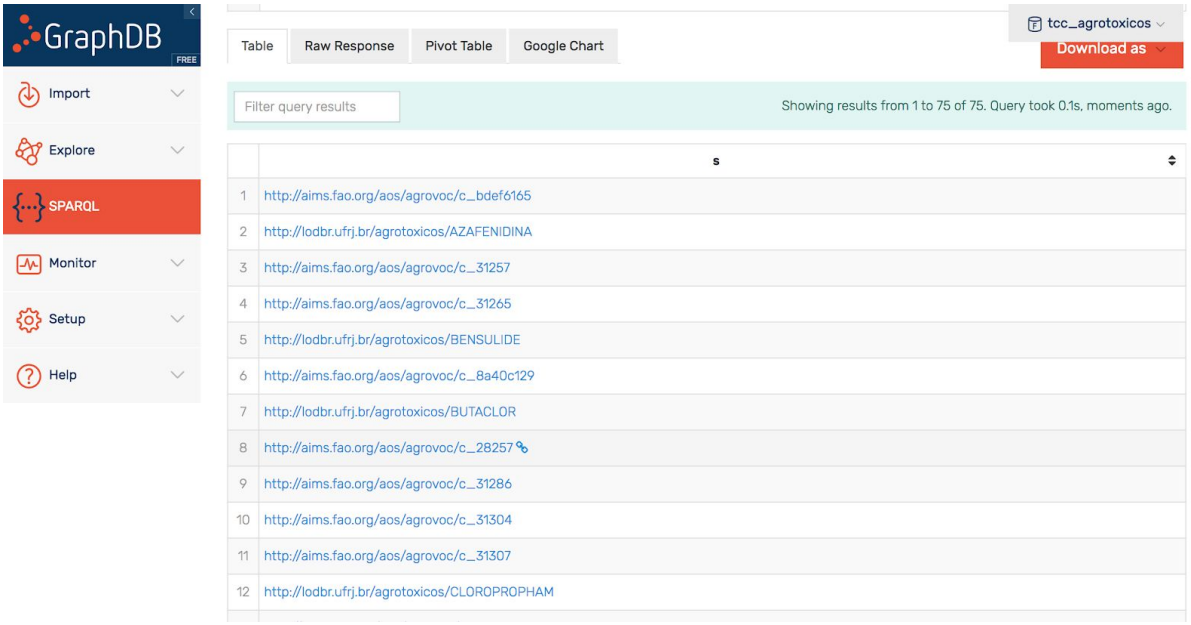


Figura 16 - Resposta à consulta realizada no GraphDB, no formato de tabela

Outra categoria de recursos oferecida pela ferramenta inclui formas mais avançadas de explorar os dados importados, entre elas, em especial, a construção automática de um grafo visual, integrando todos os dados de todas as bases importadas em um mesmo contexto, como pode ser visto na Figura 17.

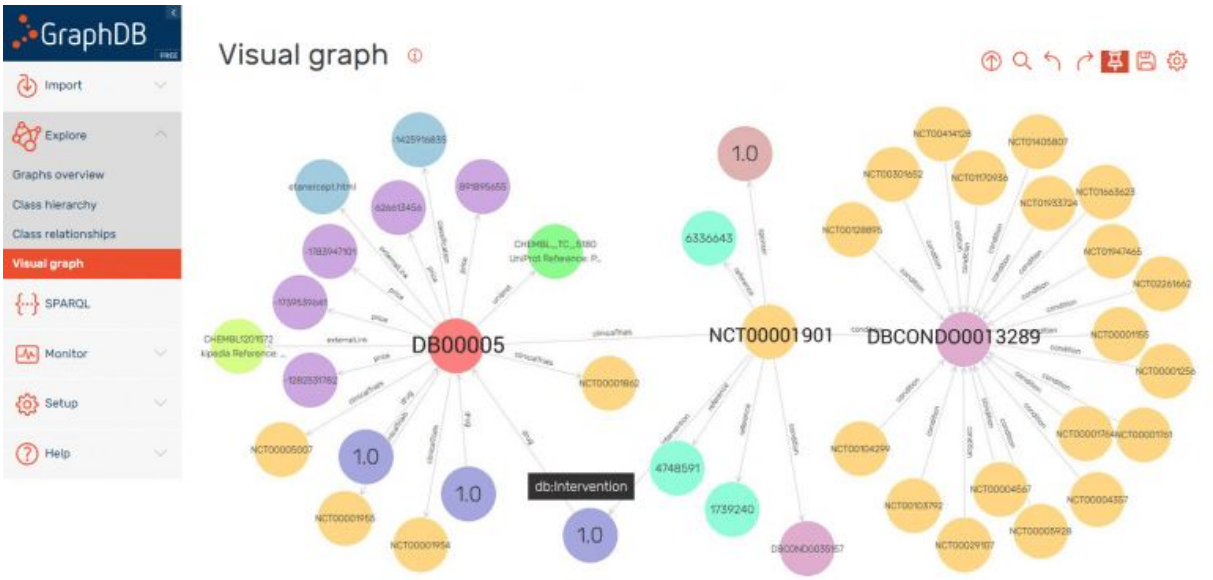


Figura 17 - GraphDB com informações expostas em grafos

3.6.6 ETL4LOD

Para implementar o fluxo de publicação de dados conectados, o projeto LinkedDataBR usou como ferramenta de ETL o Kettle ou Pentaho Data Integration. O Kettle fornece através de sua interface gráfica a opção de criar fluxos de ETL e permite que suas funcionalidades sejam estendidas através da criação de plug-ins.

O ETL4LOD, criado em 2011, é um conjunto de plug-ins que estendem as funcionalidades do Kettle para trabalhar com LOD fornecendo as funções de: conversão de dados a serem publicados, construção de consultas SPARQL, manipulação de dados armazenados em um endpoint após a publicação, anotação de dados em um determinado domínio, etc.

4 TRIPLIFICAÇÃO DE DADOS DO PORTAL DOS AGROTÓXICOS: EXPERIMENTAÇÃO

4.1. O DOMÍNIO DE AGROTÓXICOS DO PORTAL

Numerosas bases de dados que abrangem os mais diversos assuntos relacionados ao uso dos agrotóxicos no Brasil e no mundo são disponibilizadas na Internet. Entretanto, esses dados são encontrados dispersos e os estudos que já investiram em estruturá-los de forma a extrair informações mais relevantes podem, e devem, ser associados com mais dados com a finalidade de se possuir uma grande rede de conhecimento sobre os agrotóxicos. A rede de dados interligados pode esclarecer diversas questões que ainda não têm respostas nos dias atuais.

4.2 PROPOSTA DE SOLUÇÃO

Como uma prova de conceito sobre as possibilidades e vantagens da interligação de dados sobre agrotóxicos, optou-se por utilizar as bases que possam fornecer dados consistentes sobre os efeitos colaterais e sintomas que são causados em consequência da ingestão de agrotóxicos e bases que possam ajudar a posicionar o Brasil em comparação a outros países do mundo no quesito de regulamentação dos agrotóxicos. Na próxima seção deste capítulo, apresentam-se as bases de dados escolhidas que se encaixam nos critérios definidos acima como parte da etapa **Identificação e Seleção de Dados** do ciclo de vida dos dados conectados.

As bases selecionadas serão tratadas conforme a necessidade de cada uma, dependendo de sua disposição de dados inicial, língua em que os dados se encontram, forma que são distribuídas e formato das planilhas de dados. Diante deste cenário, será necessária uma limpeza na base de dados e uma organização

predisposta que facilite o entendimento e o processamento do conteúdo da base. O resultado final esperado é que estes dados contidos nas planilhas sejam transformados em triplas no formato RDF para que estes dados possam ser explorados e validados. Essas triplas serão expostas para visualização em formato de grafos e validadas conforme as questões que se deseja maiores esclarecimentos.

Atualmente no mercado há poucas ferramentas de triplificação de dados fornecidas, principalmente por se tratar de assunto relativamente recente. Dentre as opções de ferramentas gratuitas, as duas possíveis a serem utilizadas são o Open Refine e o Karma, apresentados no capítulo anterior.

Tanto o Open Refine quanto o Karma suportam a limpeza de dados e fornecem maneiras que permitem a importação de ontologias, atribuição de propriedades e classes dessas ontologias às colunas das tabelas, bem como suas respectivas URIs, e a exportação dessas tabelas corretamente mapeadas para triplas RDF. Além disso, ambas as ferramentas se propõem a atingir o objetivo de maneira simples sem maiores dificuldades.

4.2.1 Apresentação das Bases de Dados Escolhidas

Como descrito na seção anterior deste capítulo, foram escolhidas bases de dados que fornecessem noções para responder às seguintes questões:

- Quais são os problemas causados à saúde pela ingestão de agrotóxicos? Quais os sintomas imediatos e tardios da consumos dos mesmos? O quão prejudicial é viver em um ambiente contaminado por herbicidas, fungicidas e inseticidas?
- Quais agrotóxicos são permitidos para o uso no Brasil e não são permitidos no resto do mundo? E quais são proibidos aqui e permitidos no

mundo? O quanto o Brasil pode estar posicionado de forma a favorecer doenças crônicas à população consumidora de seus alimentos?

Tendo em vista essas perguntas, duas bases de dados que pudessem conceder um panorama a respeito foram selecionadas: Relação entre Agrotóxicos e Sintomas Agudos e Crônicos e *Consolidated List of Banned Pesticides*.

4.2.1.1 Relação entre Agrotóxicos e Sintomas Agudos e Crônicos

A ABRASCO, Associação Brasileira de Saúde Coletiva, tem o objetivo de atuar como mecanismo de apoio e articulação entre os centros de treinamento, ensino e pesquisa em Saúde Coletiva para fortalecimento mútuo das entidades associadas e para ampliação do diálogo com a comunidade técnico-científica e desta com os serviços de saúde, as organizações governamentais e não governamentais e a sociedade civil [ABRASCO, 2011].

Com o intuito de contribuir para o exercício do direito à saúde e consolidar políticas públicas responsáveis, a ABRASCO elaborou um dossiê para registrar e difundir a preocupação dos pesquisadores, professores e profissionais com a escalada ascendente de uso de agrotóxicos no país e a contaminação do ambiente e das pessoas dela resultante. O dossiê é dividido em três partes com focos distintos: *Agrotóxicos, Segurança Alimentar e Nutricional e Saúde*, *Agrotóxicos, Saúde e Sustentabilidade* e *Agrotóxicos, Conhecimentos e Cidadania*.

Para responder às primeiras questões descritas na seção 4.2.1 deste capítulo, é de interesse a primeira parte do dossiê ABRASCO referente a Segurança Alimentar e Nutricional e Saúde. Na seção de *Evidências Científicas: riscos para a saúde na exposição aos agrotóxicos por ingestão de alimentos* no item *B) Resíduos de agrotóxicos em alimentos e agravos à saúde* é fornecido o Quadro 1, de classificação e efeitos e/ou sintomas agudos e crônicos dos agrotóxicos.

Quadro 1 - Classificação e efeitos e/ou sintomas agudos e crônicos dos agrotóxicos:

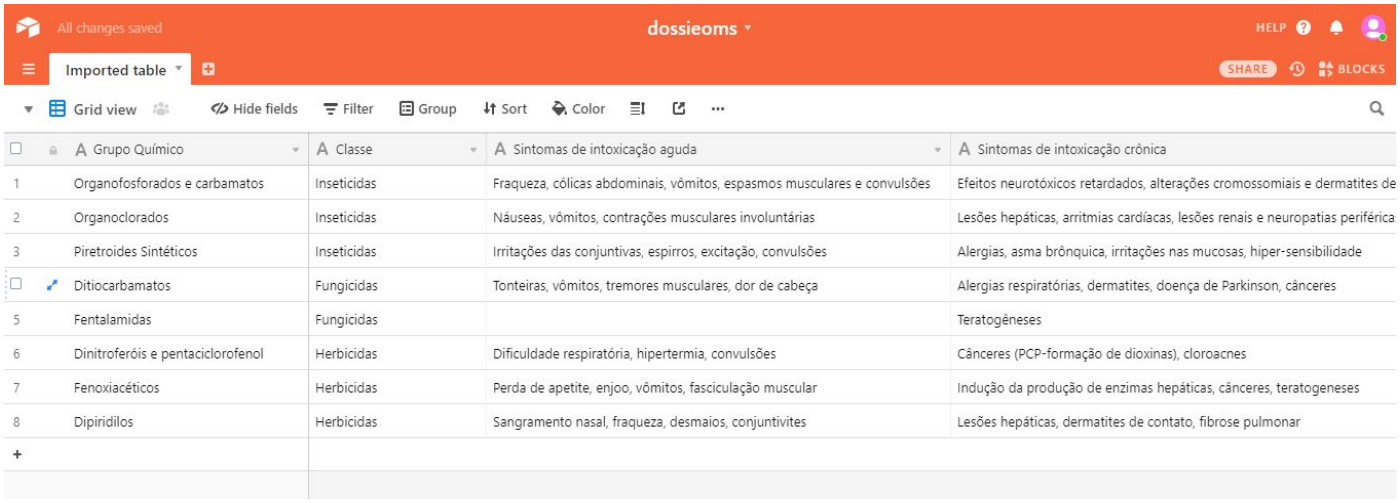
PRAGA QUE CONTROLA	GRUPO QUÍMICO	SINTOMAS DE INTOXICAÇÃO AGUDA	SINTOMAS DE INTOXICAÇÃO CRÔNICA
Inseticidas	Organofosforados e carbamatos	Fraqueza, cólicas abdominais, vômitos, espasmos musculares e convulsões	Efeitos neurotóxicos retardados, alterações cromossomiais e dermatites de contato
	Organoclorados	Náuseas, vômitos, contrações musculares involuntárias	Lesões hepáticas, arritmias cardíacas, lesões renais e neuropatias periféricas
	Piretroides sintéticos	Irritações das conjuntivas, espirros, excitação, convulsões	Alergias, asma brônquica, irritações nas mucosas, hipersensibilidade
Fungicidas	Ditiocarbamatos	Tonteiras, vômitos, tremores musculares, dor de cabeça	Alergias respiratórias, dermatites, doença de Parkinson, cânceres
	Fentalamidas	-	Teratogêneses
Herbicidas	Dinitroferóis e pentaciclorofenol	Dificuldade respiratória, hipertermia, convulsões	Cânceres (PCP-formação de dioxinas), cloroacnes
	Fenoxiacéticos	Perda de apetite, enjoo, vômitos, fasciculação muscular	Indução da produção de enzimas hepáticas, cânceres, teratogeneses
	Dipiridilos	Sangramento nasal, fraqueza, desmaios, conjuntivites	Lesões hepáticas, dermatites de contato, fibrose pulmonar

Fonte: OPAS/OMS (1996).

O quadro descreve os sintomas agudos e crônicos dos principais grupos químicos dos agrotóxicos. Os efeitos agudos são aqueles sintomas que aparecem poucas horas após a ingestão. São estes efeitos que classificam os agrotóxicos como medianamente ou pouco tóxicos. Embora haja essa categorização não se pode deixar de considerar os efeitos que podem ocorrer meses, anos e até décadas após a exposição, estes efeitos são chamados de efeitos crônicos. Cânceres, malformação congênita, distúrbios endócrinos, neurológicos e mentais são alguns dos efeitos crônicos causados expostos no quadro.

Os agrotóxicos listados foram encontrados nos alimentos analisados pelo Programa de Análise de Resíduos de Agrotóxicos em Alimentos (PARA) da ANVISA, anteriormente mostrado na Figura 1 deste trabalho.

Este quadro gerou a base de dados em CSV exposta no Portal dos Agrotóxicos [PORTAL] e parcialmente representado pela Figura 18.



	Grupo Químico	Classe	Sintomas de intoxicação aguda	Sintomas de intoxicação crônica
1	Organofosforados e carbamatos	Inseticidas	Fraqueza, cólicas abdominais, vômitos, espasmos musculares e convulsões	Efeitos neurotóxicos retardados, alterações cromossomiais e dermatites de
2	Organoclorados	Inseticidas	Náuseas, vômitos, contrações musculares involuntárias	Lesões hepáticas, arritmias cardíacas, lesões renais e neuropatias periférica
3	Piretroides Sintéticos	Inseticidas	Irritações das conjuntivas, espirros, excitação, convulsões	Alergias, asma brônquica, irritações nas mucosas, hiper-sensibilidade
	Ditiocarbamatos	Fungicidas	Tonteiras, vômitos, tremores musculares, dor de cabeça	Alergias respiratórias, dermatites, doença de Parkinson, cânceres
5	Fentalamidas	Fungicidas		Teratogêneses
6	Dinitroferóis e pentaciclorofenol	Herbicidas	Dificuldade respiratória, hipertermia, convulsões	Cânceres (PCP-formação de dioxinas), cloroacnes
7	Fenoxiacéticos	Herbicidas	Perda de apetite, enjoo, vômitos, fasciculação muscular	Indução da produção de enzimas hepáticas, cânceres, teratogeneses
8	Dipiridilos	Herbicidas	Sangramento nasal, fraqueza, desmaios, conjuntivites	Lesões hepáticas, dermatites de contato, fibrose pulmonar

Figura 18 - Base de dados retirada do dossiê em CSV

A partir deste arquivo, tem-se como objetivo final a obtenção de triplas que caracterizem a relação entre grupo químico e classe com efeitos causados. É importante ressaltar que não há interesse na diferenciação entre sintomas crônicos e agudos nesta triplificação.

4.2.1.2 Consolidated List of Banned Pesticides

A *Pesticide Action Network* (PAN) é uma rede composta por mais de 600 organizações não-governamentais, instituições e indivíduos em mais de 90 países que trabalham desde 1982 para substituir o uso de agrotóxicos por substâncias e alternativas ecologicamente corretas e em benefício à sociedade. O PAN possui

cinco centros regionais em torno do mundo, além de diversos projetos e campanhas em prol de seus ideais.

Após ser frequentemente solicitada por informações a respeito sobre agrotóxicos que foram banidos pelos países específicos do mundo, foi desenvolvida a Lista Consolidada (CL) de Pesticidas Banidos com o intuito de fornecer esta fonte de dados, que parece não estar disponíveis em nenhum outro lugar.

Essa lista é atualizada a cada seis meses e a PAN alerta que, apesar dos esforços, inevitavelmente podem haver erros visto que para alguns países são divulgadas informações conflitantes. Entretanto, a lista permite solucionar os questionamentos levantados na seção 4.2.1 deste mesmo capítulo, pois fornece exatamente a lista entre agrotóxicos banidos pelos países pelo mundo, como podemos ver na Figura 19.

A	B	C	D	E	BS	BT	BU	BV	BW	BX	BY	BZ	CA	CB	CC	CD	CE	CF	CG	CH	CI	CJ
0																						
CAS No	Active	TOTAL BANS PER ACTIVE INGREDIENT	PAN HHP	JMPM HHP	SAUDI ARABIA	SENEGAL	SOUTH AFRICA	SRI LANKA	SURINAME	SWEDEN (in addition to EU bansand not approved)	SWITZERLAND	SYRIA	THAILAND	TOGO	TRINIDAD & TOBAGO	URUGUAY	USA	VENEZUELA	VIETNAM	ZIMBABWE		
	TOTAL BANS PER COUNTRY				65	7	14	24	24	3	12	4	40	5	12	12	1	6	27	9		
7784-46-5	sodium arsenite	30	X	X																	30	
3586-60-5	asomate	29																			29	
1912-24-9	atrazine	37	X			1								1							37	
68049-83-2	azafenidin	29	X	X																	29	
35575-96-3	azamethiphos	28	X																		28	
2642-71-9	azinphos-ethyl	34	X	X	1		1						1								34	
86-50-0	azinphos-methyl	39	X	X	1								1								39	
41083-11-8	azocyclotin	28	X																		28	
22781-23-3	bendiocarb	29	X		1																29	
82560-54-1	benfuracarb	28	X																		28	
17804-35-2	benomyl	33	X	X	1																33	
741-58-2	bensulide	30	X																		30	
82657-04-3	bifenthrin	2	X																		2	
42576-02-3	bifenox	1																			1	
28434-01-7	bioresmethrin	28	X																		28	

Sheet1 Sheet2 Sheet3

Figura 19 - Base de dados Consolidated List of Banned Pesticides

A lista é disponibilizada no formato XLSX e constituída, nas linhas, pela ocorrências de 370 agrotóxicos em ordem alfabética e, nas colunas, pela União Européia, a qual possui 28 países integrantes e os outros países do mundo. Na coluna *Total Bans Per Active Ingredient* informa o número total de países que baniram cada agrotóxico em particular.

Na planilha também é exibido se os agrotóxicos listados são considerados altamente perigosos, segundo os critérios estabelecidos pelo FAO/WHO Joint Meeting in Pesticide Management (JMPM) - vide coluna *JMPM HHP*. Os agrotóxicos em azul, como podemos ver na Figura 19, são aqueles que não foram proibidos em nenhum país mas são considerados altamente perigosos, de acordo com os critérios do PAN, e não são aprovados nos países da União Europeia.

A lista consolidada não inclui os agrotóxicos proibidos considerados obsoletos pela OMS e também não inclui as restrições severas. Ou seja, as ocorrências da tabela são apenas para proibições completas, porque em alguns países mesmos os agrotóxicos totalmente proibidos ainda são usados.

Nas linhas, é possível observar as seguintes ocorrências e seus respectivos significados:

- 1: o ativo correspondente a linha é proibido
- Espaço em branco: o agrotóxico é aprovado
- ?: significa que as informações não foram disponibilizadas no país em questão

O propósito da escolha desta base para triplificação é que sejam geradas relações de proibição entre cada ocorrência de agrotóxico e os países onde o mesmo é banido. Para isso, a coluna União Europeia será desmembrada nos 28 países que a compõem e será ignorada a coluna de “Não aprovados na União Européia”. Além disto, as ocorrências desconhecidas, sinalizadas como *?*, serão consideradas como aprovadas.

4.2.2 Apresentação das Ontologias Necessárias

Fundamentalmente, os dados contidos nas bases descritas acima serão representados como um nó na Web de Dados após a triplificação e publicação dos mesmos. Estes nós serão anotados segundo vocabulários e ontologias abertas para facilitar o entendimento e a difusão desta pesquisa.

As ontologias necessárias para apoio à modelagem e anotação das bases de dados escolhidas são ChEBI, MedDRA, DBPEDIA/Country e AGROVOC. Abaixo é descrita cada uma delas.

4.2.2.1 Ontologia ChEBI

O ChEBI, *Chemical Entities of Biological Interest*, é um dicionário, disponibilizado gratuitamente, de entidades moleculares focadas em pequenos compostos químicos. “Entidades moleculares” pode se referir a moléculas, átomos, íon, par iônico, radical, etc. desde que seja uma entidade separável distinta, as quais podem ser sintéticas ou frutos da natureza. Também implementa uma classificação segundo às especificações das relações entre entidades moleculares ou classes de entidades e seus pais ou filhos.

Essa ontologia usa as nomenclaturas, simbolismos e terminologias estabelecidos pela *International Union of Pure and Applied Chemistry* (IUPAC) e Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Entretanto as moléculas como ácido nucléicos, proteínas e peptídeos caracterizadas pela codificação por genoma não são incluídas no ChEBI.

Todos os dados no banco de dados não são proprietários ou são derivados de uma fonte não proprietária. É assim livremente acessível e disponível para qualquer pessoa. Além disso, cada item de dados é totalmente rastreável e explicitamente referenciado à fonte original [ChEBI’s Documentation].

O ChEBI mostra os seguintes campos de dados:

- ChEBI Identifer: o identificador único
- Nome ChEBI: o nome recomendado para uso em bancos de dados biológicos
- ChEBI ASCII Name: o nome do ChEBI com quaisquer caracteres especiais renderizados no formato ASCII
- Classificação por estrelas: Uma classificação baseada no nível de anotação manual
- Estrutura: a representação gráfica da estrutura molecular e molécula associada
- Fórmula: Fórmula Molecular
- Carga
- Massa Média
- Ontologia ChEBI: Onde é possível ter a visão de saída e entrada e a visão em árvore da posição da entrada dentro da Ontologia ChEBI
- Nome IUPAC: nome gerado de acordo com as recomendações da IUPAC
- INN: International Nonproprietary Name, também conhecido como nome genérico, atribuído pela Organização Mundial da Saúde (OMS)
- Sinônimos: outros nomes, juntamente com uma indicação da sua fonte
- Marca: um nome comercial ou proprietário
- Links de banco de dados: referências cruzadas manualmente selecionadas para outros bancos de dados não proprietários
- Número do Registro: Número do Registro do CAS (número de registro único no banco de dados do Chemical Abstracts Service, uma divisão da Chemical American Society), Número do Registro do Beilstein, Número do Registro do Gmelin (se disponível)
- Citações: Publicações que citam a entidade junto com hiperlinks para suas entradas.

O ChEBI será usado para referenciar os dados descritos na coluna de Grupos Químicos e Classe da base de dados Relação entre Agrotóxicos e Sintomas Agudos e Crônicos.

4.2.2.2 Ontologia MedDRA

O MedDRA, *Medical Dictionary for Regulatory Activities*, é uma terminologia médica padronizada para facilitar a partilha de informações regulamentadas sobre produtos médicos usados por seres humanos a nível internacional. A terminologia é distribuída gratuitamente e está a disposição de todos para utilização relativa ao registro, documentação e monitorização da segurança de produtos médicos e doenças ou distúrbios relacionados a venda autorizada de um produto.

Esta ontologia estende-se a medicamentos, produtos biológicos, vacinas e produtos resultantes da combinação de medicamentos com dispositivos médico, além dos sintomas e doenças que os mesmos atingem.

Para além da versão original em inglês e da tradução em japonês, a terminologia MedDRA tem sido traduzida e mantida nos idiomas seguintes: chinês, checo, holandês, francês, alemão, húngaro, italiano, português e espanhol. Cada termo MedDRA possui um código numérico associado de 8 dígitos que é sempre o mesmo independentemente do idioma [MedDRA's Documentation].

A terminologia MedDra foi desenvolvida como uma validação médica para utilização durante o processo de regulamentação. Por essa razão, foi estruturada uma relação entre os termos existentes na terminologia para facilitar a organização, o entendimento e a busca de termos dentro da mesma. As relações entre os termos podem ser das seguintes categorias: Equivalência ou Hierarquia.

A relação de equivalência agrupa sinônimos ou termos equivalentes. A relação de hierarquia proporciona graus ou níveis de ordenação e subordinação dos termos, onde cada grau possui um nível de especificidade ou granulosidade. Esta

hierarquia é dividida em cinco níveis verticais: Grupo Sistêmico, Termo HLGT, Termo HLT, Termo PT e Termo LLT, respectivamente como podemos ver na Figura 20.

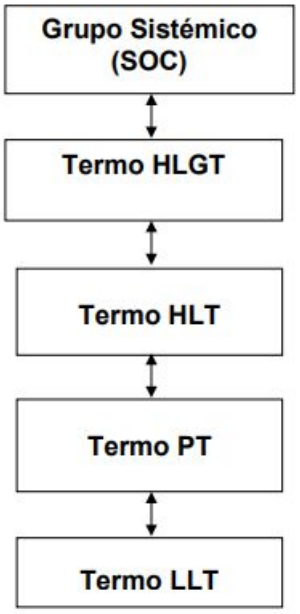


Figura 20 - Hierarquia estrutural da terminologia MedDRA

O nível mais baixo, LLT (Lowest Level Term), possui a especificidade máxima e está ligado somente a um termo PT (Preferred Term). Os termos HLT (High Level Terms) e os termos HLGT (High Level Group Terms) facilitam a recuperação e a exibição dos dados pois fornecem um agrupamento de termos por importância clínica. Ambos estão relacionados a uma anatomia, patologia, fisiologia, etiologia ou função. O grupo SOC (System Organ Classes), o nível mais alto da hierarquia, representam eixos paralelos que não se excluem mutuamente como por exemplo Infecções e infestações e Afecções gastrointestinais. Por último os termos PT, que se situam num nível mediano dos anteriormente descritos, é um elemento descritivo do conceito médico para um sintoma, sinal, doença, diagnóstico, recomendação terapêutica, etc.

No âmbito deste trabalho, o MedDRA será usado para mapear a coluna de Sintomas Crônicos e Sintomas Agudos da base Relação entre Agrotóxicos e Sintomas Agudos e Crônicos.

4.2.2.3 Ontologia DBpedia/Country

A DBpedia é um projeto cujo objetivo é extrair conteúdo das informações da Wikipedia e estruturá-las para então disponibilizá-las em uma robusta rede de ligações entre os dados, permitindo assim realizar consultas sobre o conteúdo de forma similar a um banco de dados [DBPedia Wiki].

A ontologia de País/Country fornecida pela DBpedia é uma entidade do tipo Property das classes de recursos disponíveis pelo projeto. Cada ocorrência de país na DBpedia possui as seguintes propriedades:

- Área Total
- Densidade Populacional
- Resumo: contendo um texto com informações relativas àquele país disponível em português e inglês.
- Capital
- Moeda
- Nome
- Grupos étnicos
- Bandeira: uma imagem em SVG
- Tipo de governo
- Língua
- Líderes
- Título/Cargo dos líderes
- Nome completo: Como por exemplo, República Federativa do Brasil
- Língua oficial

- Porcentagem de área com água
- População total
- Timezone
- Cidades
- Links externos

Entre outros…

Esta ontologia será usada para mapear os dados referentes aos países presentes da base de dados *Consolidated List of Banned Pesticides*.

4.2.2.4 Ontologia AGROVOC

O AGROVOC é a ontologia que cobre as áreas de interesse da *Food and Agriculture Organization* (FAO) como agricultura, silvicultura, pesca e meio ambiente [*AGROVOC Guidelines]*. O principal objetivo foi padronizar o processo de indexação para a base de dados AGRIS para facilitar a busca e torná-las mais eficiente. Esta ontologia está disponível em inglês, francês, espanhol, árabe, chinês e russo.

Seu papel é ajudar na padronização da anotação dos objetos de informação para indexar, recuperar e organizar dados de sistemas agrícolas. É amplamente utilizado por pesquisadores, bibliotecários, gestores de informação e outros profissionais que possua interesse nesta área do conhecimento.

A estrutura do AGROVOC é composta por termos que se constituem de uma ou mais palavras que simbolizam um conceito. Como o MedDRA, explicado na seção 4.2.2.2 deste capítulo, o AGROVOC também é constituído de relações hierárquicas e não hierárquicas. Para cada termo, há um conjunto de palavras que mostram essas relações. São elas:

- BT: Broader Term, termo genérico
- NT: Narrower Term, termo específico
- RT: Related Term, termo relacionado

- UF: Use For, utiliza-se para.

Por exemplo, o termo *Poluição* tem o NT *Poluição do Ar*, enquanto o termo *Poluição do Ar* têm os termos BT *Poluição*, RT *Efeito Estufa*. Usualmente também há notas de escopo ou definições para esclarecer o significado e contexto dos termos.

O AGROVOC será utilizado para mapear os dados relacionados aos agrotóxicos banidos da base de dados *Consolidated List of Banned Pesticides*.

5 O PROCESSO DE TRIPLIFICAÇÃO DE DADOS DE AGROTÓXICOS

Conforme visto na seção 4.2.1 do Capítulo 4, as bases escolhidas estão, originalmente, no formato CSV e XLSX, então o processo será descrito partindo deste princípio. Entretanto, é perfeitamente possível transformar dados de outros formatos em triplas RDF, e o processo muito se assemelha, se assim desejado, ao que será descrito a seguir.

Em linhas gerais, o processo se resume em 5 macro etapas, sendo uma delas opcional: modelagem, limpeza, tradução - sendo este relativo, a depender do objetivo a ser alcançado com a triplificação, adaptação e conversão para triplas.

Neste trabalho, foi feita também uma análise de viabilidade de conversão das bases de dados selecionadas por meio dos programas Open Refine e Karma, descritos na seção 3.3.1 e 3.3.2 do capítulo 3.

As ferramentas de triplificação que foram escolhidas e levantadas para serem aplicadas neste trabalho passaram pelos critérios de escolha que se referiam a ser um software de código aberto, haver interface gráfica para facilitar o entendimento do processo, aceitação de dados de entrada tanto em CSV como XLSX para atender ambas as bases de dados escolhidas e que, sobretudo, houvesse uma comunidade ativa que oferecesse atualizações e manutenções da ferramenta, a qual tivesse sua última versão disponibilizada pelo período máximo de 1 ano. Por estes motivos, somados à falta de familiaridade das autoras com o Kettle, o ETL4LOD foi descartado para uso neste trabalho.

5.1 MODELAGEM

As modelagens feitas para as bases de dados deste trabalho foram pensadas de modo a serem compatíveis com a modelagem do trabalho "Extração de dados de

fontes textuais: uma abordagem para enriquecimento de dados abertos interligados". [TEIXEIRA, 2018], visando uma integração entre os dados obtidos neste e naquele trabalho. Seguindo o que foi discutido na seção 4.1, abaixo será descrito o processo de modelagem para cada uma das bases de dados escolhidas como parte da etapa **Limpeza, Anotação e Transformação** do ciclo de vida dos dados conectados.

5.1.1 Modelo de Agrotóxicos Versus Sintomas Gerados

A base de dados Relação entre Agrotóxicos e Sintomas Crônicos e Agudos é composta de 4 colunas: Classe, Grupo Químico, Sintomas Crônicos e Sintomas Agudos. Fundamentalmente, neste trabalho, não há a preocupação com a natureza do sintoma, apenas com quais sintomas um agrotóxico pode causar, seja ele a curto ou longo prazo.

Um agrotóxico é composto por um ou mais grupos químicos e é caracterizado por quais tipos de pragas ele controla. Neste caso, foi definido como Entidade Química o conjunto destas informações, grupo químico e classe, as quais caracterizam um ou mais agrotóxicos. Porém, uma Entidade Química é descrita por apenas um único grupo químico e uma única classe, podendo caracterizar um ou mais agrotóxicos, os quais não estão previstos no escopo dessa base.

Ciente destes fatos, a modelagem representada na Figura 21 foi prevista.

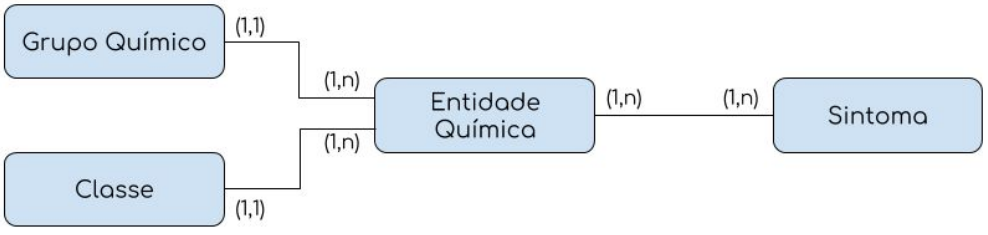


Figura 21 - Modelagem da Relação entre Sintomas e Agrotóxicos

As triplas geradas respeitarão essa modelagem e serão caracterizadas segundo as relações as quais elas descrevem.

5.1.2 Modelo de Agrotóxicos Banidos Versus País

A *Consolidated List of Banned Pesticides* é composta, fundamentalmente, em suas linhas pelos agrotóxicos banidos e em suas colunas pelos países. Basicamente, a planilha caracteriza uma matriz onde cada correspondência define se o agrotóxico daquela linha é proibido, ou não, na coluna que simboliza um país. Claramente, para geração de triplas, essa lista precisará de um grande trabalho de limpeza de dados, que será descrito na seção 5.2 deste mesmo capítulo.

Como, basicamente, o processo de triplificação visa gerar uma relação de agrotóxicos banidos e país, a modelagem é relativamente simples. Um agrotóxico pode ser banido em um ou mais países e um país tem um ou mais agrotóxicos banidos como representado na Figura 22.

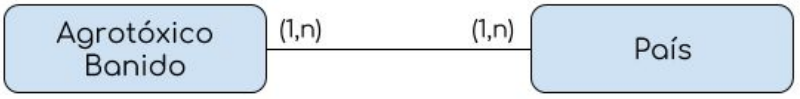


Figura 22 - Modelagem da relação entre países e agrotóxicos banidos

5.2 LIMPEZA

Nas seções seguintes será descrito o processo de limpeza para as bases selecionadas também como parte da etapa **Limpeza, Anotação e Transformação** do ciclo de vida dos dados conectados.

5.2.1 Limpeza da Relações entre Agrotóxicos e Sintomas

A limpeza dessa base foi feita usando o Airtable, mas poderia ter sido feita manualmente em um editor comum de dados CSV, desde que este editor estruturasse os dados em formato de tabela/planilha para facilitar o entendimento e discernimento de células.

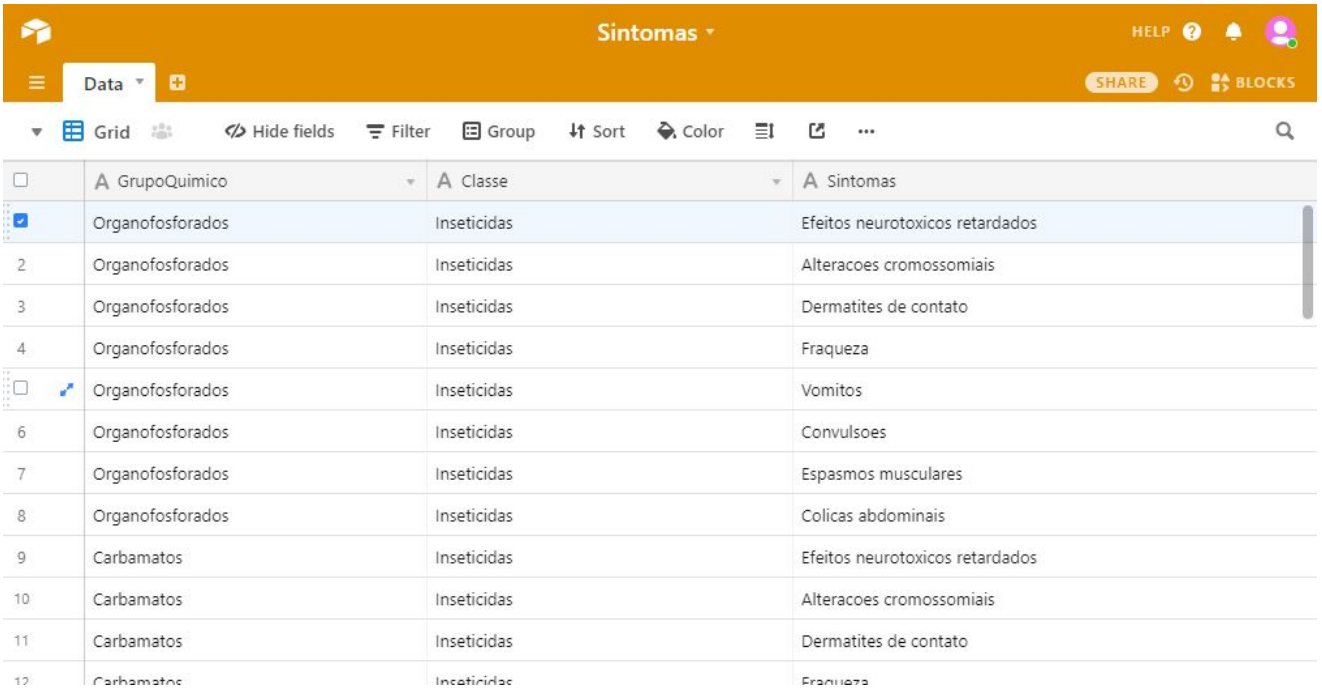
Interessam-se essencialmente por três dados principais nesta planilha: Grupo Químico, Classe e Sintoma. Como foi citado anteriormente, não será levado em consideração se o sintoma causado é agudo ou crônico, pois não há relevância, atualmente, na classificação temporal destes sintomas quanto a seu aparecimento. Logo, espera-se que após a limpeza, os dados estejam estruturados nestas três colunas base.

A primeira ação tomada foi a substituição de todos os caracteres especiais presentes na planilha, para que não houvesse a interferência de codificações no momento da triplificação de fato. O segundo passo foi a separação dos dados que estavam na mesma célula, visto que em cada célula só pode conter um dado referente àquela coluna, de modo que cada dado esteja em uma linha. Por exemplo, no Quadro 1 é possível observar que, tanto no caso de Grupo Químico quanto no caso de Sintomas, há mais de um dado numa mesma célula.

Na coluna de *Grupo Químico* são separadas as ocorrências que tratam de grupos químicos distintos. Por exemplo: embora *Organofosforados* e *Carbamatos* pertençam à mesma classe e causem os mesmos sintomas, eles deverão ser separados em duas linhas distintas mesmo que essa divisão implique na repetição de dados. Na limpeza das bases a redundância de informações não é uma restrição a ser tratada.

Nas colunas de *Sintomas Crônicos* e *Sintomas Agudos*, além de haver o mesmo problema de mais de um dado em uma célula, é preciso colocar todos os

sintomas em uma mesma coluna. Então, uma única coluna de *Sintomas* foi criada e para cada ocorrência, uma nova linha, como pode ser visto na Figura 23.



	GrupoQuimico	Classe	Sintomas
	Organofosforados	Inseticidas	Efeitos neurotoxicos retardados
2	Organofosforados	Inseticidas	Alteracoes cromossomiais
3	Organofosforados	Inseticidas	Dermatites de contato
4	Organofosforados	Inseticidas	Fraqueza
	Organofosforados	Inseticidas	Vomitos
6	Organofosforados	Inseticidas	Convulsoes
7	Organofosforados	Inseticidas	Espasmos musculares
8	Organofosforados	Inseticidas	Colicas abdominais
9	Carbamatos	Inseticidas	Efeitos neurotoxicos retardados
10	Carbamatos	Inseticidas	Alteracoes cromossomiais
11	Carbamatos	Inseticidas	Dermatites de contato

Figura 23 - Base de dados Relação entre Agrotóxicos e Sintomas após limpeza realizada

5.2.2 Limpeza da Consolidated List of Banned Pesticides

Essa base de dados, no formato XLSX, estruturou os dados em forma de matriz. Embora essa estrutura facilite o entendimento humano das informações ali presentes, não há como extrair triplas dos dados distribuídos de forma tão esparsa. Por essa razão, será necessário um processo meticuloso de limpeza da base onde será necessário extrair os conteúdos dessa matriz e dispô-los em forma de lista.

Basicamente, os dados de interesse são os agrotóxicos banidos em um determinado país, ou seja, agrotóxicos banidos e país. Logo, os dados serão estruturados em uma lista contendo essa dupla de dados (Figura 24).

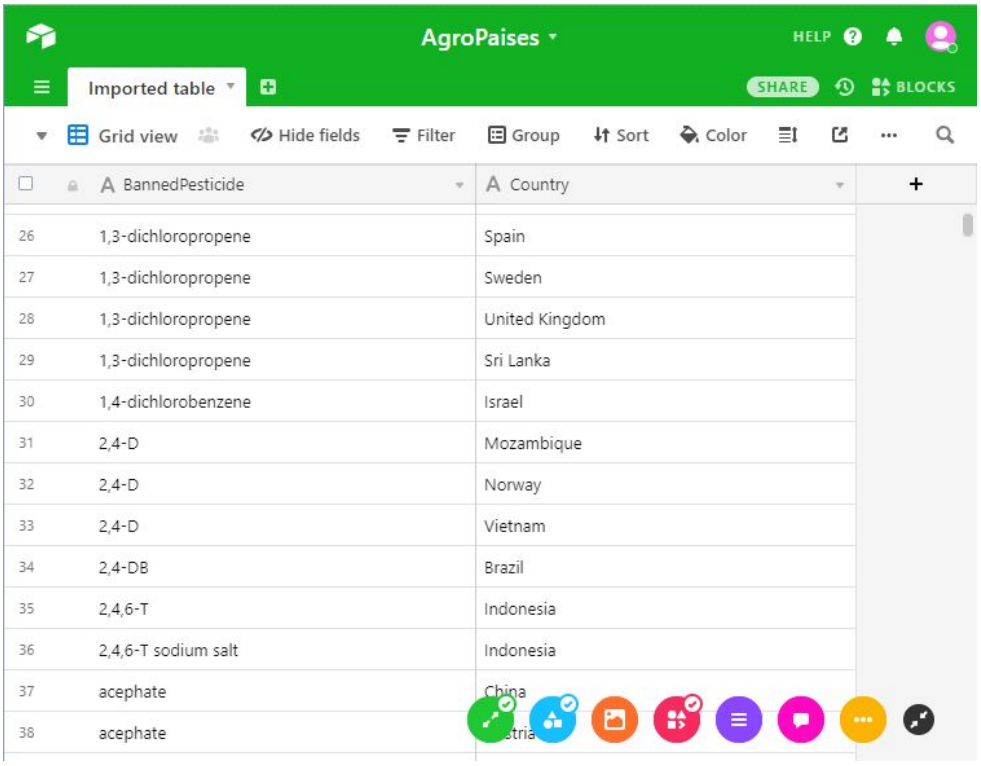


Figura 24 - Base de dados Consolidated List of Banned Pesticide após a limpeza

Para construir essa nova lista há diversas formas, como por exemplo construir um programa para ler de um arquivo CSV e devolva como saída uma nova planilha de dados estruturados. Neste caso, optou-se por uma solução mais simples e eficaz: Importar a lista consolidada em seu estado natural, como na Figura 19, para uma tabela em um banco de dados e executar uma consulta que já devolvesse as informações dispostas da forma como desejado.

A consulta utilizada para realizar a extração de um novo CSV está no Anexo 1 deste trabalho. A consulta em questão percorre a planilha célula a célula e utiliza-se do DBMS para adicionar a dupla *Agrotóxico Banido* e *País* à saída. Além disso, como uma das colunas da base de dados refere-se a União Européia, a consulta já realiza o tratamento incluindo na saída 28 linhas referente a cada país que a compõe.

5.3 TRADUÇÃO

Esta é uma etapa opcional, que também se encaixa como parte da etapa **Limpeza, Anotação e Transformação** do ciclo de vida dos dados conectados, pois depende dos objetivos a serem alcançados com o trabalho. Muitas ontologias estão disponíveis apenas em uma determinada língua, especialmente na língua inglesa, e se o objetivo do trabalho é contribuir com a disponibilização de determinados dados em Web Semântica, originalmente em outra língua, como é o caso deste, faz-se necessária a tradução da base de dados para a língua inglesa, mantendo também os dados no idioma original, no caso, português.

É possível relacionar somente os dados traduzidos para a língua inglesa com as ontologias escolhidas, interligando-os, posteriormente, com o idioma original, por meio do predicado *sameAs*, por exemplo. Esta foi a abordagem utilizada neste trabalho para associar as bases de dados em português com ontologias já existentes em língua inglesa.

O método utilizado neste trabalho para traduzir a base de dados foi escrever um programa na linguagem de programação Python, responsável por ler a base de dados CSV e, por meio de uma biblioteca responsável por fazer requisições à API do *Google Translator*, gerar um novo arquivo, ainda no formato CSV, com as colunas originais e cópias destas traduzidas para o idioma inglês. Para o caso de bases com grandes volumes de dados, entretanto, existe uma limitação no uso desta biblioteca, pois a API do *Google Translator* limita a quantidade de requisições em um curto período de tempo vinda de uma mesma origem. Deste modo, fazer um programa que leia uma base de dados extensa e a traduza de uma vez pode não ser viável através deste método, se fazendo necessário limitar o número de requisições ao permitido pela API utilizada, ou pagando pelo serviço do Google Translator, que cobra por quantidade de caracteres traduzidos.

Outra possibilidade, muito mais extensa e trabalhosa, seria expandir ou alterar o objetivo para a criação ou tradução das ontologias existentes para

português. Entretanto, isso demandaria um trabalho muito mais longo, pois a construção de ontologias é um outro campo de estudo, cheio de particularidades e dificuldades que extrapolam o objetivo deste trabalho.

Esta etapa foi necessária em toda a base de dados *Relações entre Agrotóxicos e Sintomas Crônicos e Agudos*, a qual é disponibilizada apenas em português, e na coluna *Agrotóxicos Banidos* da base de dados *Consolidated List of Banned Pesticides,* porque nem todos os agrotóxicos listados estão presentes no vocabulário do AGROVOC, sendo assim necessária a criação de uma URI em português para estas exceções.

5.4 ADAPTAÇÃO

Uma tripla de dados consiste em associar uma URI a uma outra URI ou a uma label por meio de um predicado. Sendo assim, essas URIs devem surgir em algum momento do processo de triplificação, e na abordagem utilizada neste trabalho, as URIs estão presentes na base de dados. Esta, inclusive, é uma abordagem também recomendada por alguns softwares de modelagem e transformação de dados para triplas RDF, como o *Karma* e o *Open Refine*, que serão abordados posteriormente neste capítulo. Por isso, uma das etapas citadas para o processo de triplificação consiste na etapa de adaptação da base, isto é, inserção das URIs na base de dados, para cada item da base, atuando na etapa **Interligação** do ciclo de vida dos dados conectados.

Para o caso de arquivos CSV, um item pode ser considerado uma célula, e, geralmente, itens de uma mesma categoria são agrupados em uma mesma coluna. Sendo assim, a estratégia utilizada neste trabalho foi a de criar uma coluna de URIs para cada coluna de dados das bases, que posteriormente, além de servir para as triplas de relacionamento entre os dados usando os predicados, serviu também para associar as URIs às respectivas labels, ou seja, aos dados das colunas originais.

Este processo foi feito tanto para os dados em português, que é a língua original de uma das bases de dados utilizadas, como também para os dados em inglês, obtidos na etapa de tradução.

Para realizar a etapa de adaptação, há diversas possibilidades, assim como em todas as outras abordagens. Por esta etapa ser realizada de maneira idêntica para ambas as bases de dados de interesse previamente escolhida, abaixo serão descritas as possibilidades de adaptação necessárias para cada caso encontrado neste trabalho respeitando as peculiaridades de cada informação.

Um caso observado neste trabalho é quando o dado está explícito no endereço da URI, ou seja, as labels estão contidas na URI que compartilham de um mesmo prefixo, torna-se mais fácil a geração automáticas dos endereços. Aqui o método escolhido foi utilizar o *Airtable,* já descrito na seção 3.3 do Capítulo 3, permitindo a inclusão de fórmulas lógicas para gerar novos dados. Este caso foi encontrado na coluna *Country* da segunda base de dados escolhidas, citando como exemplo a URI:

http://dbpedia.org/page/Belgium

em que o prefixo é sempre *http://dbpedia.org/page/* concatenado, no final, com o nome, em inglês, do país, já contido na base de dados original. Por isso, para este caso, basta escrever uma fórmula lógica responsável por concatenar os dados já existentes na tabela com o prefixo da URI existente na ontologia. Deste modo, também é possível criar novas URIs seguindo este mesmo padrão, para o caso de dados que não existam na ontologia escolhida ou para o caso da abordagem de mais de um idioma.

Outro caso deste trabalho, como encontrado na coluna *Sintomas* do quadro fornecido pela ABRASCO, algumas URIs precisaram ser criadas para representar dados em português e também dados não mapeados por ontologias pré-existentes, como nos dois exemplos a seguir:

http://lodbr.ufrj.br/efeito/Dermatites_de_contato

http://lodbr.ufrj.br/chemical_group/Organophosphates

onde a primeira foi criada para conter o dado em português, interligado através do predicado *sameAs* com uma URI relacionada em inglês, mapeada pela ontologia MedDRA e a segunda, criada para representar dados que não existiam na ontologia ChEBI, ambas as ontologias descritas no Capítulo 4 nas seção 4.2.2.2 e 4.2.2.1, respectivamente.

Em linhas gerais, o processo de adaptação é possível ser realizado após escrever um programa responsável por fazer a criação e associação da coluna de URIs com a coluna de labels, mas este depende das ontologias escolhidas possuírem uma API, e é também mais trabalhoso para o caso de bases que necessitem de mais de uma ontologia, pois não existe um padrão global de APIs para ontologias, então o programa deverá ser adaptado para a API de cada ontologia necessária.

Para o caso de bases de dados que não sejam muito grandes ou variem muito de itens, como é o caso deste trabalho, o uso de softwares que permitam fórmulas e também um refinamento manual em alguns casos pode ser mais eficaz, se o objetivo é obter um resultado mais amplo. Em muitos casos, os sinônimos não estão bem mapeados nas ontologias e, especialmente envolvendo tradução, muitas vezes existe uma URI para um determinado dado e o mesmo não é encontrado por não estar mapeado como sinônimo, mas ser um sinônimo. Para estes casos, a inteligência humana, nos dias de hoje, ainda é a opção mais confiável e assim foi feito em algumas URIs específicas que não foram possíveis de serem obtidas de forma automática nas bases de dados aqui trabalhadas, requerendo, assim, o tratamento manual de cada uma delas.

5.5 CONVERSÃO PARA TRIPLAS

Por fim, a etapa **Armazenamento e Publicação** do ciclo de vida dos dados conectados e, como visto, foram testados dois programas com bom suporte que oferecem a possibilidade de fazer a conversão de bases de dados para triplas RDF, o Open Refine e o Karma, conforme descrito nas seções 3.1 e 3.2 do Capítulo 3. Entretanto, nenhum deles é automático, como acontece com certos conversores de formatos, e ambos demandam a importação das ontologias e modelagem dos dados, ainda dependendo da inteligência humana para realizar a curadoria das ontologias e modelagem. Embora algumas ferramentas de triplificação sejam capazes de oferecer sugestões de modelagens, elas nem sempre surgem e muitas vezes estão incorretas, fazendo-se necessários todos os passos citados nas seções anteriores a esta.

5.5.1 Triplificação com Open Refine

Inicialmente, a conversão das bases de dados selecionadas para este trabalho seria feita com o Open Refine, devido à simplicidade da instalação e excelente suporte da comunidade. Como o Open Refine é uma ferramenta primariamente de limpeza de dados, processo este já feito nas bases de dados deste trabalho utilizando outra ferramenta, como citado na seção 5.2 deste mesmo Capítulo, restou somente aplicar a modelagem para obter a conversão de dados através do Open Refine. Para realizar a triplificação, o Open Refine é refém de sua extensão para RDF, o que a priori não é um problema, apenas demanda a instalação de um módulo no programa.

Entretanto, é necessário para a triplificação o uso de ontologias, demandando algum tipo de integração com as mesmas, seja através da busca a dados via API ou

importando as ontologias em algum formato de arquivo, como OWL ou qualquer outro que a descreva. O Open Refine não possui a capacidade de integração com nenhum tipo de API, mesmo com a extensão para RDF instalada. Sendo assim, a única maneira de importar as ontologias se dá através da importação de arquivos. Isso pode ser problemático para o caso de ontologias que não disponibilizem quaisquer arquivos, ou mesmo ontologias que disponibilizem somente arquivos que o programa não é capaz de ler, como o caso da ontologia MEDDRA, descrita na seção 4.2.2.2 do Capítulo 4, necessária para uma das bases de dados deste trabalho, por exemplo.

Nenhuma das ontologias utilizadas neste trabalho, apresentadas no Capítulo 4, foram corretamente importadas pelo Open Refine. Algumas ontologias não estavam disponíveis em formatos mapeados pelo programa e outras eram extensas demais para o programa suportar, de modo que a importação, apesar de aparentemente bem sucedida e sem nenhuma mensagem de erro, não resultava na disponibilização dos dados necessários para as bases de dados em questão. Na importação de outras ontologias em formatos teoricamente legíveis pelo programa, apesar da informação de falha na importação, a ferramenta não fornecia a razão desta irregularidade. Por todas essas razões, o Open Refine foi descartado para uso do processo de triplificação.

5.5.2 Triplificação com o Karma

Com a tabela de dados corretamente mapeada após todas as etapas acima realizadas e o descarte do Open Refine como ferramenta de triplificação, o Karma foi escolhido para realizar a geração de triplas. Embora o Karma seja uma ferramenta desenvolvida exclusivamente para triplificar dados, alguns problemas foram encontrados conforme descrito a seguir.

Para o caso de URIs que utilizam identificadores que não contenham o dado original na tabela, como acontece com muitas ontologias que fazem uso de identificadores numéricos concatenados com o prefixo da ontologia (como o MedDRA na URI http://purl.bioontology.org/ontology/MEDDRA/10047862 referente à fraqueza, por exemplo), a associação de células da base de dados original com URIs se torna inviável de ser realizada usando o Karma, pois o processo de associação das células da tabela com suas URIs não poderá ser feito de forma automática. Isso ocorre porque para descobrir de forma automática a URI que deve ser associada com cada célula da tabela, é necessário fazer uma requisição à API da ontologia escolhida para determinada coluna a cada linha que existe naquela coluna, buscando por aquele dado na ontologia, e o Karma não suporta integração com APIs verdadeiramente REST, que representam dados diferentes utilizando URLs diferentes, o que acontece com grande parte das ontologias existentes. Esta é uma funcionalidade ainda em desenvolvimento e que se torna essencial para uma maior automatização do processo de triplificação de dados dentro do Karma, em especial para o caso de bases com grandes volumes de dados, em que é muito custoso fazer manualmente esta associação.

Além da limitação para a integração com APIs, o Karma apresenta uma outra grave limitação, pois ainda que fosse decidido realizar o processo de associação de URIs às células das colunas existentes na base de dados manualmente, para isto, seria necessário importar a ontologia para o Karma, por meio de arquivos. O programa só é capaz de ler alguns formatos, entre eles, principalmente owl, mas nem todas as ontologias estão disponíveis para download neste ou em outros formatos aceitos pelo Karma, o que faz com que precisem ser representadas neste formado, transformando-as. Isto é ainda mais grave porque além de impedir o processo de associação de URIs, impede a própria modelagem em si, pois a construção dos grafos que serão responsáveis internamente, no algoritmo do Karma, por gerar as triplas RDF dependem do modelo construído visualmente pelo usuário, e este só poderá ser construído se a ontologia for importada pelo programa.

Portanto, caso as ontologias que precisam ser usadas não estejam disponíveis em formatos que o programa seja capaz de importar, o uso do Karma para o processo de triplificação se torna muito mais complexo, ainda que a base seja tratada previamente com uso de outras ferramentas e processos, pois seria necessário converter as ontologias necessárias para o formato OWL, e isto não é simples nem trivial.

Outra grave questão enfrentada no Karma é o desempenho. O programa encontra limitações no volume de dados que consegue processar, portanto, para o caso de ontologias que estejam disponíveis em arquivos muito grandes ou mesmo de bases de dados muito volumosas, seu uso se torna complexo, uma vez que o programa apresenta extrema lentidão. Uma opção para contornar esse problema seria dividir a base de dados volumosa em arquivos menores, entretanto, não foi informada pelos desenvolvedores a exata limitação de processamento do programa, o que torna esta possibilidade também experimental e complicada.

Há ainda defeitos contornáveis, como bugs ao utilizar a funcionalidade *Pytransform*, a impossibilidade de salvar os modelos visuais criados antes do programa processar e converter os dados para triplas RDF e ainda a falta de opção de excluir colunas e editar células vazias. Dentre estes, o que mais apresenta risco para o caso de modelagens trabalhosas é o fato de o programa não oferecer a possibilidade de salvar o trabalho de modelagem temporariamente realizado, obrigando o usuário a modelar tudo sem fechar o programa, pois se o usuário decidir gerar as triplas sem a modelagem estar completa, não há a opção de continuar ou alterar aquela mesma modelagem posteriormente, gerando, assim, dados RDF incompletos em relação à base de dados original.

Considerando-se todas essas limitações e falhas, mesmo tendo a tabela das bases minuciosamente limpas e detalhadas, o Karma também foi descartado para triplificar os dados. Esgotando-se as ferramentas de triplificação disponíveis que estão atualizadas e fornecem suporte aos usuários, decidiu-se então realizar a

triplificação de dados de forma semi-automática através do uso de ferramentas que não têm como objetivo realizar este processo.

5.5.3 Triplificação via SQL Developer

Como as possibilidades de triplificação das bases de dados com o uso de ferramentas únicas foram esgotadas, decidiu-se por realizar este processo de forma semi-automática, usando diversas ferramentas para diferentes etapas do processo. Havendo as duas bases de dados limpas, tratadas e com dados e suas respectivas URIs bem mapeadas, o procedimento de geração de triplas torna-se relativamente menos complexo de ser feito de forma não-automática.

Seria necessário desenvolver, primeiramente, um padrão para estruturar as triplas e em seguida uma forma de aplicar este padrão em cada ocorrência de dados na tabela fazendo-se fundamental o uso de uma ferramenta ou programa que possa percorrer a base e aplicar estes modelos. Sendo assim, foi escolhido incluir as bases de dados tratadas e limpas em uma tabela SQL e executar consultas que retornassem as triplas dispostas de forma a obedecer o modelo pré-estabelecido.

A seguir o padrão é apresentado e retratado o processo de geração de triplas para cada uma das bases citadas neste trabalho.

5.5.3.1 Triplas de Agrotóxicos Versus Sintomas

Com a base de dados *Relação Entre Agrotóxicos e Sintomas Crônicos e Agudos* pronta para passar pela etapa de geração de triplas, como pode ser visto na Figura 25, a mesma foi importada em uma tabela de um banco de dados exatamente como estruturada após seu tratamento.

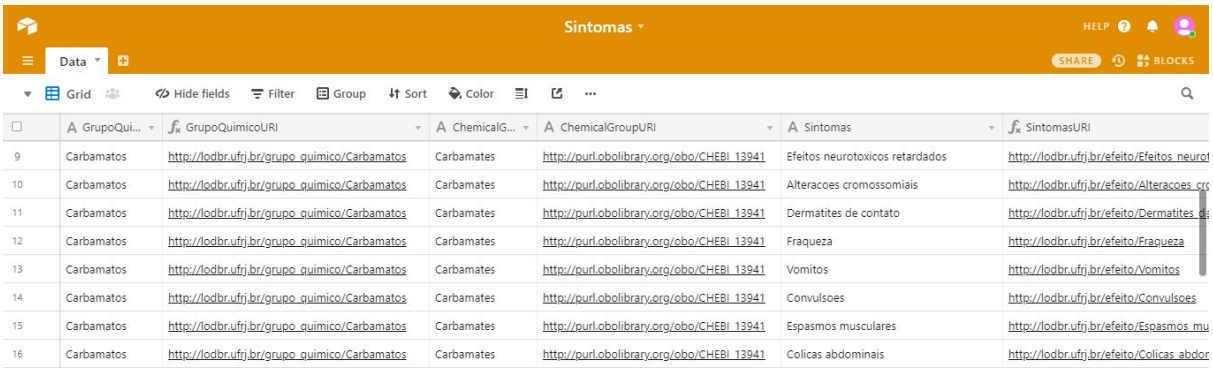


Figura 25 - Tabela de Agrotóxicos Versus Sintomas após as etapas de limpeza, tradução e adaptação

Antes de o processo em si ser executado, foi definido o seguinte padrão de triplas a serem geradas:

1. ChemicalEntityURI *is_a* CheBI/AGROTOXICO
2. ChemicalEntityURI *hasClass* ClassURI
3. ClassURI *label* Class@en
4. ChemicalEntityURI *hasComponent* ChemicalGroupURI
5. ChemicalGroupURI *label* ChemicalGroup@en
6. ChemicalEntityURI *causes* SymptomURI
7. SintomaURI *a* DBpedia/Symptom
8. SymptomURI *owl:sameAs* SintomaURI

Como informado na seção 5.1.1 deste capítulo, uma Entidade Química consiste em um agrotóxico composto pelas informações de Grupo Químico e Classe, para posterior associação direta desta entidade com os agrotóxicos existentes. Tendo em vista essa informação, segue uma explicação do padrão de triplas a serem geradas:

A primeira tripla refere-se à definição de uma Entidade Química como um agrotóxico, ou seja, um agrotóxico. A segunda e terceira tripla definem a Classe que

esta Entidade Química atua no controle de pragas. A quarta e quinta tripla definem ao Grupo Químico que a Entidade Química se caracteriza. As três última triplas, sexta, sétima e oitava, definem qual Sintoma é causado por esta Entidade Química. Como um agrotóxico causa mais de um sintoma, este trio de triplas repete-se para cada sintoma causado.

Sabendo do modelo descrito, foi executada uma consulta, apresentada no Anexo 3 deste trabalho, que percorria a tabela e, para cada linha de informações, foi gerado este conjunto de triplas através do uso do SGBD (sistema de gerenciamento de banco de dados).

Segue abaixo um exemplo de triplas correspondentes à seguinte informação: A *Entidade Química* composta pelo *Grupo Químico* Dinitroferols e *Classe* Herbicida causa o *Sintoma* Hipertermia.

1. <http://lodbr.ufrj.br/chemical_group_class/Dinitroferols_Herbicides> is_a <http://purl.obolibrary.org/obo/CHEBI_25944>.

2. <http://lodbr.ufrj.br/chemical_group_class/Dinitroferols_Herbicides> <http://aims.fao.org/aos/agrontology#hasClass> <http://purl.obolibrary.org/obo/CHEBI_24527>.

3. <http://purl.obolibrary.org/obo/CHEBI_24527> <http://www.w3.org/2000/01/rdf-schema#label> "Herbicides"@en.

4. <http://lodbr.ufrj.br/chemical_group_class/Dinitroferols_Herbicides> <http://aims.fao.org/aos/agrontology#hasComponent> <http://lodbr.ufrj.br/chemical_group/Dinitroferols>.

5. <http://lodbr.ufrj.br/chemical_group/Dinitroferols> <http://www.w3.org/2000/01/rdf-schema#label> "Dinitroferols"@en.

6. <http://lodbr.ufrj.br/chemical_group_class/Dinitroferols_Herbicides> <http://aims.fao.org/aos/agrontology#causes> <http://purl.bioontology.org/ontology/MEDDRA/10020843>.

7. <http://lodbr.ufrj.br/efeito/Hipertermia> a <http://dbpedia.org/page/Symptom>

8. <http://purl.bioontology.org/ontology/MEDDRA/10020843> <owl:sameAs> <http://lodbr.ufrj.br/efeito/Hipertermia>.

O mesmo exemplo representado em grafo é apresentado na Figura 26.

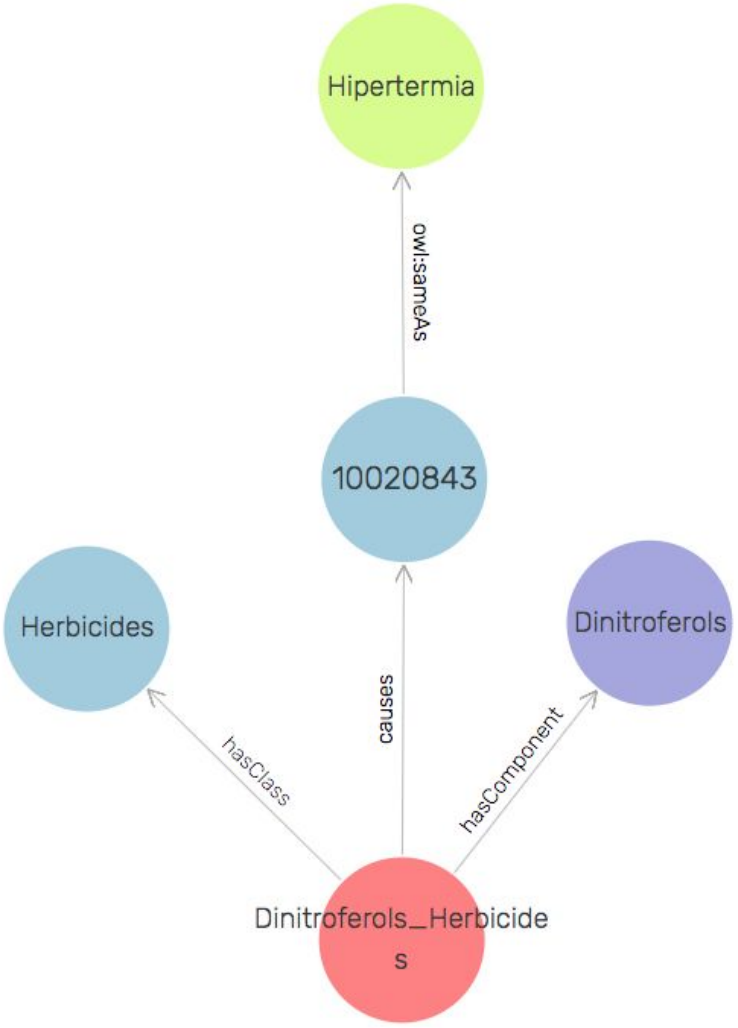
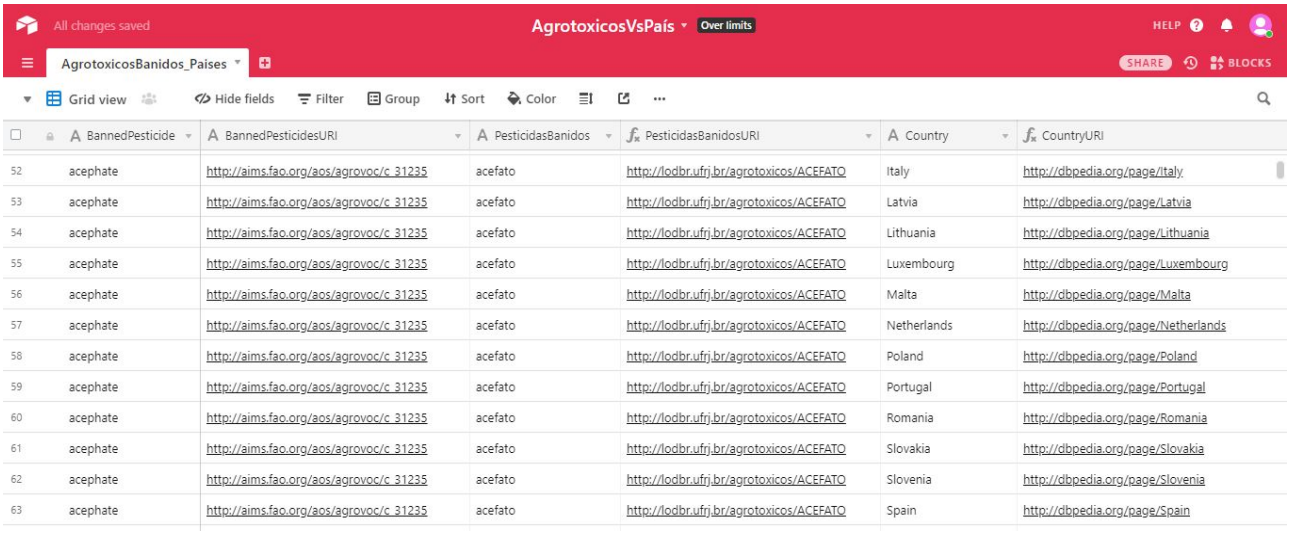


Figura 26 - Triplificação de Entidade Química causa Sintoma em Grafo gerado pelo GraphDB

5.5.3.2 Triplas de Agrotóxicos Banidos Versus Países

Após a limpeza, tradução e adaptação da base *Consolidated List of Banned Pesticides,* obteve-se uma planilha de dados mais clara e pronta para triplificação conforme é possível ver na Figura 27. O mesmo processo de modelagem de dados

foi necessário nesta etapa antes da triplificação, porém algumas exceções foram tratadas.



	BannedPesticide	BannedPesticidesURI	PesticidasBanidos	PesticidasBanidosURI	Country	CountryURI
52	acephate	http://aims.fao.org/aos/agrovoc/c_31235	acefato	http://lodbr.ufrj.br/agrotoxicos/ACEFATO	Italy	http://dbpedia.org/page/Italy
53	acephate	http://aims.fao.org/aos/agrovoc/c_31235	acefato	http://lodbr.ufrj.br/agrotoxicos/ACEFATO	Latvia	http://dbpedia.org/page/Latvia
54	acephate	http://aims.fao.org/aos/agrovoc/c_31235	acefato	http://lodbr.ufrj.br/agrotoxicos/ACEFATO	Lithuania	http://dbpedia.org/page/Lithuania
55	acephate	http://aims.fao.org/aos/agrovoc/c_31235	acefato	http://lodbr.ufrj.br/agrotoxicos/ACEFATO	Luxembourg	http://dbpedia.org/page/Luxembourg
56	acephate	http://aims.fao.org/aos/agrovoc/c_31235	acefato	http://lodbr.ufrj.br/agrotoxicos/ACEFATO	Malta	http://dbpedia.org/page/Malta
57	acephate	http://aims.fao.org/aos/agrovoc/c_31235	acefato	http://lodbr.ufrj.br/agrotoxicos/ACEFATO	Netherlands	http://dbpedia.org/page/Netherlands
58	acephate	http://aims.fao.org/aos/agrovoc/c_31235	acefato	http://lodbr.ufrj.br/agrotoxicos/ACEFATO	Poland	http://dbpedia.org/page/Poland
59	acephate	http://aims.fao.org/aos/agrovoc/c_31235	acefato	http://lodbr.ufrj.br/agrotoxicos/ACEFATO	Portugal	http://dbpedia.org/page/Portugal
60	acephate	http://aims.fao.org/aos/agrovoc/c_31235	acefato	http://lodbr.ufrj.br/agrotoxicos/ACEFATO	Romania	http://dbpedia.org/page/Romania
61	acephate	http://aims.fao.org/aos/agrovoc/c_31235	acefato	http://lodbr.ufrj.br/agrotoxicos/ACEFATO	Slovakia	http://dbpedia.org/page/Slovakia
62	acephate	http://aims.fao.org/aos/agrovoc/c_31235	acefato	http://lodbr.ufrj.br/agrotoxicos/ACEFATO	Slovenia	http://dbpedia.org/page/Slovenia
63	acephate	http://aims.fao.org/aos/agrovoc/c_31235	acefato	http://lodbr.ufrj.br/agrotoxicos/ACEFATO	Spain	http://dbpedia.org/page/Spain

Figura 27 - Base de agrotóxicos banidos e países após etapa de limpeza e adaptação

Assim como para a base de dados anterior, foi definido o seguinte padrão de triplas a serem geradas para esta base:

1. Agrovoc/Pesticide *label* Pesticide@en
2. Agrovoc/Pesticide *proibidoEm* DBPEDIA/Country

Como descrito na etapa de adaptação, todo dado necessita de uma URI que o represente, entretanto, nem todos os agrotóxicos listados na primeira coluna da base de dados foram encontrados no AGROVOC e, por essa razão, foi necessário criar uma URI própria em português.

Sabendo dessas duas naturezas de URI distintas, o padrão estabelecido foi:

- Para os agrotóxicos que possuíam identificação no AGROVOC, como o ACEPHATE, foi gerada uma tripla para a definição da label e outra tripla para cada país no qual ele é proibido. Segue o exemplo a seguir:

1. <http://aims.fao.org/aos/agrovoc/c_31235>
   <http://www.w3.org/2000/01/rdf-schema#label>
   “acephate”@en.
2. <http://aims.fao.org/aos/agrovoc/c_31235>
   <http://lodbr.ufrj.br/agrotoxicos/propriedade/proibidoEm>
   <http://dbpedia.org/page/Austria>.

- Para os agrotóxicos que não foram encontrados no AGROVOC e demandaram a criação da URI própria, a primeira tripla refere-se a definição da nova URI como um agrotóxico, na ontologia ChEBI, e a segunda tripla também para cada país no qual este agrotóxico é banido, da mesma forma que o primeiro caso descrito.

1. <http://lodbr.ufrj.br/agrotoxicos/ACROLEINA_(2-PROPENAL)>
   a
   <http://purl.obolibrary.org/obo/CHEBI_25944>.
2. <http://lodbr.ufrj.br/agrotoxicos/ACROLEINA_(2-PROPENAL)>>
   <http://lodbr.ufrj.br/agrotoxicos/propriedade/proibidoEm>
   <http://dbpedia.org/page/Saudi_Arabia>.

Um ponto importante de ser ressaltado no processo de triplificação desta base em especial é que, diferentemente do processo da base anterior, não foi possível utilizar a consulta ao SGBD porque, por a base ser extensa - cerca de 3 mil linhas, a saída da consulta estourava o máximo suportado pelo buffer do SGBD. Para contornar este problema de estouro do buffer, a solução foi inserir as triplas linha a linha em uma tabela auxiliar. Como as triplas foram geradas em uma ordem pré-definida, no momento da exportação fez-se necessário o uso de uma variável controladora para que não se perdesse a ordem da geração das triplas. Por fim, para obter essas triplas, a tabela auxiliar de triplas foi exportada para um arquivo texto.

6 EXEMPLO DE UTILIZAÇÃO DOS DADOS TRIPLIFICADOS

6.1 CONSULTA E EXPLORAÇÃO DE DADOS

Como forma de evidenciar importância da conversão das bases de dados para triplas RDF realizada neste trabalho apresenta-se a visualização e exploração das triplas geradas, fazendo uso da ferramenta GraphDB, descrita na seção 3.5.1 do Capítulo 3. Como as bases de dados escolhidas para este trabalho não se interligam diretamente, visando enriquecer as bases de dados já obtidas no trabalho "Extração de dados de fontes textuais: uma abordagem para enriquecimento de dados abertos interligados" [TEIXEIRA, 2018], fez-se, inicialmente, a importação individual de cada arquivo RDF gerado neste trabalho.

O primeiro passo consiste em garantir que a importação no GraphDB aconteça com sucesso (Figura 28). Existe uma série de verificações disponíveis durante a importação, que podem ou não ser executadas, a definir em configurações avançadas na importação (Figura 29). Como forma inicial de verificação das triplificações semi-automáticas realizadas para as bases deste trabalho, todos os recursos disponíveis no GraphDB foram utilizados, para garantir que a importação tenha ocorrido com sucesso, e consequentemente que os arquivos RDF gerados estivessem corretos e viáveis de serem explorados através da ferramenta. Vale lembrar, conforme mencionado na seção 3.5.1, que o GraphDB segue as recomendações da W3C, de modo que garantir que a importação bem sucedida no GraphDB representa também seguir as recomendações da W3C no quesito sintaxe.

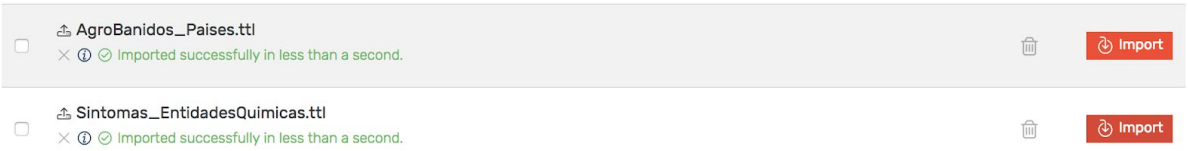


Figura 28 - Importações bem sucedidas dos arquivos RDF gerados neste trabalho

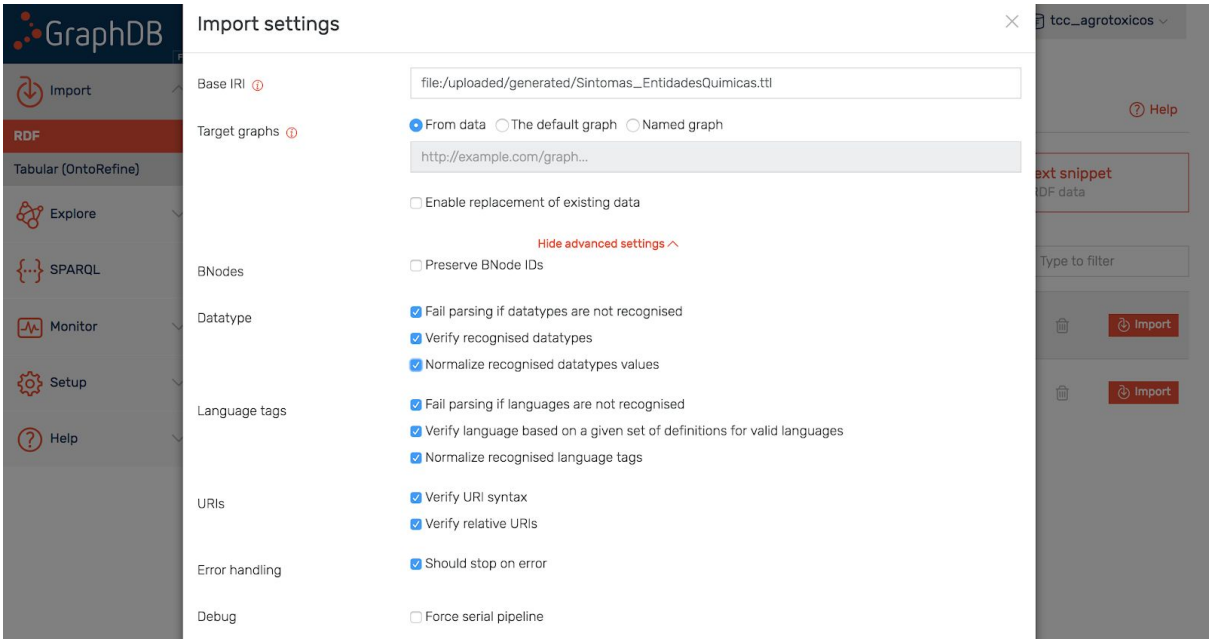


Figura 29 - Recursos avançados disponíveis durante a importação de arquivos RDF no GraphDB

Uma vez feita a importação, o passo seguinte consiste em explorar os arquivos RDF obtidos neste trabalho, um a um, por meio de consultas SPARQL que visam buscar informações básicas existentes nas bases de dados originais e que devem também, por consequência, constar no arquivo RDF obtido. Através dessas consultas básicas, é possível se certificar de que o conteúdo dos arquivos RDF estejam coerentes com o esperado. Além disso, é possível obter algumas informações interessantes, mesmo importando somente um dos arquivos gerados neste trabalho, sem ainda integrá-lo com outros arquivos RDF acerca do assunto.

Para o caso da base de dados *Consolidated List of Banned Pesticides*, apresentada na seção 4.2.1.2 do Capítulo 4, é bastante visível o ganho com a

exploração destes dados após o processo realizado neste trabalho. Originalmente esta base era de difícil interpretação humana, e demandava um grande trabalho de limpeza e transformação para que pudesse ser processada automaticamente e ter seus dados extraídos a fim de realizar algumas inferências ou mesmo consultas básicas. Após a triplificação da base e importação do arquivo RDF gerado no GraphDB, foram realizadas algumas consultas básicas de validação da triplificação. Uma das consultas está representada na Figura 30, com o objetivo de obter todos os agrotóxicos proibidos no Brasil.

```
1  prefix dbpediacountries:  <http://dbpedia.org/page/>
2  prefix lodbrproperties:  <http://lodbr.ufrj.br/agrotoxicos/propriedade/>
3  select * where {
4      ?s lodbrproperties:proibidoEm dbpediacountries:Brazil .
5  } limit 100
6
```

Figura 30 - Consulta realizada no GraphDB, para obter agrotóxicos proibidos no Brasil

O resultado obtido, no formato de tabela, está exibido na Figura 31. É possível saber quantos são os agrotóxicos proibidos no Brasil e ainda quais são eles.

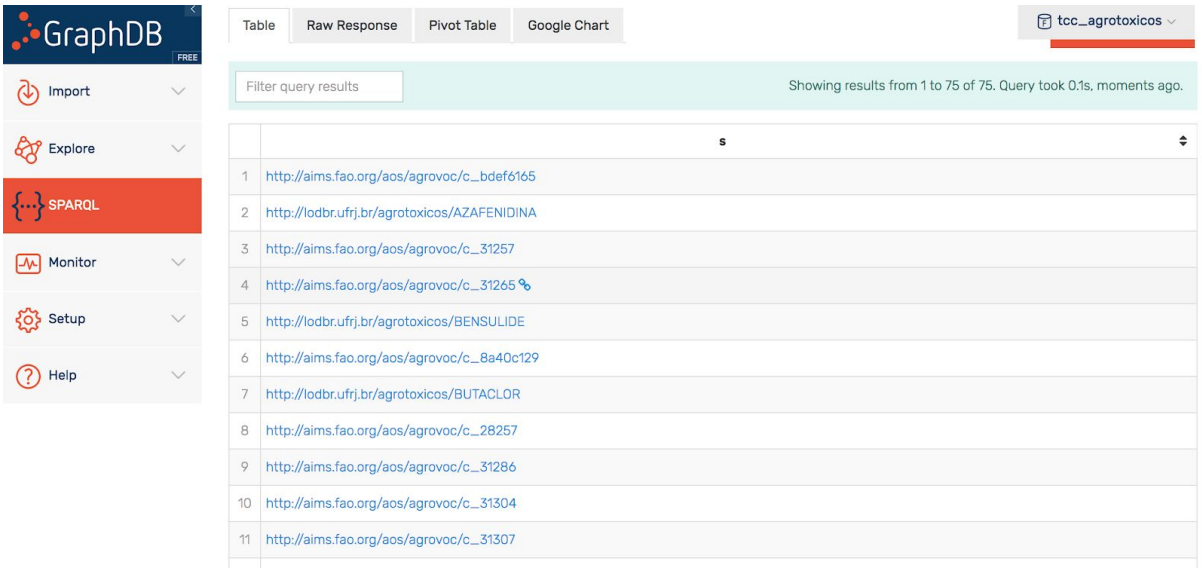


Figura 31 - Resultado de consulta realizada no GraphDB: todos os agrotóxicos proibidos no Brasil

O mesmo resultado, no formato visual de grafo gerado pelo GraphDB também é exibido na Figura 32.

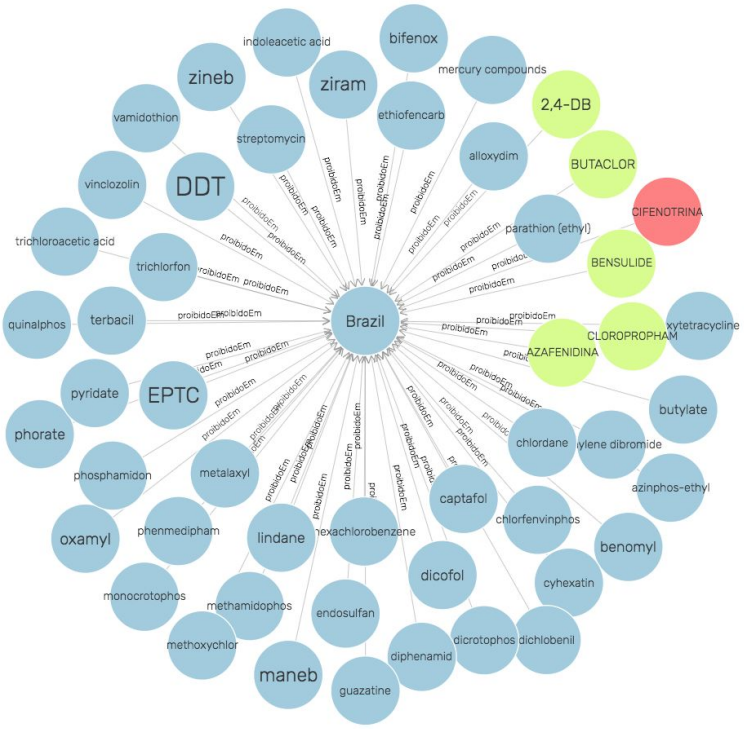


Figura 32 - Resultado visual em grafos obtido no GraphDB: todos os agrotóxicos proibidos no Brasil

Com o objetivo de validar o arquivo RDF gerado a partir da base de dados *Relação entre Agrotóxicos e Sintomas Agudos e Crônicos*, já apresentada na seção 4.2.1.1 do Capítulo 4, foi feita uma consulta em SPARQL, ilustrada na Figura 33, visando obter todas as entidades químicas associadas a uma determinada doença, no caso, Câncer. Os resultados são exibidos na Figura 34.

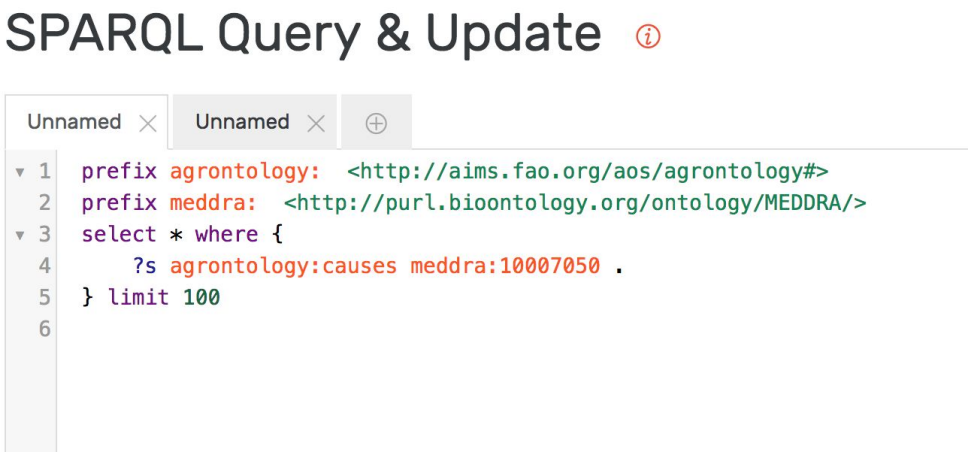


Figura 33 - Consulta realizada no GraphDB para obter entidades químicas relacionadas a Câncer

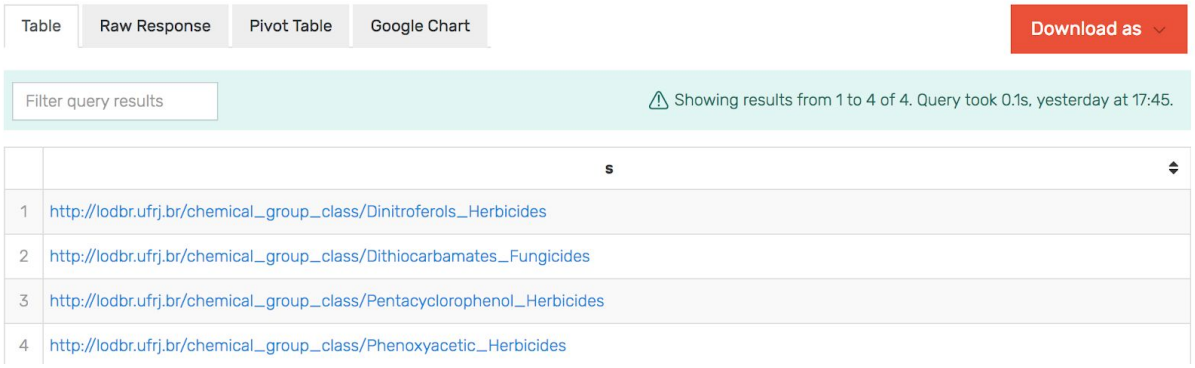


Figura 34 - Resultado de consulta realizada no GraphDB: entidades químicas relacionadas a Câncer

A disponibilidade destas informações também é útil para associar as doenças presentes na base de dados com agrotóxicos utilizados pela indústria, obtida em [TEIXEIRA, 2018]. Os grupos químicos modelados para a base *Relação entre Agrotóxicos e Sintomas Agudos e Crônicos*, selecionada neste trabalho, se relacionam diretamente com agrotóxicos existentes em bases de dados obtidas em [TEIXEIRA, 2018], de modo que, conforme explicado na seção 4.2 do Capítulo 4, este foi justamente o objetivo da criação e utilização de URIs que modelam estes grupos. Na Figura 35 é possível visualizar um pequeno fragmento do resultado de integração da base *Relação entre Agrotóxicos e Sintomas Agudos e Crônicos*, gerada neste trabalho, com bases geradas em [TEIXEIRA, 2018]: é possível visualizar o *Organofosforado*, por exemplo, um *Grupo Químico* com URI em comum

entre as bases, que integradas passam a exibir não só os agrotóxicos que pertencem a este grupo químico (em vermelho na figura), objetivo alcançado em [TEIXEIRA, 2018], mas também diversos sintomas causados por estes agrotóxicos (em amarelo), objetivo alcançado neste trabalho.

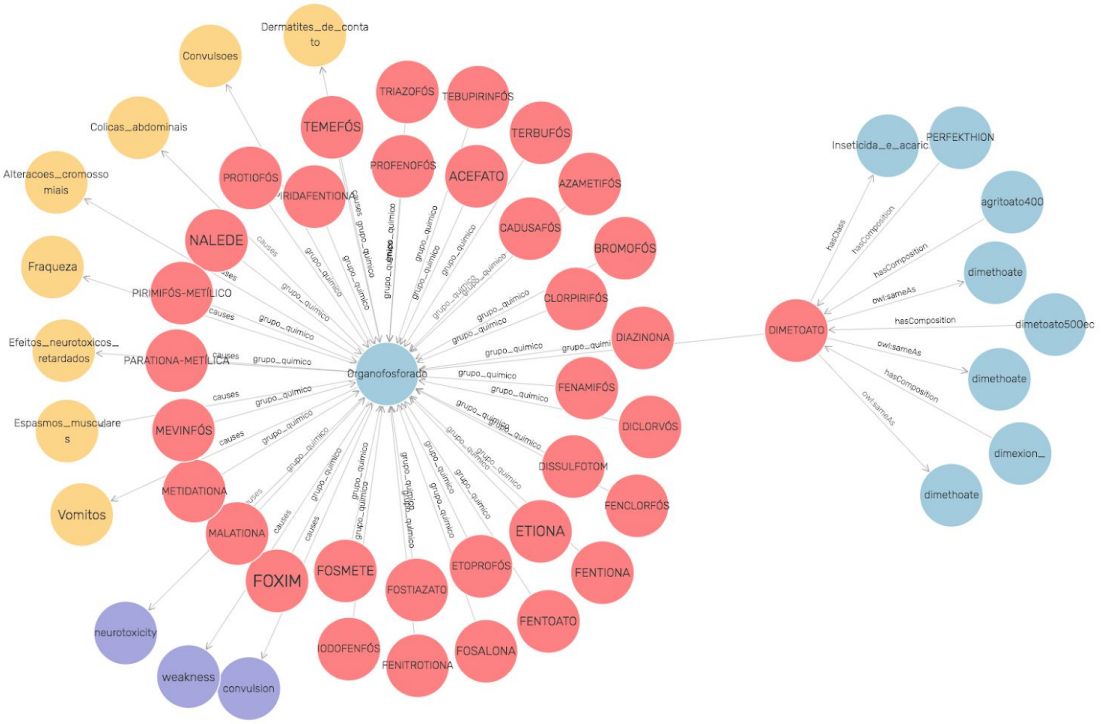


Figura 35 - Exemplificação do resultado visual da integração das bases, com Organofosforado como pivô

Através deste mesma integração é possível visualizar quais agrotóxicos que são proibidos em um determinado país e quais sintomas este mesmo agrotóxico pode causar, como é mostrado na Figura 36. A relação é possível por meio da relação entre Entidade Química que corresponde a um agrotóxico proibidos ou entre Classes e/ou Grupo Químicos comuns em ambas as bases.

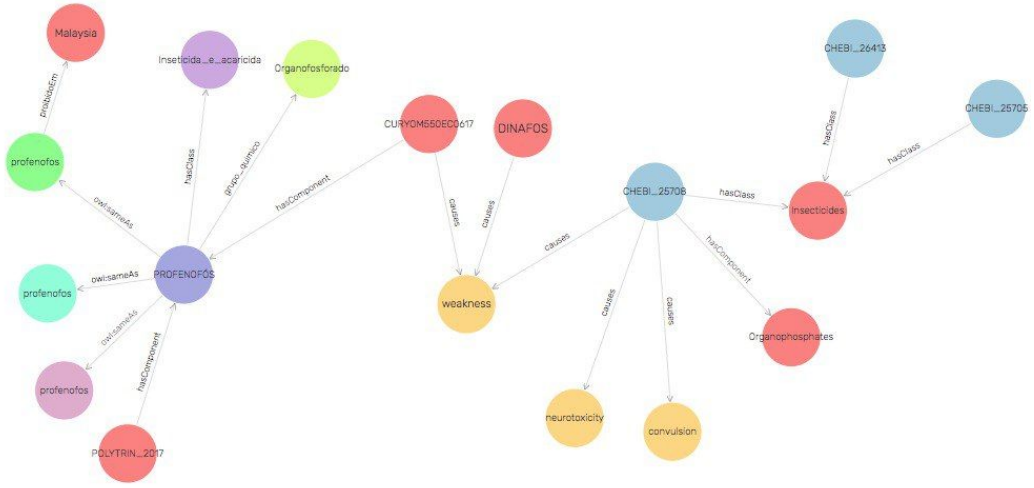


Figura 36 - Exemplo de agrotóxico banido em um determinado país e seus sintomas causados

6.2 MENSURAÇÃO DOS RESULTADOS

Para explicitar o quanto a triplificação de dados é verborrágica e demanda um certo poder de processamento, com potencial de gerar um arquivo volumoso após a triplificação, foi feito um comparativo entre a quantidade inicial de linhas na base de dados com a quantidade após a geração das triplas correspondentes. Cada tripla corresponde a uma linha no arquivo RDF gerado.

A base de dados Relação entre Agrotóxicos e Sintomas Agudos e Crônicos tinha 64 linhas após a etapa de limpeza, que aumentou o número de linhas em relação a original para que a base estivesse propícia para o processo de triplificação, necessidade explicada na seção 3.2 do Capítulo 3. Após a triplificação, foram geradas 508 triplas.

Já a base de dados *Consolidated List of Banned Pesticides* possuía, após a limpeza, 2763 linhas, e o arquivo RDF gerado após o processo de triplificação possui 3115 triplas.

7 CONCLUSÃO

7.1 CONSIDERAÇÕES FINAIS

Este trabalho foi projetado tendo em vista atender aos seguintes propósitos: (i) realizar um experimento de geração de dados conectados a partir de dados existentes acerca dos agrotóxicos no Brasil e no mundo; (ii) experimentar e avaliar ferramentas para triplificação de dados; (iii) descrever em detalhe o processo de triplificação das bases de dados escolhidas.

Como há grande disponibilização de dados sobre os agrotóxicos comercializados no Brasil e no mundo e suas influências na saúde humana, este trabalho procurou discutir a utilização de ferramentas disponíveis em um caso de uso real, com grande volume de dados, de modo a reunir mais informações e disponibilizá-las para que outras pesquisas possam ser feitas idealizando suportar a tomada de decisões de políticas públicas que tenham como foco também contribuir para a qualidade de vida da população.

Para que esses objetivos fossem alcançados, foram utilizadas bases de dados que contribuem com o tema da pesquisa e tratadas para que estivessem aptas a passar pelo processo de geração de triplas RDF. As bases foram escolhidas visando complementar o trabalho "Extração de dados de fontes textuais: uma abordagem para enriquecimento de dados abertos interligados" [TEIXEIRA, 2008], de modo que os dados tratados neste trabalho sejam integrados e complementares aos dados obtidos neste outro trabalho.

Após a escolha e o tratamento das bases ser realizada, foi realizado o estudo das ferramentas de triplificação disponíveis, entendendo seus pontos fortes e suas limitações e, por fim, compreendendo que no contexto desta pesquisa, nenhuma das ferramentas permissíveis seria capaz de lidar com o volume de dados e ontologias

extensas que seria necessário. Desta forma, tornou-se necessário triplificar de maneira semi-automática e própria para atingir a proposta da pesquisa.

Durante todo o desenvolvimento desta monografia, houve a preocupação em seguir o esquema de implementação das 5 estrelas de dados abertos, proposta por Tim Berners-Lee [BERNERS-LEE, 2001]. Os dados aqui estruturados atingem o nível da quarta estrela, onde estão disponibilizados através de URIs e prontos para atingirem a quinta estrela conectando-se a mais informações a respeito dos agrotóxicos, obtidas em "Extração de dados de fontes textuais: uma abordagem para enriquecimento de dados abertos interligados" [TEIXEIRA, 2018]. Em todo o processo de modelagem de dados houve a intenção de padronizar os dados de forma que, uma vez dispostos na Web, os mesmos contivessem propriedades que proporcionam a ligação com dados de outros trabalhos.

Por fim, conclui-se que a partir de todo o processo percorrido até a disponibilização dos dados propostos, alcançou-se as intenções de produzir informações relevantes sobre o uso de agrotóxicos no país e fornecer um panorama sobre os meios de triplificação de dados atualmente existentes.

7.2 DIFICULDADES

As principais dificuldades encontradas para viabilização deste trabalho evidenciam o fato de que triplificar está longe de ser uma tarefa banal, especialmente se depender das características das bases de dados definidas. Inclusive a modelagem necessária para tornar o processo viável pode ser bem complexa dependendo da combinação da base e do conhecimento prévio sobre o assunto, demandando possivelmente interações com especialistas do domínio.

Ainda em relação ao processo de triplificação, outra dificuldade é a disponibilidade de ontologias que não são feitas de forma padronizada. Algumas ontologias podem ser oferecidas no formato OWL, outras em TTL e algumas não

disponibilizam seus arquivos, apenas uma API, para que através da mesma se obtenham as informações necessárias. Apesar disso, ainda faltam ontologias que capazes de descrever muitos dados presentes em várias bases, e para contornar este problema criar uma ontologia própria é algo bem complexo. Muitas vezes, como no caso deste trabalho, é necessária a criação de URIs que não possuam uma descrição ainda disponível na Web e isso não é ideal, pois fere um dos princípios da disponibilização de dados Web semânticos proposta por Tim Berners-Lee, explicitada no Capítulo 2.

Outro fator que adicionou complexidade a este trabalho foi a falta de ferramentas que suportem bem o processo de triplificação para dados mistos e massivos em grande escala. Embora uma ferramenta se descreva como ideal e completa, pode haver defeitos nas próprias funcionalidades disponíveis como na importação de ontologias complexas ou no suporte de bases muito extensas.

A última dificuldade foi encontrada devido ao tamanho das bases escolhidas. Triplificar bases muito grandes por meio de consultas SQL requer um cuidado na parte de suporte do programa usado, como foi o caso da base dados *Consolidated List of Banned Pesticides* a qual foi necessário a inserção numa outra tabela de dados porque o máximo de tamanho do buffer de saída padrão do SQL Developer não suportou gerar tantos caracteres.

7.3 TRABALHOS FUTUROS

O principal eixo para o desenvolvimento de trabalhos futuros é popular as informações contidas nas URIs que precisaram ser criadas exclusivamente para o este trabalho por não haver informações dispostas nas ontologias já existentes. É necessário que essas URIs retornem mais informações a respeito do dado que elas descrevem a apontam para que essas informações possam ser utilizadas futuramente para outras pesquisas.

Uma sugestão é a triplificação de mais dados do portal para que essa rede de informações acerca dos agrotóxicos possa ser cada vez mais completa e ampla, podendo assim, permitir a resposta de perguntas progressivamente mais relevantes. Seria igualmente pertinente usar outras ferramentas de triplificação destes dados como objetivo de realizar mais experimentações de ferramentas disponíveis no mercado. Nesta mesma linha, recomenda-se a experimentação de ferramentas para triplificação específicas para triplas a partir de bases relacionais baseada em R2RML (RDB to RDF *mapping language*) - uma linguagem para expressar esses mapeamentos de RDB para RDF, tornando o processo de montagem das consultas mais simples.

Também pode-se triplificar dados da saúde e integrá-los com os dados referentes a agrotóxicos atuais, visando extrair informações mais precisas e polêmicas, como o número de doenças por ano e a quantidade do uso de agrotóxico nos países. Assim, seria possível inferir se a liberação de agrotóxicos em cada país pode estar causando mais doenças à população que consome estes próprios agrotóxicos, interligando, ainda, estes dados obtidos com bases de dados externas inclusive utilizadas neste trabalho, como DBPEDIA, Agrovoc etc.

REFERÊNCIAS

TEIXEIRA, K. *Extração de dados em fontes textuais: uma abordagem para enriquecimento de dados abertos conectados,* 2018.

TYGEL, A. F. ; GONÇALVES, L. G. ; SANTOS, M. ; MARQUES, G. ; CAMPOS, M.L.M. *Informação para Ação: Desenvolvimento de um Portal de Dados Abertos Sobre Agrotóxicos*, 2015

GUIZZARDI, G. *Desenvolvimento para e com reuso: Um estudo de caso no domínio de vídeo sob demanda*. Master's thesis, Universidade Federal do Espírito Santo, 2000.

EQUIPE GT, *LinkedDataBR: Exposição, Compartilhamento e Conexão de Recursos de Dados Abertos na Web (Linked Open Data) - Manual do Usuário,* 2011.

ANVISA, "Relatório das análises de amostras monitoradas no período de 2013 a 2015". Disponível em:<http://portal.anvisa.gov.br/documents/111215/0/Relat%C3%B3rio+PARA+2013-2015_VERS%C3%83O-FINAL.pdf/494cd7c5-5408-4e6a-b0e5-5098cbf759f8>. Acesso em 01/08/2018.

BERNERS-LEE, T. ; HENDLER, J. ; and LASSILA , O.. *The Semantic Web* https://www-sop.inria.fr/acacia/cours/essi2006/Scientific%20American_%20Feature%20Article_%20The%20Semantic%20Web_%20May%202001.pdf

ABRASCO. "Dossiê Abrasco: um alerta sobre os impactos dos agrotóxicos na saúde". Disponível em: <https://www.abrasco.org.br/dossieagrotoxicos/wp-content/uploads/2013/10/DossieAbrasco_2015_web.pdf>. Acesso em 28/07/2018.

AIMS. "Agrovoc". Disponível em: <http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus>. Acesso em 30/07/2018.

BERNERS-LEE, Tim. "Linked data". Disponível em: <https://www.w3.org/DesignIssues/LinkedData.html>. Acesso em 29/07/2018.

W3C. "Resource Description Framework (RDF)". Disponível em: <https://www.w3.org/RDF/>. Acesso em 13/08/2018.

FAO. "AGROVOC Guidelines". Disponível em: <http://aims.fao.org/standards/agrovoc/editorial_guidelines>. Acesso em 20/09/2018.

DBPedia. "DBPedia Wiki". Disponível em: <https://pt.wikipedia.org/wiki/DBpedia>. Acesso em 02/09/2018.

GraphDB. “Ontotext GraphDB’s Documentation”. Disponível em: <http://graphdb.ontotext.com/documentation/standard/index.html>. Acesso em 02/09/2018.

Kettle (Pentaho Data Integration). Disponível em: <https://www.hitachivantara.com/en-us/products/big-data-integration-analytics/pentaho-data-integration.html>. Acesso em 06/10/2018.

ETL4LOD. Disponível em: <https://github.com/rogersmendonca/ETL4LOD>. Acesso em 06/10/2018.

DICIO. “Dicionário Online de Português”. Disponível em: <https://dicio.com.br/>. Acesso em 10/09/2018.

W3C BRASIL. "Web Semântica". Disponível em: <http://www.w3c.br/Padroes/WebSemantica>. Acesso em 18/08/2018.

HAUSENBLAS, Michael. "5-star Open Data". Disponível em: <https://5stardata.info/en/>. Acesso em: 22/07/2018.

KNOBLOCK, C., SZEKELY, P., AMBITE, J., GOEL A., GUPTA S., LERMAN K., MUSLEA M., TAHERIYAN, M., MALLICK, P. "Semi-Automatically Mapping Structured Sources into the Semantic Web". Disponível em: <https://www.isi.edu/integration/papers/knoblock12-eswc.pdf >. Acesso em: 25/07/2018.

GRECO. "Portal de Dados Abertos sobre Agrotóxicos". Disponível em: <http://dados.contraosagrotoxicos.org/>. Acesso em: 23/07/2018.

Câmara dos Deputados. "PL 6299/2002". Disponível em: <https://www.camara.gov.br/proposicoesWeb/fichadetramitacao?idProposicao=46249>. Acesso em: 13/08/2018.

NICBR, "Guia de Web Semântica". Disponível em: <https://nic.br/media/docs/publicacoes/13/Guia_Web_Semantica.pdf>. Acesso em: 03/08/2018.

"Open Refine". Disponível em: <http://openrefine.org/>. Acesso em: 01/08/2018.

University of Southern California. "Karma. A Data Integration Tool". Disponível em: <http://usc-isi-i2.github.io/karma/>. Acesso em 30/07/2018.

W3C. "Turtle". Disponível em: <https://www.w3.org/TR/turtle/>. Acesso em: 17/08/2018.

W3C. "N-Triples". Disponível em: <https://www.w3.org/TR/n-triples/>. Acesso em: 15/08/2018.

W3C. "XML Syntax". Disponível em: <https://www.w3.org/TR/rdf-syntax-grammar/>. Acesso em: 15/08/2018.

PORTAL, "Portal de Dados Abertos sobre Agrotóxicos". Disponível em: <http://dados.contraosagrotoxicos.org/>. Acesso em: 15/07/2018.

ANEXOS

1. Consulta de extração do CSV a partir da *Consolidated List of Banned Pesticides*

```
DECLARE
BEGIN
FOR X IN (SELECT * FROM SIEBEL.TCC)
LOOP
    IF X.ANTIGUA_BARBUDA IS NOT NULL THEN

DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Antigua_and_Barbuda;');
    END IF;
    IF X.ARMENIA IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Armenia;');
    END IF;
    IF X.AUSTRALIA IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Australia;');
    END IF;
    IF X.BANGLADESH IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Bangladesh;');
    END IF;
    IF X.BENIN IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Benin;');
    END IF;
    IF X.BRAZIL IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Brazil;');
    END IF;
    IF X.BULGARIA IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Bulgaria;');
    END IF;
    IF X.BURKINA_FASO IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Burkina_Faso;');
    END IF;
    IF X.CAMBODIA IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Cambodia;');
    END IF;
    IF X.CAMEROON IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Cameroon;');
    END IF;
    IF X.CANADA IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Canada;');
    END IF;
```

```
IF X.CAPO_VERDE IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Capo_Verde;');
END IF;
IF X.CHAD IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Chad;');
END IF;
IF X.CHILE IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Chile;');
END IF;
IF X.CHINA IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';China;');
END IF;
IF X.COLOMBIA IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Colombia;');
END IF;
IF X.COSTA_RICA IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Costa_Rica;');
END IF;
IF X.COTE_DIVOIRE IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Cote_Divoire;');
END IF;
IF X.CUBA IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Cuba;');
END IF;
IF X.DOMINICAN_REPUBLIC IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Dominican_Republic;');
END IF;
IF X.ECUADOR IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Ecuador;');
END IF;
IF X.EU IS NOT NULL AND <> '?' THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Austria;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Belgium;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Bulgaria;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Croatia;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Cypru;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Czech_Republic;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Denmark;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Estonia;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Finland;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';France;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Germany;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Greece;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Hungary;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Ireland;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Italy;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Latvia;');
```

```
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Lithuania;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Luxembourg;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Malta;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Netherlands;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Poland;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Portugal;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Romania;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Slovakia;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Slovenia;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Spain;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Sweden;');
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';United_Kingdom;');
END IF;
IF X.FIJI IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Fiji;');
END IF;
IF X.GAMBIA IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Gambia;');
END IF;
IF X.GUINEA IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Guinea;');
END IF;
IF X.GUINEA_BISSAU IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Guinea_Bissau;');
END IF;
IF X.GUYANA IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Guyana;');
END IF;
IF X.HUNGARY IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Hungary;');
END IF;
IF X.INDIA IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';India;');
END IF;
IF X.INDONESIA IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Indonesia;');
END IF;
IF X.IRAQ IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Iraq;');
END IF;
IF X.IRAN IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Iran;');
END IF;
IF X.ISRAEL IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Israel;');
END IF;
IF X.JAMAICA IS NOT NULL THEN
```

```
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Jamaica;');
END IF;
IF X.JAPAN IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Japan;');
END IF;
IF X.JORDAN IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Jordan;');
END IF;
IF X.KOREA IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Korea;');
END IF;
IF X.KYRGYZSTAN IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Kyrgyzstan;');
END IF;
IF X.LAO_DPR IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Lao_Dpr;');
END IF;
IF X.MALAWI IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Malawi;');
END IF;
IF X.MALAYSIA IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Malaysia;');
END IF;
IF X.MALI IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Mali;');
END IF;
IF X.MAURITANIA IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Mauritania;');
END IF;
IF X.MEXICO IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Mexico;');
END IF;
IF X.MONGOLIA IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Mongolia;');
END IF;
IF X.MOROCCO IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Morocoo;');
END IF;
IF X.MOZAMBIQUE IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Mozambique;');
END IF;
IF X.MYANMAR IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Myanmar;');
END IF;
IF X.NEPAL IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Nepal;');
END IF;
```

```
IF X.NETHERLANDS IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Netherlands;');
END IF;
IF X.NEW_ZEALAND IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';New_Zealand;');
END IF;
IF X.NICARAGUA IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Nicaragua;');
END IF;
IF X.NIGER IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Niger;');
END IF;
IF X.NIGERIA IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Nigeria;');
END IF;
IF X.NORWAY IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Norway;');
END IF;
IF X.OMAN IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Oman;');
END IF;
IF X.PAKISTAN IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Pakistan;');
END IF;
IF X.PALESTINE IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Palestine;');
END IF;
IF X.PANAMA IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Panama;');
END IF;
IF X.PAPUA_NEW_GUINEA IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Papua_New_Guinea;');
END IF;
IF X.PARAGUAY IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Paraguay;');
END IF;
IF X.PERU IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Peru;');
END IF;
IF X.PHILIPPINES IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Philippines;');
END IF;
IF X.ROMANIA IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Romania;');
END IF;
IF X.SAUDI_ARABIA IS NOT NULL THEN
   DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Saudi_Arabia;');
```

```
    END IF;
    IF X.SENEGAL IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Senegal;');
    END IF;
    IF X.SOUTH_AFRICA IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';South_Africa;');
    END IF;
    IF X.SRI_LANKA IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Sri_Lanka;');
    END IF;
    IF X.SURINAME IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Suriname;');
    END IF;
    IF X.SWEDEN IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Sweden;');
    END IF;
    IF X.SWITZERLAND IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Switzerland;');
    END IF;
    IF X.SYRIA IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Syria;');
    END IF;
    IF X.THAILAND IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Thailand;');
    END IF;
    IF X.TOGO IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Togo;');
    END IF;
    IF X.TRINIDAD_TOBAGO IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Trindad_and_Tobago;');
    END IF;
    IF X.URUGUAY IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Uruguay;');
    END IF;
    IF X.USA IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';United_States;');
    END IF;
    IF X.VENEZUELA IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Venezuela;');
    END IF;
    IF X.VIETNAM IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Vietnam;');
    END IF;
    IF X.ZIMBABWE IS NOT NULL THEN
       DBMS_OUTPUT.PUT_LINE(X.AGROTOXICO||';Zimbabwe;');
    END IF;
END LOOP;
```

```
END;
```

2. Consulta de geração de triplas da base *Relação entre Agrotóxicos e Sintomas*

```
DECLARE
agrotoxico VARCHAR2(100);
a int;
BEGIN
    a := 1;
    FOR X IN (SELECT DISTINCT BANNEDPESTICIDE, BANNEDPESTICIDEURI,
PESTICIDASBANIDOS, PESTICIDASBANIDOSURI FROM TCC2 ORDER BY
BANNEDPESTICIDE ASC)
        LOOP
            agrotoxico := X.BANNEDPESTICIDE;
            IF X.BANNEDPESTICIDEURI IS NOT NULL THEN
                a:=a+1;
                INSERT INTO TCC4 VALUES(a,
'<'||X.BANNEDPESTICIDEURI||'>
<http://www.w3.org/2000/01/rdf-schema#label>
<'||X.BANNEDPESTICIDE||'>@en.');
            ELSE
                a:=a+1;
                INSERT INTO TCC4 VALUES(a,
'<'||X.PESTICIDASBANIDOSURI||'> a
<http://purl.obolibrary.org/obo/CHEBI_25944>.');
            END IF;
            FOR Y IN (SELECT * FROM TCC2 WHERE BANNEDPESTICIDE =
agrotoxico ORDER BY COUNTRY ASC)
                LOOP
                    IF Y.BANNEDPESTICIDEURI IS NOT NULL THEN
                        a:=a+1;
                        INSERT INTO TCC4 VALUES(a,
'<'||Y.BANNEDPESTICIDEURI||'>
<http://lodbr.ufrj.br/agrotoxicos/propriedade/proibidoEm>
<'||Y.COUNTRYURI||'>.');
                    ELSE
                        a:=a+1;
                        INSERT INTO TCC4 VALUES(a,
'<'||Y.PESTICIDASBANIDOSURI||'>
<http://lodbr.ufrj.br/agrotoxicos/propriedade/proibidoEm>
<'||Y.COUNTRYURI||'>.');
                    END IF;
                END LOOP;
        END LOOP;
        COMMIT;
END;
```

3. Consulta de geração de triplas da base *Consolidated List of Banned Pesticides*

```
DECLARE
agrotoxico VARCHAR2(100);
a int;
BEGIN
    a := 1;
    FOR X IN (SELECT DISTINCT BANNEDPESTICIDE, BANNEDPESTICIDEURI,
PESTICIDASBANIDOS, PESTICIDASBANIDOSURI FROM TCC2 ORDER BY
BANNEDPESTICIDE ASC)
        LOOP
            agrotoxico := X.BANNEDPESTICIDE;
            IF X.BANNEDPESTICIDEURI IS NOT NULL THEN
                a:=a+1;
                INSERT INTO TCC4 VALUES(a,
'<'||X.BANNEDPESTICIDEURI||'>
<http://www.w3.org/2000/01/rdf-schema#label>
<'||X.BANNEDPESTICIDE||'>@en.');
            ELSE
                a:=a+1;
                INSERT INTO TCC4 VALUES(a,
'<'||X.PESTICIDASBANIDOSURI||'> a
<http://purl.obolibrary.org/obo/CHEBI_25944>.');
            END IF;
            FOR Y IN (SELECT * FROM TCC2 WHERE BANNEDPESTICIDE =
agrotoxico ORDER BY COUNTRY ASC)
                LOOP
                    IF Y.BANNEDPESTICIDEURI IS NOT NULL THEN
                        a:=a+1;
                        INSERT INTO TCC4 VALUES(a,
'<'||Y.BANNEDPESTICIDEURI||'>
<http://lodbr.ufrj.br/agrotoxicos/propriedade/proibidoEm>
<'||Y.COUNTRYURI||'>.');
                    ELSE
                        a:=a+1;
                        INSERT INTO TCC4 VALUES(a,
'<'||Y.PESTICIDASBANIDOSURI||'>
<http://lodbr.ufrj.br/agrotoxicos/propriedade/proibidoEm>
<'||Y.COUNTRYURI||'>.');
                    END IF;
                END LOOP;
        END LOOP;
        COMMIT;
END;
```

Consulta de geração de triplas da base *Relação entre Agrotóxicos e Sintomas Crônicos e Agudos*

```
DECLARE
grupo_quimico VARCHAR2(100);
BEGIN
    FOR X IN (SELECT DISTINCT GRUPOQUIMICO, GRUPOQUIMICOURI,
CHEMICALGROUP, CHEMICALGROUPURI, CLASSE, CLASSEURI, CLASS,
CLASSURI, GRUPOQUIMICOCLASSEURI, CHEMICALENTITYURI FROM TCC ORDER
BY CHEMICALENTITYURI ASC)
        LOOP
            grupo_quimico := X.GRUPOQUIMICO;
            DBMS_OUTPUT.PUT_LINE('<'||X.CHEMICALENTITYURI||'> is_a
<http://purl.obolibrary.org/obo/CHEBI_25944>.');
            DBMS_OUTPUT.PUT_LINE('<'||X.CHEMICALENTITYURI||'>
<http://aims.fao.org/aos/agrontology#hasClass>
<'||X.CLASSURI||'>.');
            DBMS_OUTPUT.PUT_LINE('<'||X.CLASSURI||'>
<http://www.w3.org/2000/01/rdf-schema#label> "'||X.CLASS||'"@en.');
            DBMS_OUTPUT.PUT_LINE('<'||X.CHEMICALENTITYURI||'>
<http://aims.fao.org/aos/agrontology#hasComponent>
<'||X.CHEMICALGROUPURI||'>.');
            DBMS_OUTPUT.PUT_LINE('<'||X.CHEMICALGROUPURI||'>
<http://www.w3.org/2000/01/rdf-schema#label>
"'||X.CHEMICALGROUP||'"@en.');
            FOR Y IN (SELECT * FROM TCC WHERE GRUPOQUIMICO =
grupo_quimico ORDER BY CHEMICALENTITYURI ASC)
                LOOP

DBMS_OUTPUT.PUT_LINE('<'||Y.CHEMICALENTITYURI||'>
<http://aims.fao.org/aos/agrontology#causes>
<'||Y.SYMPTOMURI||'>.');
                    DBMS_OUTPUT.PUT_LINE('<'||Y.GRUPOQUIMICOURI||'>
<http://aims.fao.org/aos/agrontology#causes>
<'||Y.SYMPTOMURI||'>.');
                    DBMS_OUTPUT.PUT_LINE('<'||Y.GRUPOQUIMICOURI||'>
<http://aims.fao.org/aos/agrontology#causes>
<'||Y.SINTOMASURI||'>.');
                    DBMS_OUTPUT.PUT_LINE('<'||Y.SINTOMASURI||'> a
<http://dbpedia.org/page/Symptom>.');
                    DBMS_OUTPUT.PUT_LINE('<'||Y.SYMPTOMURI||'>
<owl:sameAs> <'||Y.SINTOMASURI||'>.');
                END LOOP;
                DBMS_OUTPUT.PUT_LINE(' ');
        END LOOP;
END;
```