UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE CIÊNCIA DA INFORMAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO & ORGANIZAÇÃO DO CONHECIMENTO

RAFAEL ROCHA

Integração Semântica de Dados Tabulares em CSV: proposta de arcabouço comparativo de ferramentas

BELO HORIZONTE

2021

RAFAEL ROCHA

Integração Semântica de Dados Tabulares em CSV: proposta de arcabouço comparativo de ferramentas

Dissertação apresentada ao Programa de Pós-Graduação em Gestão & Organização do Conhecimento, Escola de Ciência da Informação da Universidade Federal de Minas Gerais para obtenção do grau de Mestre, área de concentração Ciência da Informação.

Linha de Pesquisa: Gestão & Tecnologia da Informação e Comunicação

Orientador: Prof. Dr. Marcello Peixoto Bax

BELO HORIZONTE

2021

R672c Rocha, Rafael

Integração semântica de dados tabulares em CSV [recurso eletrônico]: proposta de arcabouço comparativo de ferramentas. / Rafael Rocha. - 2021.

1 recurso eletrônico (63 f. : il., color): pdf.

Orientador: Marcello Peixoto Bax.
Dissertação (Mestrado) – Universidade Federal de Minas Gerais, Escola de Ciência da Informação.
Referências: f. 59-63.
Exigências do sistema: Adobe Acrobat Reader.

1. Ciência da Informação – Teses. 2. Web semântica – Teses. I. Título. II. Bax, Marcello Peixoto. III. Universidade Federal de Minas Gerais, Escola de Ciência da Informação.

CDU:004.738.5

Ficha catalográfica: Maianna Giselle de Paula CRB:2642
Biblioteca Profª Etelvina Lima, Escola de Ciência da Informação da UFMG.

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO

FOLHA DE APROVAÇÃO

Integração Semântica de Dados Tabulares em CSV: proposta de arcabouço comparativo de ferramentas.

RAFAEL ROCHA

Dissertação submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, como requisito para obtenção do grau de Mestre em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, área de concentração CIÊNCIA DA INFORMAÇÃO, linha de pesquisa Gestão e Tecnologia da Informação e Comunicação.

Aprovada em 22 de fevereiro de 2021, pela banca constituída pelos membros:

Prof(a). Marcello Peixoto Bax (Orientador)
ECI/UFMG [por videoconferência]

Prof(a). Frederico Cesar Mafra Pereira
ECI/UFMG [por videoconferência]

Prof(a). Renata Maria Abrantes Baracho Porto
Escola de Arquitetura/UFMG [por videoconferência]

Prof(a). Elisângela Cristina Aganette
ECI/UFMG [por videoconferência]

Belo Horizonte, 22 de fevereiro de 2021.

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO

ATA DA DEFESA DA DISSERTAÇÃO DO ALUNO

RAFAEL ROCHA

Realizou-se, no dia 22 de fevereiro de 2021, às 09:30 horas, Videoconferência, da Universidade Federal de Minas Gerais, a defesa de dissertação, intitulada *Integração Semântica de Dados Tabulares em CSV: proposta de arcabouço comparativo de ferramentas*, apresentada por RAFAEL ROCHA, por videoconferência, número de registro 2019663532, graduado no curso de TECNOLOGIA EM PROCESSAMENTO DE DADOS, como requisito parcial para a obtenção do grau de Mestre em GESTÃO E ORGANIZAÇÃO DO CONHECIMENTO, à seguinte Comissão Examinadora: Prof(a). Marcello Peixoto Bax - ECI/UFMG [por videoconferência] (Orientador), Prof(a). Frederico Cesar Mafra Pereira - ECI/UFMG [por videoconferência], Prof(a). Renata Maria Abrantes Baracho Porto - Escola de Arquitetura/UFMG [por videoconferência], Prof(a). Elisângela Cristina Aganette - ECI/UFMG [por videoconferência].

A Comissão considerou a dissertação:

(x) Aprovada

( ) Reprovada

Finalizados os trabalhos, lavrei a presente ata que, lida e aprovada, vai assinada por mim e pelos membros da Comissão.
Belo Horizonte, 22 de fevereiro de 2021.

Prof(a). Marcello Peixoto Bax

Prof(a). Frederico Cesar Mafra Pereira

Prof(a). Renata Maria Abrantes Baracho Porto

Prof(a). Elisângela Cristina Aganette

Resumo

A web semântica representa o conhecimento de maneira legível para seres humanos e computadores. “Dados conectados” (*linked data*) semanticamente associam conceitos de diversas fontes, e o reúso de dados e vocabulários enriquece diversos sistemas de informação na web, sobretudo aqueles voltados para organizar dados de pesquisas científicas. Tais sistemas manipulam dados tabulares em grades bidimensionais (arquivos CSV - *Comma Separated Values*) com seus metadados localizados no cabeçalho do arquivo (primeira linha do arquivo). Em geral, os metadados do CSV são insuficientes para possibilitar integração e interoperabilidade semântica, i.e., a capacidade dos sistemas de se comunicar de forma transparente (ou o mais próximo disso) com outros sistemas (semelhantes ou não). Para contribuir nas pesquisas, os arquivos CSV devem ser integrados semanticamente e armazenados em repositórios de dados. Os dados tabulares precisam ter seus significados explicitados de modo que os conceitos tratados não se percam ou tenham seus significados distorcidos. O processo de integração semântica de dados é realizado por ferramentas que automatizam o processo, com o intuito de sistematizar e agilizar o trabalho, minimizando os erros humanos. Essas ferramentas possuem características e implementações distintas e os recursos (ou funcionalidades) disponíveis em cada uma delas impactam na sua capacidade de integrar os dados gerando dados conectados para a web semântica. Um determinado projeto de integração de dados pode fracassar caso a ferramenta escolhida para gerar dados conectados não possua os recursos necessários ao projeto. Diversos arcabouços comparativos foram propostos para avaliar ferramentas na geração de dados conectados, mas nenhum deles utiliza uma escala de valores que simplifique a avaliação e a sumarização dos resultados das análises. Propõe-se nesta pesquisa um arcabouço comparativo para ferramentas para integração semântica de dados tabulares em CSV. Os recursos do arcabouço são fundamentados na literatura científica com pontos dispostos em uma reta numérica positiva. No percurso metodológico utiliza-se um arquivo em CSV no processo de integração semântica, em seguida as ferramentas são avaliadas à luz do arcabouço comparativo. Assim, dispondo as ferramentas em uma reta numérica positiva é possível saber quais delas possuem os recursos mais adequados para um dado projeto de integração ou ainda os recursos mais bem avaliados. Os resultados deste trabalho são úteis para todos aqueles que necessitam avaliar ferramentas em seus projetos de integração semântica de dados, principalmente nas pesquisas científicas, uma vez que os dados conectados conceitualmente contribuem sobremaneira para as mesmas.

Palavras-chaves: Integração Semântica. Dados Tabulares. Ferramentas para Integração Semântica. Arcabouço Comparativo. Web Semântica. Dados Conectados.

Abstract

The semantic web represents knowledge in a human-readable and machine-readable form. Linked data semantically associate concepts from different sources, and the reuse of data and vocabularies enriches information systems on the web, especially those aimed at organizing scientific research data. Such systems manipulate tabular data in two-dimensional degrees (CSV files - Comma Separated Values) with their associated metadata in the file header (first line of the file). In general, CSV metadata is insufficient to enable semantic integration and interoperability, that is, the ability of systems to communicate transparently (or as closely as possible) with other similar systems (or not). To contribute to research, CSV files must be integrated semantically and stored in data repositories. Tabular data must have its meanings made explicit so that the concepts treated are not lost or have their meanings distorted. The process of semantic data integration is performed by tools that automate the process, in order to systematize and streamline the work, minimizing human errors. These tools have different characteristics and implementations and the features (or functionalities) available in each of them impact on their ability to integrate data generating linked data for semantic web. A given data integration project can fail if the tool chosen to generate linked data does not have the features available to the project. Several comparison frameworks have been proposed to evaluate tools in the generation of linked data, but none of them uses a scale of values that simplifies the evaluation and summary of the results of the analyzes. This research proposes a comparison framework for semantic integration tools of tabular data in CSV. The features of the framework are based on the scientific literature with points arranged on a positive number line. In the methodological path, a CSV file is used in the semantic integration process, then the tools are evaluated in the light of the comparison framework. Thus, having the tools on a positive number line, it is possible to know which of them have the most adequate features for a given integration project or even the best evaluated features. The results of this work are useful for all those who need to evaluate tools in their semantic data integration projects, especially in scientific research, since the data connected conceptually contribute greatly to them.

Keywords: Semantic Integration. Tabular Data. Semantic Integration Tools. Comparison Framework. Semantic Web. Linked Data.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ASCII	*American Standard Code for Information Interchange*
CAFe	Comunidade Acadêmica Federada
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CSV	*Comma-Separated Values*
DL	*Description Language*
DSV	*Delimiter-Separated Values*
ETL	*Extract, Transformation, Load*
GCOP	Grafo de Conhecimento dos Óbitos de Policiais
HADatAc	*Human-Aware Data Acquisition*
LD	*Linked Data*
LD-R	*Linked Data Reactor*
LDIF	*Linked Data Integration Framework*
LOD	*Linked Open Data*
LP-ETL	*LinkedPipes ETL*
OBDA	*Ontology-Based Data Access*
OWL	*Web Ontology Language*
RD	Recursos para metaDados
RDF	*Resource Description Framework*
RDFS	*Resource Description Framework Schema*
RM	Recursos para Mapeamento
RP	Recursos para Processamento
RR	Recursos para peRsistência
RS	Recursos para Serialização

Silk	*Silk Link Discovery Framework*
SPARQL	*SPARQL Protocol and RDF Query Language*
TSV	*Tab-Separated Values*
Turtle	*Terse RDF Triple Language*
URI	*Uniform Resource Identifier*
URL	*Uniform Resource Locator*
URN	*Uniform Resource Name*
W3C	*World Wide Web Consortium*
WOD	*Web Of Data*
XML	*eXtensible Markup Language*
XSD	*XML Schema Definition*

Sumário

1 INTRODUÇÃO

A web semântica representa conhecimentos em estrutura legível por humanos e computadores na web convencional. Além disso, conceitos publicados na web podem ser reaproveitados para enriquecer expressões semânticas (BERNERS-LEE; HENDLER; LASSILA, 2001). Nesse sentido, a integração semântica de dados se beneficia com o reúso dos conceitos e informações, ademais mantém a interoperabilidade de significado das diversas fontes de dados com visão unificada (IZZA, 2009). A visão se materializa por meio de um grafo de conhecimento que também é utilizado para modelagem conceitual para explicitar os significados. Os grafos de conhecimentos são redes de vértices e arestas que representam entidades do mundo real e suas relações. Com esta estrutura é possível manter qualquer conhecimento (EHRLINGER; WÖSS, 2016). Os dados tabulares não possuem semântica compartilhada, esse formato é uma grade bidimensional organizada em linhas e colunas, dessa maneira, tabelas de banco de dados relacionais; planilhas eletrônicas; arquivos *Comma-Separated Values*[1] (CSV) são alguns exemplos desse formato (ADELFIO; SAMET, 2013).

Os padrões da web semântica aplicada na integração semântica de dados permite a construção de um repositório com informação baseada em conceitos reutilizáveis. As pesquisas científicas se favorecem dos dados tabulares disponibilizados, já que a investigação contará com mais informação. A ciência de dados[2] (*data science*) disponibiliza milhares de conjuntos de dados publicamente, já os dados abertos governamentais estimulam a governança por meio da transparência, assim alguns dos conjuntos estão disponibilizados em CSV[3] (ATTARD *et al.*, 2015). O CSV é um formato padronizado que armazena os dados em uma estrutura tabular: a primeira linha é o cabeçalho do arquivo, que representa os metadados; as linhas em sequência contêm os dados, assim, os dados e cabeçalhos são separados por vírgulas, formando as colunas; todas as linhas (registros) de uma coluna (atributos) pertencem ao mesmo conjunto de valores (SHAFRANOVICH, 2019).

O CSV não favorece a interoperabilidade semântica dos dados, uma vez que seus metadados descrevem de forma limitada seus atributos. Assim, para extrair mais valor[4] dos dados em CSV é necessário integrá-los com outros dados em um repositório (STEIN; MORRISON, 2014). Um repositório armazena os dados em seu estado "natural", ou seja, sem impor um esquema coercitivo de informação. Extrair inteligência de negócio (*Business Intelligence*) ou informação analítica (*analytics*) é atrativo quando ao longo do tempo o repositório concentra

[1] Os formatos como *Delimiter-Separated Values* (DSV) e *Tab-Separated Values* (TSV) serão considerados neste trabalho como variações do CSV.

[2] Disponível em: https://www.kaggle.com. Acesso em 11/07/2020.

[3] Alguns conjuntos de dados estão disponíveis em: https://data.europa.eu/euodp/en/data/dataset; https://dados.gov.br/dataset. Acesso em 15/12/2020.

[4] No preenchimento do campo de um metadado, o valor descreve e/ou detalha o significado do dado.

volume e diversidade de dados (MILOSLAVSKAYA; TOLSTOY, 2016). Neste contexto, ao acrescentar os registros de um CSV em repositório, primeiramente é necessário enriquecê-los com semântica para que os significados dos dados se mantenham íntegros.

Uma maneira de se conseguir o enriquecimento dos dados tabulares em CSV é utilizar as tecnologias semânticas para conectar os dados, gerando "dados conectados" ou *Linked Data* (LD) (ISOTANI; BITTENCOURT, 2015). O LD pode representar o conhecimento com a ajuda de um modelo de declaração semântica em forma de tripla que é constituído por sujeito, predicado e objeto (BIZER; HEATH; BERNERS-LEE, 2011). Desse modo, cada coluna do arquivo CSV é mapeada para um recurso da tripla, enriquecendo semanticamente os dados tabulares. O *Uniform Resource Identifier* (URI) confere identificação única a cada recurso de uma tripla para que não ocorra conflito no conhecimento representado. Além disso, uma declaração semântica pode receber valores literais, ou seja, números ou textos. Logo, o LD não perde o conhecimento representado, pois, em seu conjunto de definições possuem padrões compartilhados de semântica única, que permanecerão inalterados desde a sua origem até ao fim de uma comunicação. Além disso, pode-se manter a compreensão compartilhada do significado dos dados, tanto por pessoas, quanto por computadores, uma vez que ele é um modelo semântico.

Borgida *et al.* (1989) elencam três pontos em que o modelo semântico é superior a outros modelos: (i) Quando objetos complexos são o jeito natural de descrevê-los; (ii) quando a informação do domínio está incompleta ou é incrementalmente disponível; (iii) quando o banco de dados deve ter uma função mais ativa e conseguir deduzir relacionamentos ao invés de ser um mero repositório de dados. Neste último ponto, Hammer e McLeod (1978) destacam que o banco de dados é um retrato do estado de um sistema em determinado momento, assim esses repositórios devem representar algo mais que uma mera coleção de informações.

1.1 Problema e Justificativa

As ferramentas para integração semântica automatizam o contínuo processamento dos dados (Biffl *et al.*, 2014). A automação é importante, pois novos dados são gerados a todo momento, e, além de proporcionar maior confiabilidade, reduz erros e retrabalho na execução frequente das tarefas. Cada ferramenta possui fatores que influenciam na integração semântica de dados, tais como recursos implementados, parametrização e configuração. Caso uma decisão seja tomada sem se conhecer os detalhes da ferramenta semântica, pode-se colocar em risco um projeto. Logo, é necessário conhecer todos os recursos que a ferramenta oferece antes de ser adotada. A ciência da informação endereça diversas soluções para os problemas na integração de dados (TARAPANOFF; JÚNIOR; CORMIER, 2000; GALVÃO; RICARTE, 2011). Os trabalhos na área realizam tratamento da criação dos dados até a recuperação da informação. A curadoria da informação se beneficia dos instrumentos criados na ciência da informação. Uma proposta de método de avaliação das ferramentas é verificar a capacidade de mapeamento do LD por

meio de modelos de bancos de dados relacionais. A aferição se dá por meio da análise da diferença do resultado esperado, em contraste com o resultado obtido (SEQUEDA; PRIYATNA; VILLAZóN-TERRAZAS, 2012; BAGUI; BOURESSA, 2014). A deficiência desse método é a falta de critérios objetivos para avaliar as abordagens utilizadas em cada ferramenta.

Uma alternativa para isso é adotar um arcabouço comparativo, composto por um conjunto de critérios objetivos de avaliação. Os arcabouços comparativos são estruturas conceituais construídas com itens avaliativos para examinarem problemas específicos. Os arcabouços comparativos de ferramentas semânticas avaliam os recursos presentes em cada ferramenta. (HERT; REIF; GALL, 2011; MICHEL; MONTAGNAT; ZUCKER, 2014; JUNIOR *et al.*, 2017; RASHID *et al.*, 2020). Além disso, avaliar cada recurso com valor categórico não permite uma sumarização eficaz dos resultados, assim necessitando de conversões de valores (HERT; REIF; GALL, 2011; MICHEL; MONTAGNAT; ZUCKER, 2014; JUNIOR *et al.*, 2017; DIMOU *et al.*, 2018). Ao passo que pontuar em uma reta numérica positiva[5] permite uma constatação eficaz de qual ferramenta possui mais recursos ou recursos mais avançados. Na literatura científica não foi encontrado um arcabouço comparativo para avaliar ferramentas para integração semântica de dados.

Cada projeto de integração de dados é único devido aos participantes ou por causa dos dados utilizados. O sucesso ou fracasso de um projeto pode ser atribuído a adoção de uma determinada ferramenta. Este trabalho é relevante, pois propõe um método pragmático de avaliação de ferramentas. A reta numérica positiva permitirá transformar o arcabouço comparativo em gráficos ou simplesmente sumarizar os resultados. Além disso, as ferramentas estão mais próximas do funcionamento do dia a dia do projeto com a utilização de dados do mundo real. Os resultados são úteis para todos os projetos que necessitam comparar as ferramentas para integração semântica. Em especial a ciência que produzirá pesquisas mais ricas para a sociedade com os dados conectados. Exposta a problematização e a justificativa, esta pesquisa busca responder à seguinte questão: Quais os recursos são essenciais para a avaliação de ferramentas para integração semântica de dados tabulares em CSV em um arcabouço comparativo com resultados que possam ser medidos numericamente?

1.2 Objetivo

O objetivo geral deste trabalho é propor um arcabouço comparativo de ferramentas para a integração semântica de dados tabulares em CSV utilizando uma reta numérica positiva para avaliação dos recursos.

Os objetivos específicos são: (i) construir um grafo de conhecimento a partir de dados tabulares como linha de base para integração semântica; (ii) determinar, com base na literatura

[5] Uma reta numérica positiva sempre possui seu início no número zero e cresce para o infinito positivo. Disponível: https://en.wikipedia.org/wiki/Number_line. Acesso em 05/01/2021.

científica, quais ferramentas possuem características para integração semântica de dados tabulares; (iii) mapear insumos na literatura para construir o arcabouço comparativo para avaliação de ferramentas para integração semântica de dados; (iv) comparar as ferramentas que realizam integração semântica de dados.

1.3 Estrutura

O Capítulo 1 introduz e contextualiza a temática para apresentar as motivações para o estudo, em seguida apresenta o problema que a pesquisa se propõe e a justificativa para sua realização. Por fim, descrevem-se os objetivos geral e específicos, que sintetizam o propósito desta pesquisa. O Capítulo 2 examina as linguagens e tecnologias na web semântica para obter o LD. Apresenta como se constitui uma tripla e alguns vocabulários utilizados. Além disso, é discutido o processo de elevação semântica e a integração dos dados de um arquivo CSV. No Capítulo 3, são discutidos os métodos para avaliar ferramentas semânticas. Realiza-se a revisão dos recursos necessários nos arcabouços comparativos. O Capítulo 4 apresenta a caracterização da pesquisa e as etapas do percurso metodológico para, logo após, discutir no Capítulo 5 os resultados obtidos nos procedimentos metodológicos. Por fim, o Capítulo 6 apresenta as considerações finais.

2 WEB SEMÂNTICA

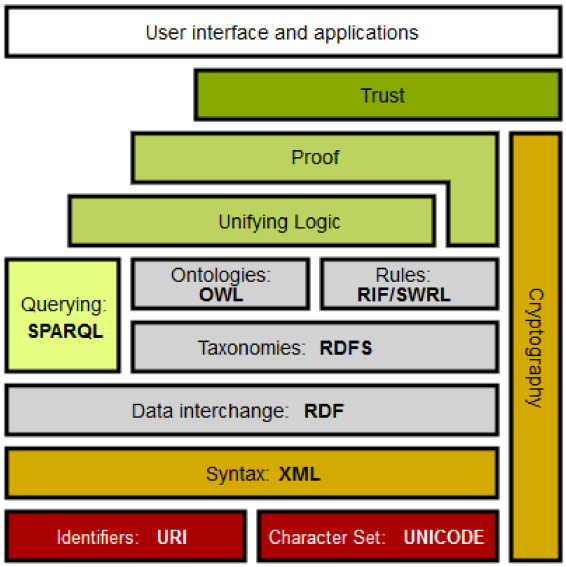
Este capítulo apresenta os conceitos da web semântica e discute como estruturar um LD. A web semântica estende a web convencional, ou seja, reaproveita os hipertextos e atribui significado aos dados e suas ligações (*link*) (BERNERS-LEE, 1998; BERNERS-LEE; HENDLER; LASSILA, 2001). Os hipertextos são marcados com etiquetas (*tags*) da web semântica, assim tanto pessoas quanto computadores interpretam de maneira única uma expressão. Outra alternativa é o *eXtensible Markup Language* (XML) que é a linguagem de marcação utilizada para descrever textualmente as expressões semânticas.

As expressões semânticas, também conhecidas como triplas, possuem estrutura com: sujeito, predicado e objeto. O LD permite que cada parte de uma expressão semântica reaproveite conceitos de outro local na web (BIZER; HEATH; BERNERS-LEE, 2011). O *Resource Description Framework* (RDF) é o arcabouço descritivo para codificar as triplas. Já o *Resource Description Framework Schema* (RDFS) e o *Web Ontology Language* (OWL) conferem mais vocabulário para as triplas.

Os vocabulários são conjuntos de conceitos com identificadores únicos para que não ocorra conflito na representação do conhecimento. Além do vocabulário, o OWL permite representar conceitos ontológicos (Seção 2.4.1). Este trabalho não lida com as tecnologias para tornar o LD seguro, tal como a criptografia; também não contempla a linguagem de consulta de triplas *SPARQL Protocol and RDF Query Language* (SPARQL).

A breve introdução desse Capítulo é condensada conforme a Figura 1, a web semântica possui uma pilha de linguagens, tecnologias e protocolos para representar o conhecimento.

Figura 1 – Pilha hierárquica da web semântica



Fonte: (Wikipedia contributors, 2019)

Este Capítulo traz a seguinte ordem: A Seção 2.1 discute a composição de uma tripla na web semântica. A Seção 2.2 discute os vocabulários do RDFS e OWL. O LD é discutido na Seção 2.3. A Seção 2.4 discute o LD oriundo dos dados tabulares em um arquivo CSV. Finalmente na Seção 2.5 apresenta a integração semântica de dados e os meios necessários para obtenção de uma visão unificada de conhecimento.

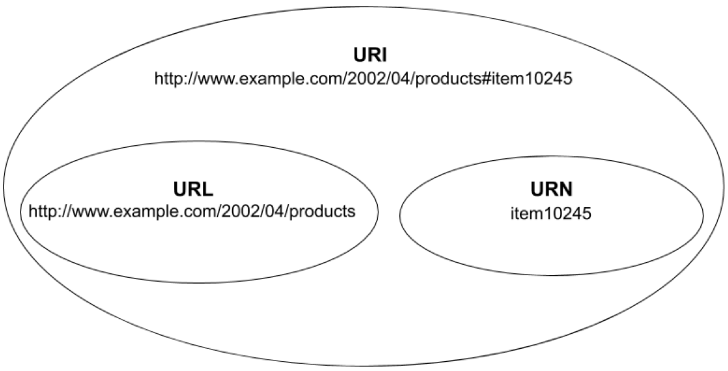
2.1 Triplas

Os conceitos na web semântica precisam ser identificados unicamente para que não ocorra conflito no conhecimento representado. Uma vez que no mundo real existem homônimos, faz-se necessário a utilização do URI (BERNERS-LEE, 2019). Ademais, o *Uniform Resource Locator* (URL) e o *Uniform Resource Name* (URN) que compõem o URI, geram a desambiguação necessária para identificação dos recursos na web semântica, em virtude da impossibilidade de se registrar diversos URI's idênticos na internet. Portanto, o URI *http://www.example.com/ 2020/ products#item10245* é formado pelo URL *http://www.example.com/ 2020/ products* e o URN *item10245*. A Figura 2 ilustra o URI e seus componentes.

O XML é uma linguagem de marcação extensível hierárquica que permite a criação customizável de elementos, destaque para as *tags* e atributos (BAX, 2001; BRAY *et al.*, 2008). Desse modo, os conflitos de elementos são resolvidos com a utilização de *namespace*[1]. Além

[1] Declaração do namespace xhtml e sua respectiva URI segue a seguinte forma: xmlns:xhtml="http://www.w3.org/1999/xhtml"

Figura 2 – Composição do URI



Fonte: Elaborado pelo autor

disso, *tags* e atributos podem receber valores e tipos (complexo ou primitivo). O *XML Schema Definition*[2] (XSD) é o *namespace* que define os tipos suportados. Assim as representações *"0"^^xsd:integer* e *"0"^^xsd:string*, são respectivamente um *integer* (inteiro) e uma *string* (cadeia de caracteres) (EMMONS *et al.*, 2011).

A interoperabilidade de significado é mantida pela representação dos conceitos em forma de uma tripla RDF (GARDNER, 2005). Uma tripla RDF é constituída por três elementos: sujeito, o conceito no qual a tripla está declarando; predicado, o tipo de relação; objeto, a asserção de valor. Salienta-se que a declaração de uma tripla utiliza XML, por recomendação do *World Wide Web Consortium* (W3C), sendo compreendida tanto por pessoas, quanto por computadores (LASSILA; SWICK, 1999). Desse modo, as triplas e seus recursos são abstrações virtuais de objetos do mundo real. A Figura 3 demonstra as triplas em forma gráfica.

Figura 3 – Triplas em RDF na forma gráfica

Fonte: (POWERS, 2003, p. 33)

As triplas em RDF possuem sua representação em XML, mas outras serializações[3] são recomendadas pelo W3C, tal como o *Terse RDF Triple Language* (Turtle) que fornece uma forma textual compacta de expressar uma tripla e *N-Triples* que fornece uma sintaxe de triplas

[2] As definições do *schema* estão disponíveis em: http://www.w3.org/2001/XMLSchema. Acesso em 22/12/2020.
[3] Uma serialização é a representação em texto de algo na memória do computador.

por linha (W3C Recommendation, 2020b; W3C Recommendation, 2020a). A Figura 4 possui conhecimentos expressos em triplas em XML, na sequência a Figura 5 expressa os mesmos conhecimentos em Turtle (a) e por fim em *N-Triples* (b).

Figura 4 – Triplas serializadas em XML

```xml
<?xml version="1.0"?>
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:schema="http://schema.org/"
   xmlns:foaf="http://xmlns.com/foaf/0.1/"
>
   <rdf:Description rdf:about="https://www.w3.org/People/Berners- Lee/">
      <rdf:type rdf:resource="foaf:Person"/>
      <foaf:name>Tim Berners-Lee</foaf:name>
      <schema:birthDate>1955-06-08</schema:birthDate>
      <schema:birthPlace rdf:resource="http://dbpedia.org/resource/London"/>
   </rdf:Description>
</rdf:RDF>
```

Fonte: Elaborado pelo autor

O sujeito na serialização em Turtle não precisa ser declarado em cada tripla, para isso, utiliza-se “;” como indicador que o sujeito é compartilhado. Finalmente, a indicação de término de uma tripla é simbolizado pelo carácter “.”. Ao passo que o *N-Triples* declara em cada linha uma tripla completa, o carácter “.” indica o final da tripla.

Figura 5 – Triplas serializadas em Turtle e N-Triples

(a) Serialização em Turtle

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix schema: <http://schema.org/> .

<https://www.w3.org/People/Berners-Lee/>
   a foaf:Person ;
   foaf:name "Tim Berners-Lee" ;
   schema:birthDate "1955-06-08" ;
   schema:birthPlace <http://dbpedia.org/resource/London> .
```

(b) Serialização em N-Triples

```
<https://www.w3.org/People/Berners-Lee/> <http://www.w3.org/1999/02/22-rdf-syntax-ns
    ↪ #type> <foaf:Person> .
<https://www.w3.org/People/Berners-Lee/> <http://xmlns.com/foaf/0.1/name> "Tim
    ↪ Berners-Lee" .
<https://www.w3.org/People/Berners-Lee/> <http://schema.org/birthDate> "1955-06-08"
    ↪ .
<https://www.w3.org/People/Berners-Lee/> <http://schema.org/birthPlace> <http://
    ↪ dbpedia.org/resource/London> .
```

Fonte: Elaborado pelo autor

Além das triplas com recurso declarado, o RDF possui a concepção de *Blank nodes* que são nodos que não possuem URI (CHEN *et al.*, 2012). Estes recursos não identificados são

utilizados para representar diversas estruturas de dados. A Figura 6 retrata um *blank node* com uma associação n-ária.

Figura 6 – *Blank node* em RDF

```
@prefix ns0: <http://xmlns.com/foaf/0.1#> .
@prefix ns1: <http://example.org/data#> .
<https://www.w3.org/People/Berners-Lee/>;
ns0:name "Tim Berners-Lee" ;
ns1:birth [
   ns1:city "London" ;
   ns1:date "1955-06-08"
] .
```

Elaborado pelo autor

O *Named graph* permite nomear triplas com um determinado agrupamento. Essa concepção na web semântica atribui contexto às triplas (CARROLL *et al.*, 2005). A Figura 7 representa o *named graph G1* e suas respectivas triplas agrupadas por chaves.

Figura 7 – Exemplo de *named graph*

```
   :G1{
    _:Monica ex:name "Monica Murphy" .
    _:Monica ex:email <mailto:monica@murphy.org> .
    :G1 pr:disallowedUsage pr:Marketing
   }
```

Fonte: (CARROLL *et al.*, 2005)

As triplas na web semântica podem representar a proveniência dos dados. A camada de confiança (*Trust*) prevê mecanismos para que os conhecimentos representados sejam realmente da procedência esperada ou que no mínimo forneça recursos suficientes para aferir a qualidade dos mesmos (HARTIG, 2009). Ademais, é necessário metadados que representam: qualidades de precisão, atualidade, confiança e credibilidade. Em Ding *et al.* (2010), os autores demonstram que a utilização de vocabulários ontológicos traz mais robustez para checar a proveniência dos dados. Dessa forma, várias ontologias podem ser utilizadas para proveniências, além da possibilidade de estendê-las.

2.2 Vocabulários

O RDFS introduz mais vocabulário[4] para as triplas. Os vocabulários que se destacam são: *Class*, define quem é a classe pai do recurso, ou relação de subsunção; *Literal*, são todos os tipos de dados que não são recursos; *domain*, define de qual recurso origina-se o predicado; *range*, qual recurso aplica-se o predicado (MCBRIDE, 2004). A Figura 8 demonstra a sintaxe das triplas enriquecidas com o vocabulário RDFS.

Figura 8 – Triplas em RDF enriquecidas com RDFS serializadas em XML

```
<?xml version="1.0"?>
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
      xmlns:pstcn="http://burningbird.net/postcon/elements/1.0/">

      <rdfs:Class rdf:about="http://burningbird.net/postcon/elements/1.0/
          ↪ Article">
            <rdfs:subClassOf rdf:resource="http://www.w3.org/2000/01/rdf-schema
                ↪ #Resource"/>
      </rdfs:Class>
</rdf:RDF>
```

Fonte: Adaptado de (POWERS, 2003, p. 90)

O OWL insere mais expressividade às triplas com seu vocabulário, possibilitando expressar conceitos ontológicos (W3C Recommendation, 2019a). O vocabulário do OWL possui recursos[5] suficientes para descrever: igualdade (*sameAs*), disjunção (*AllDifferent*), característica de propriedade (*TransitiveProperty*), restrição de propriedade (*someValuesFrom*), cardinalidade (*cardinality*), interseção de classe (*intersectionOf*), versionamento (*versionInfo*) e anotações (*AnnotationProperty*). Além disso, é possível expressar conceitos e instanciá-los com o OWL indivíduos (*NamedIndividual*). Por fim, uma linguagem ontológica deve possuir cinco características (ANTONIOU; HARMELEN, 2004):

1) Sintaxe bem-definida: as regras estruturais das declarações não se contradizem e não são ambíguas.

2) Semântica bem-definida: os significados das declarações não se contradizem e não são ambíguas.

3) Suporte a um raciocinador eficiente: raciocinador que permita checar a consistência do conhecimento representado.

[4] Um vocabulário atribui completude no significado ao anotar os dados com seus respectivos conceitos. A lista completa dos vocabulários RDFS está disponível em: https://www.w3.org/TR/rdf-schema/. Acesso em 22/12/2020.

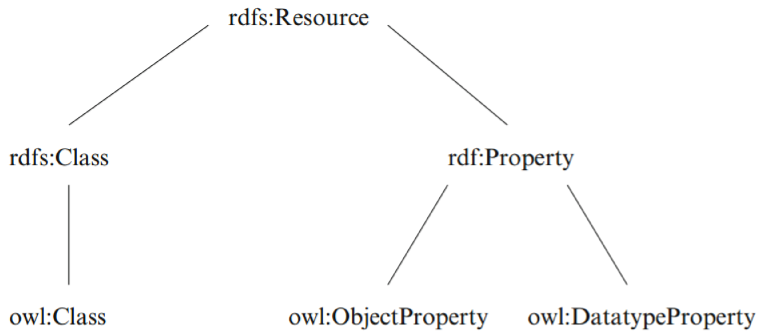
[5] A lista completa dos recursos do vocabulário do OWL está disponível em: https://www.w3.org/TR/owl-features/

4) Suficiente poder expressivo: uma linguagem ontológica precisa ser capaz de expressar escopo, disjunção de classes, combinação booleana, cardinalidade de restrições e característica das propriedades de uma classe.

5) Conveniência nas expressões: permitir expressões simples a complexas conforme a necessidade do usuário.

A interoperabilidade do OWL é garantida com reaproveitamento dos vocabulários do RDFS, dessa maneira é possível escalar em significado as triplas que foram criadas originalmente com RDF. A Figura 9 demonstra parte da herança do OWL.

Figura 9 – Hierarquia de classes



Fonte: (ANTONIOU; HARMELEN, 2004)

OWL possui três sublinguagens em sua taxonomia, definidas como: OWL *Full*, OWL *Description Language* (DL) e OWL *Lite* (W3C Recommendation, 2019b).

- OWL *Full* é sua forma mais expressiva por incorporar todas as outras sublinguagens de OWL. Em OWL *Full*, toda *owl:Class* é equivalente à classe *rdfs:Class*, sendo assim todo documento RDF válido é considerado um documento OWL *Full* válido. Em OWL *Full*, os valores de dados e indivíduos estão no mesmo nível, ou seja, são subclasses de *rdf:Property* sendo assim, *owl:DataTypeProperty* e *owl:ObjectProperty* são disjunção da classe *rdf:Property*. Por esse poder expressivo maior é difícil a implementação de raciocinadores em OWL *Full*.

- OWL DL é mais expressiva que OWL *Lite* e menos expressiva que o OWL *Full*, nesta forma há a implementação da lógica de descrição, sendo a lógica de descrição mais expressiva que a lógica proposicional e mais resolvível que a lógica de primeira ordem. Os raciocinadores podem levar a novos conhecimentos por meio das inferências do modelo ontológico ou até mesmo encontrar alguma inconsistência na ontologia. OWL DL para ser eficaz dependente das propriedades, axiomas e cardinalidade para a resolução da ontologia.

- OWL *Lite* é uma sublinguagem com menor expressividade traz consigo uma curva de aprendizagem menor e a facilidade de utilização em tecnologias semânticas. Permite classificação e taxonomia com restrições simples. Pode ser utilizado para a interoperabilidade de banco de dados, ferramentas e sistemas de informação, por exemplo.

A utilização de uma sublinguagem de OWL não é excludente, como visto na própria recomendação. Por tanto há interoperabilidade entre as sublinguagens.

> Every legal OWL Lite ontology is a legal OWL DL ontology. Every legal OWL DL ontology is a legal OWL Full ontology. Every valid OWL Lite conclusion is a valid OWL DL conclusion. Every valid OWL DL conclusion is a valid OWL Full conclusion. (W3C Recommendation, 2019c)

2.3 Linked Data

LD são dados que se conectam a outros com semântica explícita, além disso, o conjunto de LD disponível na *internet* é conhecido como *Web Of Data* (WOD) (BIZER *et al.*, 2008). Dessa forma, a disponibilização pública de dados assume o conceito de *Linked Open Data* (LOD) e o mesmo possui uma classificação que pode assumir uma estrela de até cinco estrelas conforme a Tabela 1. Portanto, o LD é obtido a partir do RDF, mas possui mais expressividade com OWL.

Tabela 1 – Classificação do LOD

Classificação	Descrição
⋆	Informação disponível na Web (qualquer formato) sob uma licença aberta
⋆⋆	Informação disponível como dados estruturados (ex. excel no lugar de imagem escaneada)
⋆⋆⋆	Formatos não-proprietários utilizados (ex. CSV e não excel)
⋆⋆⋆⋆	URI's é utilizado para identificar recursos. Isto ajuda as pessoas a apontarem para os recursos
⋆⋆⋆⋆⋆	Dados conectados (LD) com outros dados para prover contexto

Fonte: Adaptado de (FLORIAN; MARTIN, 2012)

2.4 Elevação Semântica

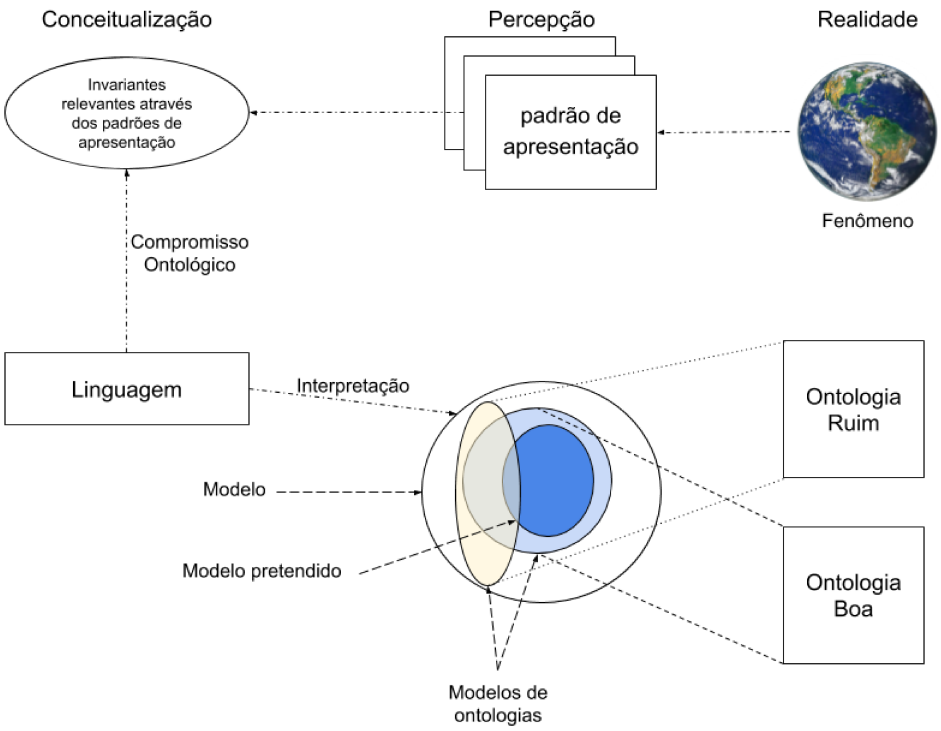
A elevação semântica adiciona significado aos metadados existentes em um conjunto de dados não semânticos. Assim, esse processo aplicado ao arquivo CSV se beneficia dos vocabulários ontológicos e do auxílio de um mapeamento (BRENNAN; FEENEY; GAVIN, 2013; WU *et al.*, 2017; JUNIOR; DEBRUYNE; O'SULLIVAN, 2018). Neste primeiro ponto, Lopez *et al.* (2015) reforçam que, é importante o reaproveitamento dos vocabulários do OWL de modo a manter o reúso do conhecimento. A Seção 2.4.1 discute sobre a modelagem ontológica;

já a Seção 2.4.2 apresenta os conceitos do grafo de conhecimento e como obtê-lo; por fim, a Seção 2.4.3 apresenta a discussão de um arquivo CSV do mundo real que é utilizado como caso de uso.

2.4.1 Modelagem Ontológica

Ontologia é um termo oriundo da filosofia que é o esforço de classificar sistematicamente a existência das coisas em todos seus aspectos (GRUBER, 1993). As ontologias de domínio geram artefato do fenômeno a ser mapeado, sendo necessária a observação da realidade e utilização de uma linguagem que consiga preservar todas as características desejadas em semântica única. A conceitualização deve conter a essência epistêmica do modelo escolhido (ALMEIDA; BAX, 2003b; GUARINO; OBERLE; STAAB, 2009). Resultará em uma ontologia ruim a utilização de elementos ontológicos que não sejam a compreensão total do conjunto. Na modelagem, não se deve alcançar o máximo da especialização, mas generalizar, no limite do possível, o fenômeno observado. Na conceitualização é preciso obter os elementos identificadores compartilhados dos conjuntos. Uma modelagem ontológica satisfatória é alcançada quando o domínio do modelo pretendido está contido dentro da ontologia. A modelagem ontológica segue conforme a Figura 10.

Figura 10 – Modelagem ontológica da realidade



Fonte: Adaptado de (GUARINO; OBERLE; STAAB, 2009)

O artefato obtido pela modelagem deve ser de mesma interpretação tanto pelo ser humano quanto pela máquina. McCusker, Luciano e McGuinness (2011) definem que a interoperabilidade na modelagem ontológica deve possuir três requisitos:

1) Conversão: seja possível converter indivíduos com certo nível de fidelidade.

2) Mapeamento: permita automatização do mapeamento sem perda semântica.

3) Pesquisa: por meio de uma linguagem de consulta, seja possível alcançar qualquer classe, indivíduo ou predicado.

2.4.2 Grafo de Conhecimento

Um grafo de conhecimento possui uma estrutura similar a uma tripla da web semântica. Cada vértice do grafo é equivalente a um sujeito ou objeto, enquanto a aresta representa o predicado (GANGEMI *et al.*, 2016). O termo "grafo de conhecimento" foi cunhado pela empresa Google e desde então assumiu o papel de representar qualquer grafo que possua semântica compartilhada, bem como a representação de entidades e seus relacionamentos ainda que enviesados por diversos tópicos de domínios (EHRLINGER; WÖSS, 2016). Segundo Paulheim (2017), é esperado que os grafos de conhecimentos cubram a maior parte das *coisas* do mundo real e não se limite a um determinado domínio, ainda que relações arbitrárias ocorram. Conforme Auer *et al.* (2018), ontologia não é excludente ao grafo de conhecimento, uma vez que ontologia é uma conceitualização explícita.

De acordo com Uschold e King (1995), para construção de uma ontologia é necessário: (i) identificar o propósito por meio de questões de competência; (ii) capturar a ontologia identificando os conceitos chaves e o relacionamento dos seus respectivos termos; (iii) codificar a ontologia por meio de uma linguagem ontológica; (iv) integrar ontologias existentes; (v) avaliar o resultado.

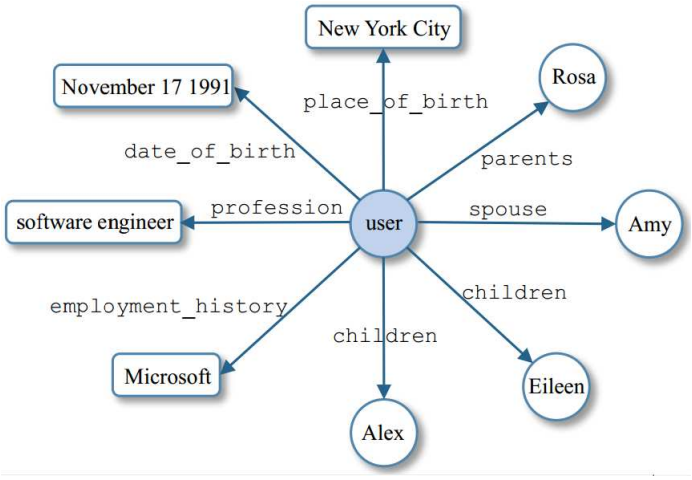
A Figura 11 é a representação gráfica de um grafo de conhecimento sobre um usuário (*user*). O sujeito (*user*) possui literais (retângulos), objetos (círculos) e predicados (setas) (Seção 2.2). Esse grafo de conhecimento informa qual a profissão e local de trabalho. Também expressa o local e data de nascimento. Por fim, indica que o usuário possui uma esposa além de ter dois filhos.

2.4.3 Caso de Uso em CSV

O caso de uso neste trabalho é o registro de policiais mortos em serviço[6], dados esses coletados dos incidentes ocorridos aos profissionais dos Estados Unidos da América. O CSV é a forma tabular dos dados em texto puro, a primeira linha é o cabeçalho e as linhas seguintes são

[6] Disponível em: https://github.com/fivethirtyeight/data/tree/master/police-deaths. Acesso em 01/12/2019.

Figura 11 – Grafo de conhecimento



Fonte: (LI *et al.*, 2014)

os registros de dados. Por último, as colunas ou atributos representam a divisão dos dados e seus respectivos metadados indicados no cabeçalho. A Tabela 3 contém os 3 primeiros registros de dados do CSV após a remoção dos atributos redundantes e limpeza de registros[7].

Tabela 3 – CSV com registros de óbito de policiais

#	person	cause_short	date	dept_name	state
1	Constable Darius Quimby	Gunfire	1791-01-03	Albany County Constable's Office	NY
2	Sheriff Cornelius Hogeboom	Gunfire	1791-10-22	Columbia County Sheriff's Office	NY
3	Deputy Sheriff Isaac Smith	Gunfire	1792-05-17	Westchester County Sheriff's Department	NY

Fonte: Elaborado pelo autor

O atributo # é o identificador único-sequencial dos registros de dados, ao passo que o atributo *person* é o metadado que representa o policial, notadamente o nome; o atributo *cause_short* possui os registros da causa do óbito; *date* possui as datas das ocorrências; *dept_name* são os departamentos que cada policial pertencia; *state* é a unidade federativa que se localiza o departamento. Logo, o #1, ou seja, o primeiro registro nos diz que o policial Constable Darius Quimby do departamento Albany County Constable's Office em NY foi morto por disparo de arma de fogo em 03/01/1791.

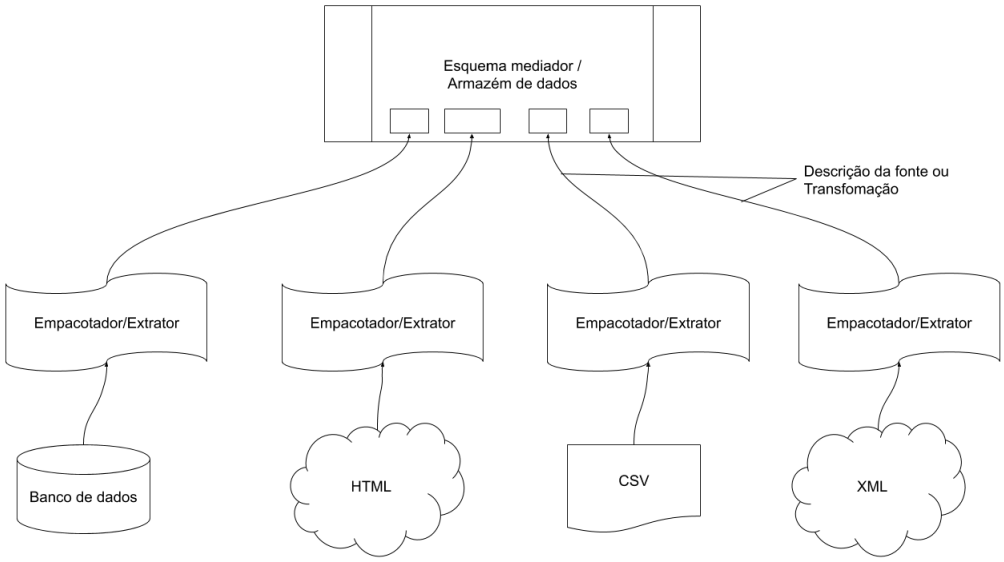
Feita a compreensão dos dados, a discussão sobre o arquivo CSV será retomada na Seção 5.1 com a elevação semântica. O procedimento é realizado com a metodologia da Seção 2.4.2 para construção do grafo de conhecimento. A integração semântica de dados utiliza o grafo para a intermediação do dado tabular para o LD.

[7] É realizado a exclusão de registros com valores vazios e registros de óbitos de cães policiais, pois neste trabalho serão considerados somente humanos.

2.5 Integração Semântica de dados

A integração de dados visa criar uma visão unificada da heterogeneidade dos diversos conjuntos de dados existentes (ALMEIDA; BAX, 2003a; XU *et al.*, 2003; IZZA, 2009). Uma vez que as instituições cada vez mais automatizadas geram diariamente dados estruturados, sendo estes tabulares ou hierárquicos, existem ainda os dados não estruturados que podem ser texto, documento, *email*, imagem, áudio ou vídeo. Os sistemas de informações que utilizam tecnologias diversas, também armazenam os dados em locais diferentes com estrutura respectiva. Antes de tudo, cada técnica de unificação possui um agente de integração (empacotador ou extrator) que obtém os dados das suas devidas fontes de modo a concentrá-los em um armazém de dados ou em um esquema mediador. O processo de integração de dados segue conforme a Figura 12.

Figura 12 – Arquitetura básica para aplicação geral de integração de dados



Fonte: Adaptado de (DOAN; HALEVY; IVES, 2012, p. 10)

A integração de dados tradicional preza pela interoperabilidade da forma, ou seja, todas as tecnologias no processo comunicam-se utilizando um esquema. Ao passo que na integração semântica de dados, além de possuir um esquema, também possui um mediador que mantém o significado dos dados. Izza, Vincent e Burlat (2008) ratificam, a interoperabilidade semântica provê conformidade de significado entre aplicações, ao passo que a interoperabilidade sintática provê uma interface para comunicação. Para realizar a integração semântica de dados em arquivo CSV é necessária a representação semântica do dado tabular (Seção 2.4).

A representação semântica precisa ser mapeada em uma ferramenta para extrair ou empacotar os dados tabulares. Como definido por Bizer, Heath e Berners-Lee (2011), o sucesso do LD está atrelado a facilidade dos usuários em ligar os dados e navegar por eles. Na medida em que

as páginas na web se tornaram mais intuitivas e mais amigáveis, mais adeptos foram angariados para sua utilização. Neste contexto, podemos afirmar que o sucesso da web atual é pelo fato dos usuários finais conseguirem publicar conteúdo. O compreendimento da informação disponível na web é de fácil consumo, gerando mais publicação de hipertextos nesta retroalimentação. Segundo Stone *et al.* (2005), uma boa interface utiliza a experiência dos usuários para ergonomia do *software*. Uma boa ergonomia resulta em uma usabilidade que contribui para a entrega dos resultados almejados. Davies *et al.* (2010) demonstram que a contribuição de usuários finais não técnicos pode aumentar a publicação de mais LD quando as ferramentas oferecem interface adequada. Uma estratégia é encapsular as linguagens complexas de mapeamento e processamento, com uma interface amigável (HEYVAERT *et al.*, 2016).

A integração de dados tradicional utiliza banco de dados relacional para persistência. As tuplas das tabelas estão limitadas ao par atributo/valor (LUDÄSCHER *et al.*, 2006). Para expansão de conhecimento é preciso escalar em metadados e em redundância de dados. Ao passo que uma tripla possui um esquema livre, pois utiliza uma estrutura em forma de grafo que contém sujeito, predicado e objeto, como visto no Seção 2.1. Diante disso, um *triplestore* se faz necessário como repositório para atender a escalabilidade do grafo (SEGARAN; EVANS; TAYLOR, 2009). As triplas são recuperadas por meio da linguagem de consulta *SPARQL Protocol and RDF Query Language* (SPARQL).

3 TRABALHOS CORRELATOS

Este capítulo faz a revisão dos trabalhos que avaliam recursos presentes nas ferramentas que contribuem para integração semântica de dados tabulares, bem como a metodologia utilizada para avaliá-las.

As primeiras linguagens de mapeamento foram feitas para converter bancos de dados relacionais em LD. Em virtude disso, as linguagens de mapeamento foram comparadas com um caso de uso do mundo real na criação de um banco de dados. Desta forma os trabalhos avaliam a utilização de padrões que se destacam na criação do banco de dados e seu equivalente em linguagem de mapeamento para gerar LD (SEQUEDA; PRIYATNA; VILLAZóN-TERRAZAS, 2012; BAGUI; BOURESSA, 2014).

Em Thomas, O'Sullivan e Brennan (2009), as metas (*goals*) avaliam a capacidade das ferramentas em gerar LD. A primeira meta (*ability to express a mapping relation*) diz respeito à habilidade de tratar mapeamento das relações, além disso, é preciso avaliar operadores e funções que expresse a conversão estrutural. A segunda meta (*computationally efficient to proccess*) avalia eficiência computacional para implementar a interoperabilidade ontológica. Já a terceira meta (*sharing and reuse of existing mappings*) é relativa à capacidade da ferramenta compartilhar e reutilizar mapeamentos. A avaliação é aplicada em ferramentas de mapeamento e em outras ferramentas correlatas. Assim o resultado da avaliação é sumarizado em percentual de cumprimento das metas estabelecidas.

Em Hert, Reif e Gall (2011), um arcabouço é construído a partir de recursos que permitem o mapeamento de bancos de dados relacionais para RDF. Dessa forma, linguagens de mapeamento que façam a conversão são analisadas e recebem uma avaliação conceitual em cada recurso, sendo: suporte completo, suporte parcial e sem suporte. Ademais, as linguagens de mapeamento são categorizadas como: *Direct Mapping* quando a linguagem de mapeamento apresenta alguns recursos com ganho de significado, no entanto, não possui recursos adicionais com mais expressividade; *Read-only General-purpose Mapping* têm alto poder expressivo em gerar significado, mas com capacidade unidirecional de mapeamento; *Read-Write General-purpose Mapping* são as linguagens com as características da categoria anterior, mas possui suporte bidirecional de mapeamento; *Special-purpose Mapping* possuem expressividade de significado, contudo, sua aplicabilidade é restrita ao modelo para qual foi desenvolvido. Em suma, os recursos avaliados nas linguagens de mapeamento são: *Logical Table to Class* permite o mapeamento de tabelas que não estão materializadas no banco de dados, por exemplo, *views*; *M:N Relationships* são relações muitos para muitos que necessitam de tabelas de junção, esse recurso avalia a capacidade de mapear as junções de várias tabelas com a anotação semântica correspondente; *Project Attributes* permite projetar somente atributos escolhidos pelo usuário, ou seja, atributos

que armazenam dados sensíveis ou irrelevantes podem ser excluídos do mapeamento; *Select Conditions* permite filtrar registros nas tabelas do banco de dados por meio de uma condição, isso é útil para não gerar LD dos dados que estão desatualizados; *User-defined Instance URIs* permite ao usuário definir a sintaxe que formará o URI das instâncias dos recursos; *Literal to URI* permite gerar URI a partir de valores literais armazenados no banco de dados; *Vocabulary Reuse* permite utilizar vocabulário existente, em contraste à utilização dos nomes das tabelas e atributos para formar o vocabulário; *Transformation Functions* permite aplicar funções de transformação para converter dados para um novo formato ou adequação de uma unidade de medida; *Datatypes* permite explicitamente definir o tipo de dados de destino no mapeamento; *Named Graphs* permite gerar contexto no grafo ao nomear as triplas; *Blank Nodes* permite gerar recursos sem URI, essa técnica é utilizada para representar dados mais complexos; *Integrity Constraints* permite descrever, no mapeamento, explicitamente as restrições de integridade definidas no banco de dados; *Static Metadata* permite incluir metadados que não estão presentes no banco de dados, de modo a inserir informação de licenciamento ou proveniência; *One Table to n Classes* permite desmembrar uma tabela para diversas classes, esse recurso se faz necessário a fim de corrigir erros na modelagem do banco de dados; *Write Support* possui a capacidade de gravar informação no banco de dados.

Michel, Montagnat e Zucker (2014) analisam ferramentas que fazem conversão do banco de dados para RDF, classificando-as conforme as seguintes abordagens: *Direct Mapping*, quando a ferramenta converte diretamente para RDF os dados do banco de dados, utilizando processo automatizado para identificar os metadados que estarão presentes no RDF; *Augmented Direct Mapping* é a elevação da abordagem anterior, ou seja, é utilizado um processo semi-automatizado para enriquecer os dados com semântica; *Domain Semantics-Driven Mapping*, quando a ferramenta utiliza uma linguagem de mapeamento com semântica explícita que define as triplas RDF a partir dos metadados do banco de dados. Os autores ainda definem que independentemente de qual abordagem de mapeamento utilizado, as ferramentas podem utilizar *Data Materialisation* e *On-Demand Mapping* como método para gerar RDF, o primeiro método é conversão de todo o banco de dados em RDF, enquanto o segundo é a conversão de parte dos dados a cada solicitação. Já a recuperação dos dados pode ser na forma de *Query-based access*, quando o retorno é feito na forma de uma consulta. Por fim, na forma *Linked Data* quando o retorno se constitui de triplas. Os autores definem como salientes para as ferramentas, os seguintes recursos: *generation of user defined unique Ids*, *logical table*, *column selection*, *column renaming*, *select conditions*, *vocabulary reuse*, *1 table to n classes*, *many-to-many relation to simple triples*, *blank nodes*, *data types*, *data transformation*, *named graphs*, *user-defined namespaces* e *static metadata*. Todos os recursos citados possuem correspondência aos recursos discutidos no trabalho de Hert, Reif e Gall (2011). Assim, os recursos das ferramentas são avaliadas conceitualmente como: suporte completo, suporte parcial e sem suporte.

Junior *et al.* (2017) avaliam as ferramentas que elevam semanticamente arquivos CSV, para isso utilizam caso de uso do mundo real. Para cada recurso testado é aplicado uma avaliação

utilizando dados tabulares como linha de base. As ferramentas podem oferecer suporte completo, suporte parcial e sem suporte dos seguintes recursos: *M:N Relationships*, *Additional Data*, *Select*, *Filter*, *Literal to IRI*, *Vocabulary Reuse*, *Transformation Functions*, *Data types*, *Named Graphs*, *Blank Nodes* e *Reusability*. Todos os recursos possuem correlação com os recursos avaliados em Hert, Reif e Gall (2011), exceto os recursos *Filter* e *Reusability*. *Filter* provê uma nova abordagem com critérios mais complexos para filtrar registros, ao passo que *Reusability* é a serialização do mapeamento para reuso.

Dimou *et al.* (2018) discutem que o desenho das ferramentas vai além dos fatores inerentes aos algoritmos. Neste sentido, os autores apresentam requisitos não funcionais que influenciam a geração do LD. Assim, utilizam um sistema de transporte, como caso de uso, para avaliar como os fatores agem nos respectivos LD obtidos. As ferramentas podem oferecer suporte total, suporte parcial ou sem suporte aos fatores. O propósito (*purpose*) é o fator que guia a geração do LD, pode ser *production*, quando o LD é gerado em sua origem para disponibilização pública de uma demanda em comum, ou *consumption*, quando o LD é obtido do dado em estado bruto para atender alguma demanda específica com baixa probabilidade de ser reaproveitado em outro projeto. A geração do LD pode assumir a direção (*direction*) *target-centric* quando existe uma ontologia global que mapeia todas as fontes de dados para um único destino, ou *source-centric* quando existe uma ontologia para cada fonte de dados. O fator de materialização (*materialization*) das triplas pode ser *dumping*, quando as triplas são persistidas no *triplestore*, ou *on-the-fly*, quando as triplas são geradas na requisição. Os elementos necessários para gerar o LD podem estar em localizações (*location*) diferentes, por exemplo, os dados podem estar em um local e o processamento ocorrer em outro. Esse fator é *in-situ* se o mesmo servidor possui os dados e a ferramenta, ou seja, a geração do LD é local. Ao passo que a localização é *remote* quando a ferramenta acessa os dados de forma remota para o processamento. O fator que guia (*driving force*) os elementos na geração do LD é *mapping-driven* se a ferramenta utiliza um mapeamento adequado para cada fonte de dados no processamento requisitado. Enquanto o *data-driven* é o envio do dado bruto em forma de requisição para geração do LD. O gatilho (*trigger*) é o fator para iniciar o processamento. A tarefa é *real-time* se disparada por algum evento ou resposta imediata, ou *on-demand* quando um agente dispara o processamento. A dinamicidade (*dynamicity*) das fontes de dados pode afetar a geração do LD. O fator é *static data* quando a dimensão dos dados é previsível. Ou *dynamic data* quando a dimensão dos dados pode crescer ou encolher. A diversidade (*diversity*) é o fator que se refere à capacidade de lidar com os vários formatos de dados existentes. A diversidade é *homogeneity*, quando o processamento ocorre com determinado formato de dado. Ao passo que *heterogeneity* ocorre quando o processamento com diversos formatos de dados. A complexidade (*complexity*) *data* ocorre quando os dados de diferentes tamanhos exigem estratégias de processamento distintas, ou seja, grandes conjuntos de dados devem ser processados diferentemente de pequenos conjuntos de dados. A complexidade *rules* ocorre quando as regras são complexas e o processamento é penalizado.

Para Rashid *et al.* (2020), o valor de cada métrica avaliada está compreendida entre 0 e 1,

dessa forma as abordagens (*approaches*) de integração semântica de dados são organizados em arcabouço de números decimais, além dos valores serem sumarizados por agrupamentos.

Os autores analisam se as abordagens ingerem (*ingestible*) e harmonizam (*harmonize*) os dados, ou seja, existe uma representação coesa do conhecimento das colunas nos conjuntos de dados. Além de permitir a seleção de subconjunto (*subset selection*) de linhas/colunas e possuir suporte na definição de tipos (*data type assignment*). As abordagens necessitam explicitar objetos (*object elicitation*) e explicitar relações (*relation elicitation*) que estão implícitas no conjunto de dados. Os dados devem ser: anotados por valor (*value annotation*); anotação temporal (*time annotation*); anotação espacial (*space annotation*). As abordagens devem reutilizar ontologias ou vocabulários (*Domain knowledge support*). Além de prover utilização de ontologia de topo (*Top-level ontology foundation*). O LD deve ser consultável (*queryable*) e um grafo deve ser materializado (*graph materialization*).

Cada abordagem deve utilizar identificadores únicos persistentes que possam ser acessíveis (*accessible*) na web. Os dados publicados devem ser localizáveis (*findable*). As linguagens devem ser interoperáveis (*interoperable*). O licenciamento dos dados deve ser reusável (*reusable*), ou seja, com uso irrestrito. Os experimentos devem ser reproduzíveis (*reproducible*) conforme a metodologia proposta. O processo da representação de conhecimento deve ser transparente (*transparent*).

A abordagem é agnóstica de domínio (*domain-agnostic*) quando não é restrita a um determinado domínio. Ao passo que a abordagem é agnóstica de ontologia (*ontology-agnostic*) quando há a utilização de diversas ontologias,.

O processo deve produzir proveniência (*provenance*) dos dados, tais como, a obtenção dos dados e equipamentos utilizados. A abordagem deve possuir documentação (*documentation*) das tarefas, além de todos os insumos serem legíveis por máquinas (*machine-readable*), tais como, mapeamento e saídas.

Portanto, esse capítulo fez a compreensão de como os arcabouços comparativos são construídos e qual a metodologia utilizada para comparar as ferramentas ou as abordagens. A discussão dos recursos apresentados nesse capítulo será retomada na Seção 5.3, onde é apresentado o arcabouço comparativo.

4 METODOLOGIA

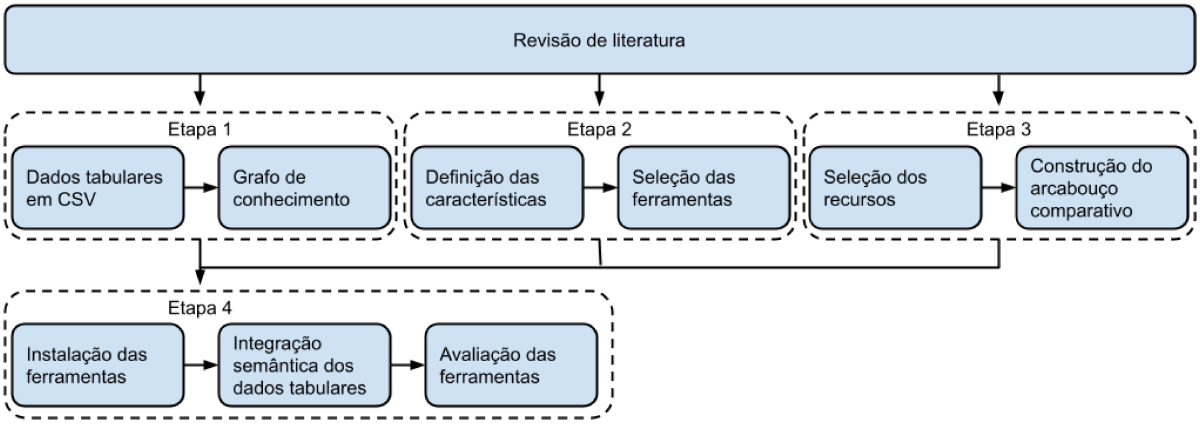
Neste capítulo apresenta-se o percurso metodológico que foi adotado para alcançar os objetivos estabelecidos nesta pesquisa, que é propor um arcabouço comparativo que utiliza uma reta numérica positiva para avaliação. Segundo Gil (2008, p. 8), o que faz o "conhecimento científico distinto dos demais é que tem como característica fundamental a sua verificabilidade", e, para tanto, "torna-se necessário identificar as operações mentais e técnicas que possibilitam a sua verificação", assim como "determinar o método que possibilitou chegar a esse conhecimento".

Seguindo a orientação de (GIL, 2002), esta pesquisa se caracteriza como exploratória e descritiva, no que diz respeito ao objetivo geral da pesquisa, uma vez que recorreu à pesquisa bibliográfica para obter as informações necessárias sobre a classificação dos recursos no contexto da integração semântica, descrevendo os insumos encontrados. Quanto à abordagem do problema, esta pesquisa se caracteriza como qualitativa, pois interpreta os dados que foram levantados (GIL, 2002), usando, para isso, o método procedimental comparativo. Esta pesquisa tem caráter técnico e circunstancial, e, assim, também se caracteriza como aplicada, já que objetiva uma aplicação prática ao pretender prover um arcabouço comparativo como produto final. A base lógica deste trabalho é o método indutivo.

O método comparativo, segundo Gil (2008), é adotado quando se quer analisar dados concretos de fatos, indivíduos, classes ou fenômenos, deduzindo desses dados as similaridades e diferenças entre eles, destacando "os elementos constantes, abstratos e gerais" (LAKATOS; MARCONI, 2007, p. 107).

A Figura 13 demonstra o percurso metodológico adotado. Os retângulos sólidos representam os procedimentos realizados pelo autor. Os retângulos pontilhados são os agrupamentos dos procedimentos que formam as etapas. As setas direcionadas indicam o fluxo e determinam a dependência dos procedimentos. Cada procedimento fornece insumos para outro procedimento, igualmente as etapas fornecem insumo para outra etapa.

Figura 13 – Percurso metodológico



Fonte: Elaborado pelo autor

O início da pesquisa é a revisão de literatura e o compreendimento das bases teóricas de cada etapa. A *etapa 1* busca dados tabulares em CSV para a construção de um grafo de conhecimento que é utilizado na integração semântica. Esta etapa é fundamental, pois cria uma linha de base para o processamento dos dados. Já a *etapa 2* define os critérios que as ferramentas necessitam para a posterior avaliação. Em seguida é realizada pesquisa nos repositórios científicos em busca por ferramentas. Elas são instaladas e um teste de integração semântica é realizado. A *etapa 3* é feita a discussão dos recursos presentes nas ferramentas que foram avaliados nos trabalhos correlatos (Capítulo 3). Os recursos que possuem compatibilidade para manipulação de arquivos CSV são selecionados para formar o arcabouço comparativo. Ademais, os recursos recebem uma classificação e uma pontuação baseada em uma reta numérica positiva. Por fim, na *etapa 4*, as ferramentas são avaliadas à luz do arcabouço comparativo com base no seu desempenho em realizar a integração semântica de dados tabulares em CSV.

5 RESULTADOS E DISCUSSÕES

Neste capítulo apresenta-se os resultados dos procedimentos metodológicos compostos por quatro etapas. A Seção 5.1 apresenta a construção do grafo de conhecimento utilizado na integração semântica dos dados em CSV. Já a Seção 5.2, discute as características necessárias das ferramentas, assim selecioná-las na literatura científica. Em seguida, a Seção 5.3 apresenta a construção do arcabouço comparativo com os recursos obtidos na literatura científica. A Seção 5.4 realiza a avaliação das ferramentas e a sumarização dos resultados. Por fim, a Seção 5.5 discute os resultados dos procedimentos metodológicos.

5.1 Etapa 1: Grafo de Conhecimento

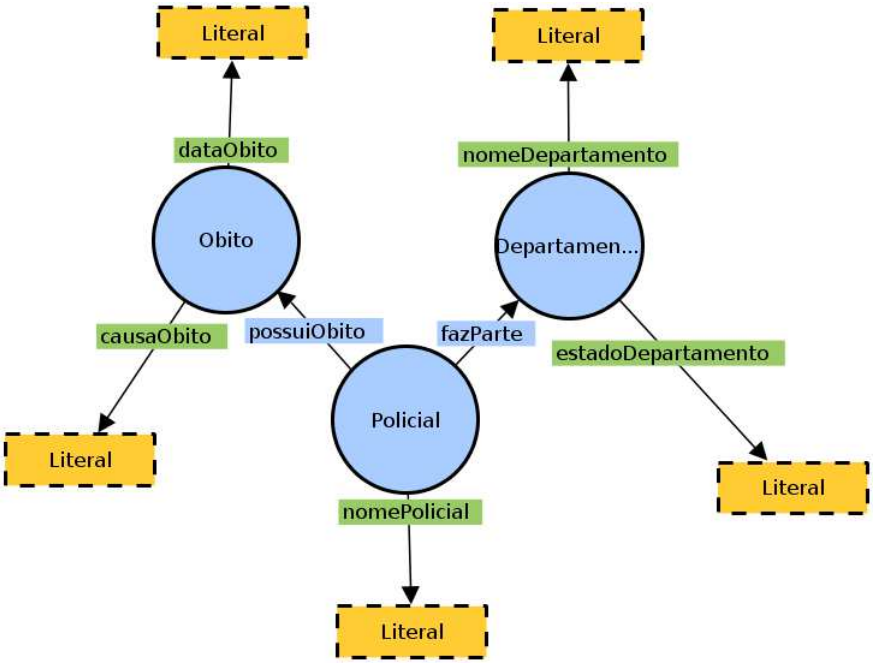
O grafo de conhecimento é utilizado como linha de base na integração semântica, visto que para se integrar dados tabulares é preciso uma representação semântica (Seção 2.4). O arquivo CSV apresentado na Seção 2.4.3 é utilizado como caso de uso. O propósito do grafo de conhecimento é manter o significado de todas as ocorrências de óbitos policiais. Além disso, as questões de competência estão definidas como a seguir. Dado um policial, pergunta-se: Qual o motivo da sua morte?; Quando ocorreu sua morte?; Qual é o departamento que trabalhava quando ocorreu o incidente? Dado um departamento, pergunta-se: Quais os principais motivos de óbitos?; Quais os policiais que morreram?

Nos atributos do CSV foi identificada a classe *Policial* que possui o literal: nome do policial. Depois, a classe *Obito*[1] que representa a morte do policial contendo os literais: data; causa do ocorrido. Por fim, a classe *Departamento*, local onde o policial trabalhava, assim, essa classe possui os literais: nome; estado do departamento. O vocabulário do OWL é base para o mapeamento semântico ou ontológico. A Figura 14 representa graficamente o Grafo de Conhecimento dos Óbitos de Policiais (GCOP).

Outra forma de se obter um grafo de conhecimento é por meio do mapeamento direto ou canônico. Este método preza pela simplicidade no mapeamento ao utilizar os metadados do arquivo CSV na representação semântica (MICHEL; MONTAGNAT; ZUCKER, 2014; BERNABÉ-DÍAZ *et al.*, 2019). A Figura 15 representa graficamente o GCOP utilizando o mapeamento direto.

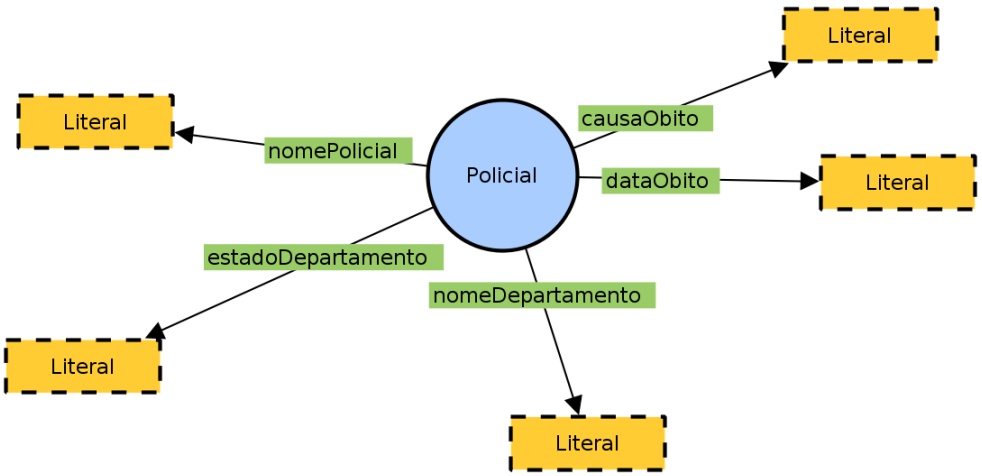
[1] Não se utiliza acentuação, pois parte do sistema de codificação computacional é baseado no *American Standard Code for Information Interchange* (ASCII), na qual não contempla acentuação em sua tabela de caracteres.

Figura 14 – Mapeamento semântico dos óbitos de policiais



Fonte: Elaborado pelo autor

Figura 15 – Mapeamento direto dos óbitos de policiais



Fonte: Elaborado pelo autor

5.2 Etapa 2: Ferramentas para Integração Semântica

Essa etapa seleciona as ferramentas para integração semântica de dados. As ferramentas devem: (i) processar arquivo CSV nativamente[2], ou seja, sem utilização de *software* de terceiros; (ii) possuir interface gráfica para o usuário final; (iii) integrar os dados semanticamente utilizando LD; (iv) ter o código fonte aberto (*open source*). Foram executadas pesquisas nos repositórios científicos que estão disponibilizadas por meio do Portal de Periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)[3] através do acesso remoto via Comunidade Acadêmica Federada[4] (CAFe) com a *string* de busca: (semantic AND data AND tool) OR (“linked data” AND tool) OR (“linked data” AND framework) OR (“linked data” AND map* AND tool) OR (“tabular data” AND “linked data”) OR (OBDA AND CSV). Dessa maneira, foram realizadas leituras no resumo e na introdução de cada trabalho por mais informação sobre cada ferramenta semântica. Após a fase de leitura, realizou-se a instalação de cada ferramenta para testes iniciais para verificar se possuem os critérios exigidos neste trabalho.

As ferramentas a seguir foram selecionadas por atender todos os critérios de seleção. O *Linked Data Reactor*[5] (LD-R) possui uma interface adaptativa ao contexto de uso e a semântica dos dados, assim fornecendo melhor usabilidade para realizar as tarefas (KHALILI, 2016). O *LinkedPipes ETL*[6] (LP-ETL) executa o *Extract, Transformation, Load* (ETL) por meio de fluxos de trabalhos, ou seja, em um determinado fluxo de trabalho programado a ferramenta extrai da fonte de dados os registros, após aplicar as transformações necessárias e por fim faz a carga no repositório de destino (KLÍMEK; ŠKODA; NEČASKÝ, 2016). O *Human-Aware Data Acquisition*[7] (HADatAc) integra semanticamente os dados fornecendo gráficos analíticos (PINHEIRO *et al.*, 2018). *Karma*[8] possibilita visualizar os dados tabulares e o mapeamento semântico simultaneamente. O mapeamento é representado por um grafo em formato de árvore, além disso, a ferramenta constrói o mapeamento baseado em mapeamento anterior (KNOBLOCK *et al.*, 2011). *Morph-CSV*[9] é uma ferramenta que promove melhoria de performance na recuperação dos dados em CSV utilizando a abordagem *Ontology-Based Data Access* (OBDA) (CHAVES-FRAGA *et al.*, 2020). *OpenRefine*[10] realiza tratamento de dados, além de permitir a combinação e transformação de fontes de dados. A ferramenta necessita do *plug-in RDF-Extension*[11] para manipular

[2] Nativo é relativo às ferramentas utilizarem algoritmos ou *plug-ins* de forma que seja transparente para o usuário, ou seja, não é necessário configurar outros *softwares* além da própria ferramenta.
[3] Disponível: http://www.periodicos.capes.gov.br. Acesso em 01/09/2020.
[4] Disponível: https://www.rnp.br/servicos/servicos-avancados/cafe. Acesso em 01/09/2020.
[5] Utilizada a versão 1.3.10. Disponível em: https://github.com/ali1k/ld-r. Acesso em 01/12/2020.
[6] Utilizada a versão em desenvolvimento. Disponível em: https://github.com/linkedpipes/etl. Acesso em 06/12/2020
[7] Utilizada a versão 1.2.4. Disponível em: https://github.com/paulopinheiro1234/hadatac. Acesso em 07/12/2020
[8] Utilizada a versão 2.4. Disponível em: https://github.com/usc-isi-i2/Web-Karma. Acesso em 01/12/2020.
[9] Utilizada a versão 1.1.0. Disponível em: https://github.com/oeg-upm/morph-csv. Acesso em 07/12/2020.
[10] Utilizada a versão 3.4.1. Disponível em: https://github.com/OpenRefine/OpenRefine. Acesso em 05/12/2020.
[11] Utilizada a versão 1.3.0. Disponível em: https://github.com/stkenny/grefine-rdf-extension. Acesso em 05/12/2020.

as triplas e o *OpenRefine Metadata Extension*[12] para suporte aos *triplestores* (KUSUMASARI *et al.*, 2016). *Silk Link Discovery Framework*[13] (Silk) cria ligação de diversas fontes de dados, além de executar transformações nos dados (VOLZ *et al.*, 2009).

As ferramentas a seguir foram recuperadas nas buscas, no entanto, na utilização apresentaram problemas. *Datalift* é mencionado na literatura, mas não foi localizado o código fonte da ferramenta e nenhum instalador, uma vez que todos os endereços eletrônicos estão "quebrados" (KEPEKLIAN; BIHANIC; TRONCY, 2014). O *Linked Data Integration Framework* (LDIF) foi instalado, no entanto, o mapeamento para fonte de dados em CSV não foi possível (SCHULTZ *et al.*, 2011). A documentação do LDIF não foi suficiente para realizar a integração de dados. O autor tentou contato com o desenvolvedor do LDIF, no entanto, sem sucesso. *Ontop* é outra ferramenta baseada em OBDA que permite a utilização de CSV como fonte de dados, mas isso se deve com auxílio de sistema federado de acesso a dados (CALVANESE *et al.*, 2017). Por este motivo não atende ao segundo critério de seleção.

5.3 Etapa 3: Arcabouço Comparativo

Essa etapa discute o arcabouço comparativo, bem como seus recursos. Os recursos foram obtidos nos trabalhos do Capítulo 3. Entretanto, os recursos não fazem parte do arcabouço comparativo caso eles estejam intrinsecamente ligados a uma tecnologia, por exemplo banco de dados relacional, e não possua a mesma aplicabilidade aos arquivos CSV. Os recursos estão agrupados por semelhança, além de pontuá-los em uma reta numérica positiva.

5.3.1 Recursos para Metadados

Recursos para metaDados (RD) contém funcionalidades voltadas ao gerenciamento da informação sobre os dados. *Filtrar metadados* (RD.I) é um recurso que permite selecionar colunas ou atributos no arquivo CSV que não estarão presentes no LD. Este recurso se faz útil para não integrar dados com informações sensíveis ou atributos sem relevância. A ferramenta recebe 1 ponto caso possua o recurso e 0 caso contrário. Nos trabalhos relacionados o RD.I é citado como: *project attributes* em Hert, Reif e Gall (2011); *column selection* em Michel, Montagnat e Zucker (2014); *filter* em Junior *et al.* (2017); *subset selection* em Rashid *et al.* (2020).

Criar metadados (RD.II) se faz útil para enriquecer o LD com informação que não está presente no arquivo CSV. A ferramenta recebe 1 ponto caso possua o recurso e 0 caso contrário. Nos trabalhos relacionados o RD.II é citado como: *static metadata* em Hert, Reif e Gall (2011); *static metadata* em Michel, Montagnat e Zucker (2014); *additional data* em Junior *et al.* (2017).

[12] Utilizada a versão 1.6.0. Disponível em: https://github.com/FAIRDataTeam/OpenRefine-metadata-extension. Acesso em 05/12/2020.

[13] Utilizada a versão 3.2.1. Disponível em: https://github.com/silk-framework/silk. Acesso em 06/12/2020

Proveniência (RD.III) é o recurso utilizado para registrar histórico dos dados, como foram obtidos e outras informações pertinentes à sua origem (Seção 2.1). A ferramenta recebe 1 ponto caso possua o recurso e 0 caso contrário. Nos trabalhos relacionados o RD.III é citado como: *static Metadata* em Hert, Reif e Gall (2011); *additional Data* em Junior *et al.* (2017); *provenance* em Rashid *et al.* (2020).

Tipar literal (RD.IV) é necessário, pois os dados no LD utilizam XSD (Seção 2.1) para tipar, diferentemente dos tipos implícitos presentes no CSV. A ferramenta recebe 1 ponto caso possua o recurso e 0 caso contrário. Nos trabalhos relacionados o RD.IV é citado como: *datatypes* em Hert, Reif e Gall (2011); *data types* em Michel, Montagnat e Zucker (2014); *data types* em Junior *et al.* (2017); *data type assignment* em Rashid *et al.* (2020).

Blank nodes (RD.V) é uma representação sem identificar os recursos no LD (Seção 2.1). A ferramenta recebe 1 ponto caso possua o recurso e 0 caso contrário. Nos trabalhos relacionados o RD.V é citado como: *blank nodes* em Hert, Reif e Gall (2011); *blank nodes* em Michel, Montagnat e Zucker (2014); *blank nodes* em Junior *et al.* (2017).

Named graphs (RD.VI) é um agrupamento nomeado de triplas (Seção 2.1). Este recurso enriquece a representação do conhecimento. Assim a ferramenta recebe 1 ponto caso possua o recurso e 0 caso contrário. Nos trabalhos relacionados o RD.VI é citado como: *named graphs* em Hert, Reif e Gall (2011); *named graphs* em Michel, Montagnat e Zucker (2014); *named graphs* em Junior *et al.* (2017).

A distribuição sumarizada de pontos do RD.I, RD,II, RD.III, RD.IV e RD.V segue conforme a Tabela 4.

Tabela 4 – Distribuição de pontos do RD.I, RD,II, RD.III, RD.IV, RD.V e RD.VI

Recurso	Sigla	Ausente	Presente
Filtrar metadados	RD.I	0	1
Criar metadados	RD.II	0	1
Proveniência	RD.III	0	1
Tipar literal	RD.IV	0	1
Blank nodes	RD.V	0	1
Named graphs	RD.VI	0	1

Fonte: Elaborado pelo autor

URI (RD.VII) é a identificação dos recursos no LD (Seção 2.1). Assim, uma ferramenta pode fornecer método *automático* (RD.VII.1), quando a ferramenta define a construção do URI, ou permite ao *usuário* (RD.VII.2) definir um modelo de construção do URI. O identificador customizado tem a maior nota. Pois, possui mais adaptabilidade no processo ao considerar que atende as peculiaridades inerentes ao projeto. A ferramenta recebe 1 ponto caso forneça o RD.VII.1 e 2 pontos para RD.VII.2. Nos trabalhos relacionados o RD.VII é citado como: *literal to URI* em Hert, Reif e Gall (2011); *generation of user defined unique ids* em Michel, Montagnat

e Zucker (2014); *literal to IRI* em Junior *et al.* (2017). A distribuição de pontos do RD.VII segue conforme a Tabela 5.

Tabela 5 – Distribuição de pontos do RD.VII

Recurso	Sigla	Automático	Usuário
URI	RD.VII	1	2

Fonte: Elaborado pelo autor

5.3.2 Recursos para Mapeamento

Recursos para Mapeamento (RM) estão relacionados à representação da conversão do dado tabular sem semântica explícita para LD. Nos trabalhos relacionados o RM é citado como: *mapping* em Hert, Reif e Gall (2011); *direct Mapping*, *augmented direct mapping* e *domain semantics-driven mapping* em Michel, Montagnat e Zucker (2014); *mapping-driven* e *data-driven* em Dimou *et al.* (2018); *approaches* em Rashid *et al.* (2020). O RM.I define o *tipo de mapeamento* realizado, ou seja, *mapeamento direto*, *mapeamento direto aumentado* ou *mapeamento semântico* (Seção 2.4.2). O mapeamento *direto* (RM.I.1) é a replicação da estrutura tabular para o LD. A ferramenta que oferece este recurso recebe 1 ponto. O mapeamento *direto aumentado* (RM.I.2) é o mapeamento *direto*. Entretanto, a ferramenta oferece algum enriquecimento semântico nos dados tabulares, como a explicitação de algumas relações, tal como a relação de subsunção. Logo a ferramenta que oferece este recurso recebe 2 pontos. O mapeamento é *semântico* (RM.I.3) quando a ferramenta mapeia os dados com conceitos de uma ontologia de domínio (Seção 2.4.1), recebendo assim 3 pontos a ferramenta que disponibiliza este recurso. A Tabela 6 possui a sumarização dos pontos do RM.I.

Tabela 6 – Distribuição de pontos do RM.I

Recurso	Sigla	Direto	Direto aumentado	Semântico
Tipo de mapeamento	RM.I	1	2	3

Fonte: Elaborado pelo autor

A interface gráfica no mapeamento (RM.II) auxilia o usuário minimizando os erros ao realizar a intermediação do dado tabular para LD. Dessa forma, a ferramenta recebe 1 ponto no RM.II caso ofereça o recurso e 0 caso contrário. O RM.II não é citado nos trabalhos relacionados, no entanto, se faz necessário por auxiliar o usuário no processo.

O mapeamento pode ocorrer em dois momentos antes da importação dos dados tabulares, chamado de *mapeamento pré* (RM.III), ao passo que o *mapeamento pós* (RM.IV) ocorre quando o dado tabular já está carregado. Os itens RM.III e RM.IV não são excludentes, assim a ferramenta recebe 1 ponto para cada recurso presente e 0 caso contrário. A importância da ordem de

execução de determinados recursos é salientado como *generation's execution* em Dimou *et al.* (2018). A Tabela 7 sintetiza a distribuição de pontos do RM.II, RM.III e RM.IV

Tabela 7 – Distribuição de pontos do RM.II, RM.III e RM.IV

Recurso	Sigla	Ausente	Presente
Interface gráfica	RM.II	0	1
Mapeamento pré	RM.III	0	1
Mapeamento pós	RM.IV	0	1

Fonte: Elaborado pelo autor

5.3.3 Recursos para Processamento

Os Recursos para Processamento (RP) são os métodos para integrar semanticamente os dados. O *processamento em lote* (RP.I) é acionado para integrar os dados sem interação com usuário. É utilizado para integrar um volume de dados que pode levar horas, dessa forma libera o usuário para outras atividades. O *processamento online* (RP.II) permite interação do usuário e disponibiliza *feedback* durante e após o processamento. Assim, o usuário pode fazer experimentações com imediata visualização dos resultados. A ferramenta recebe 1 ponto para cada recurso presente e 0 caso contrário, assim a distribuição de pontos segue em conformidade com a Tabela 8. Os recursos de processamento são discutidos como *trigger* em Dimou *et al.* (2018).

Tabela 8 – Distribuição de pontos do RP.I e RP.II

Recurso	Sigla	Ausente	Presente
Processamento em lote	RP.I	0	1
Processamento *online*	RP.II	0	1

Fonte: Elaborado pelo autor

5.3.4 Recursos para Persistência

Os Recursos para peRsistência (RR) são referentes a gravação do LD em um *triplestore* (Seção 2.5). Registros inconsistentes ou desatualizados no arquivo CSV não deve ser integrado semanticamente, por isso *filtrar registros* (RR.I) deve ser um recurso presente na ferramenta. Este recurso é salientado como *select conditions* em Hert, Reif e Gall (2011); *select conditions* ems Michel, Montagnat e Zucker (2014); *filter* em Junior *et al.* (2017); *subset selection* em Rashid *et al.* (2020). Caso a ferramenta possua o PR.I recebe 1 ponto e 0 caso contrário. Neste sentido a Tabela 9 representa a distribuição.

Tabela 9 – Distribuição de pontos do RR.I

Recurso	Sigla	Ausente	Presente
Filtrar registro	RR.I	0	1

Fonte: Elaborado pelo autor

A ferramenta deve estabelecer nativamente conexão com o *triplestore*, assim é analisado se a ferramenta possui suporte a um único *triplestore* (RR.II.1) ou se possui suporte a múltiplos *triplestores* (RR.II.2). Os trabalhos relacionados não lidam com a integração semântica de dados, portanto, o RR.II não se baseia em nenhuma avaliação prévia, mas se faz necessário para avaliar a capacidade das ferramentas. Dessa forma a ferramenta recebe 1 ponto para o RR.II.1 e 2 pontos para RR.II.2. A Tabela 10 sumariza a distribuição de pontos.

Tabela 10 – Distribuição de pontos do RR.II

Recurso	Sigla	Único	Múltiplos
Suporte ao *triplestore*	RR.II	1	2

Fonte: Elaborado pelo autor

5.3.5 Recursos para Serialização

Os Recursos para Serialização (RS) estão ligados à materialização dos dados em memória para arquivo em disco. O recurso de *serialização do CSV* (RS.I) está presente quando a ferramenta permite salvar os dados tabulares em seu estado original, ao passo que a *serialização do LD* (RS.II) é o armazenamento em disco dos dados semânticos. Por fim, a *serialização do mapeamento* (RS.III) salva em disco a representação da conversão do dado tabular para LD. A ferramenta recebe 1 ponto para cada recurso presente e 0 caso contrário. Assim a distribuição de pontos segue em conformidade com a Tabela 11. A serialização é discutida como: *reusability* em Junior *et al.* (2017); *materialization* em Dimou *et al.* (2018); *graph materialization* em Rashid *et al.* (2020).

Tabela 11 – Distribuição de pontos do RS.I, RS.II e RS.III

Recurso	Sigla	Ausente	Presente
Serialização do CSV	RS.I	0	1
Serialização do LD	RS.II	0	1
Serialização do mapeamento	RS.III	0	1

Fonte: Elaborado pelo autor

5.4 Etapa 4: Avaliação das ferramentas

Essa seção verifica cada recurso existente nas ferramentas (Seção 5.2) por meio de testes, aferições e execução das tarefas necessárias. Cada ferramenta recebe a nota (Seção 5.3) pela presença do recurso ou como o recurso foi implementado.

Todas as ferramentas, ou seja, *Karma*; *HADatAc*; *LD-R*; *LP-ETL*; *Morph-CSV*; *OpenRefine*; *Silk* possuem recurso para filtrar metadados (RD.I), assim é possível selecionar quais colunas do arquivo CSV estará no LD. As ferramentas recebem 1 ponto no RD.I.

O RD.II avalia se a ferramenta permite criar metadados que não estão presentes no arquivo CSV. Nesse contexto, o *LD-R* não possui este recurso, portanto, não pontua no RD.II. As demais ferramentas - *HADatAc*; *Karma*; *LP-ETL*; *Morph-CSV*; *OpenRefine*; *Silk* - possuem o recurso, recebendo cada uma 1 ponto no RD.II.

Todas as ferramentas permitem gerar a proveniência (RD.III) dos dados. Assim, cada ferramenta recebe 1 ponto neste recurso. Igualmente, todas as ferramentas permitem tipar literais (RD.IV), pontuando 1 ponto no recurso.

Recurso para *blank nodes* (RD.V) não está disponível no *HADatAc* e no *LD-R*, logo não recebem ponto neste recurso, enquanto as ferramentas restantes recebem 1 ponto.

As ferramentas *LP-ETL* e *Morph-CSV* possuem suporte ao RD.VI (Named Graphs), portanto, cada ferramenta recebe 1 ponto, enquanto as ferramentas restantes não recebem ponto.

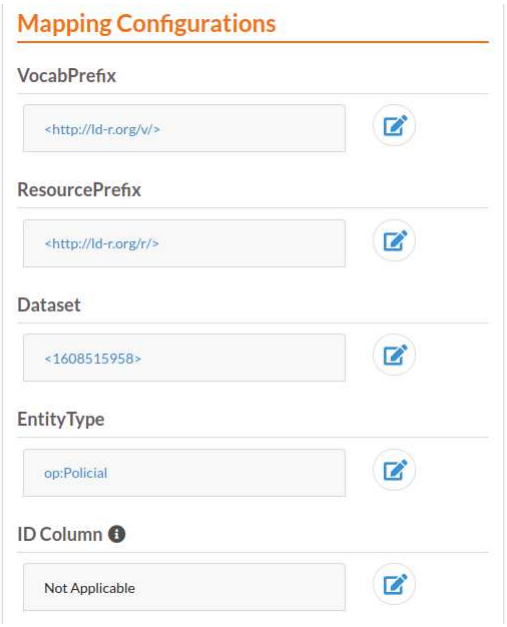
A estratégia de gerar o URI (RD.VII) automaticamente é adotado por: *Karma* e *HADatAc*. Por isso, cada ferramenta recebe 1 ponto, ao passo que as restantes das ferramentas permitem ao usuário definir a construção do URI, assim *LD-R*; LP-ETL; Morph-CSV; OpenRefine; Silk recebem 2 pontos no RD.VII.

O *LD-R* é a única ferramenta que utiliza o mapeamento (RM.I) direto aumento, recebendo 2 pontos. As restantes das ferramentas, ou seja, *HADatAc*; *LP-ETL*; *Karma*; *Morph-CSV*; *OpenRefine*; *Silk* recebem 3 pontos no RM.I. Cada ferramenta implementa o mapeamento utilizando métodos distintos. Conforme a Figura 16, não é possível para o usuário especificar como os literais devem compor o LD, ou seja, automaticamente o *LD-R* gera os predicados para os literais. Já a relação de subsunção é definida pelo campo *EntityType*.

O *HADatAc* utiliza um dicionário semântico de dados para construir o GCOP. Conforme a Figura 17, o dicionário é uma tabela contendo: *Column* onde declara os atributos do arquivo CSV, ou uma classe que deve iniciar com dupla interrogação; *Attribute* define qual o tipo do literal; *attributeOf* especifica a qual classe pertence o sujeito em *Column*; *Entity* é a classe que estabelece a relação de subsunção da classe com dupla interrogação.

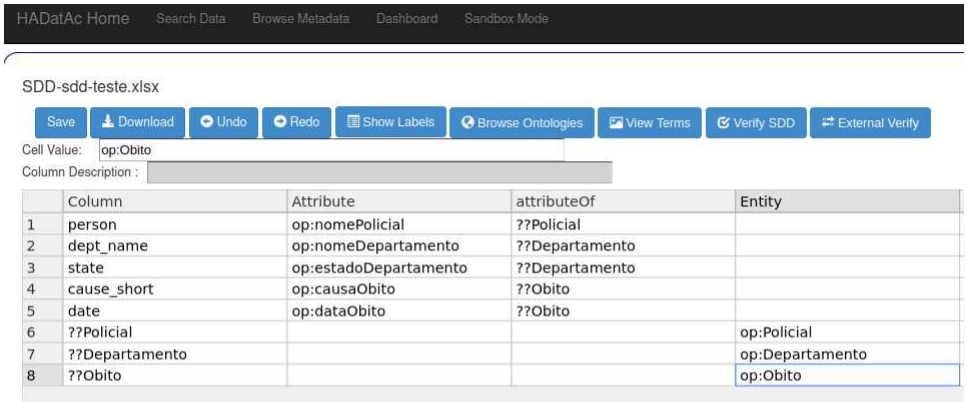
A utilização de um grafo em formato de árvore foi adotado pelo *Karma*. Conforme a Figura 18, a árvore se conecta aos dados tabulares, os vértices correspondem às classes ou aos

Figura 16 – Mapeamento do GCOP pelo LD-R



Fonte: Elaborado pelo autor

Figura 17 – Mapeamento do GCOP pelo HADatAc

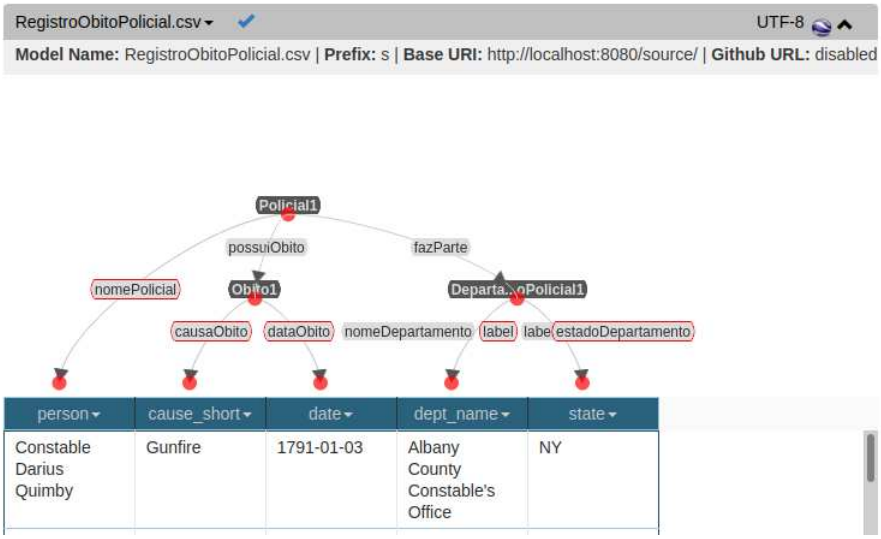


	Column	Attribute	attributeOf	Entity
1	person	op:nomePolicial	??Policial	
2	dept_name	op:nomeDepartamento	??Departamento	
3	state	op:estadoDepartamento	??Departamento	
4	cause_short	op:causaObito	??Obito	
5	date	op:dataObito	??Obito	
6	??Policial			op:Policial
7	??Departamento			op:Departamento
8	??Obito			op:Obito

Fonte: Elaborado pelo autor

literais, ao passo que as arestas são os predicados. O ajuste fino dos elementos do mapeamento é realizado em diversas telas de configuração.

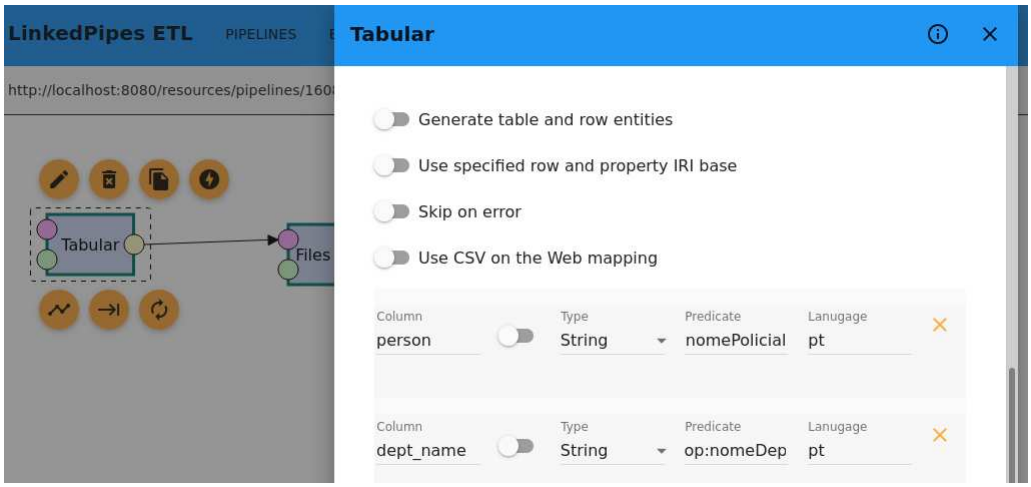
Figura 18 – Mapeamento do GCOP pelo Karma



Fonte: Elaborado pelo autor

O *LP-ETL* utiliza uma cadeia de transformação de valores permitindo o mapeamento semântico. A Figura 19 demonstra uma "caixa de transformação". Cada caixa manipula o dado para construir o grafo de conhecimento desejado. Assim, várias transformações são necessárias para se obter o GCOP.

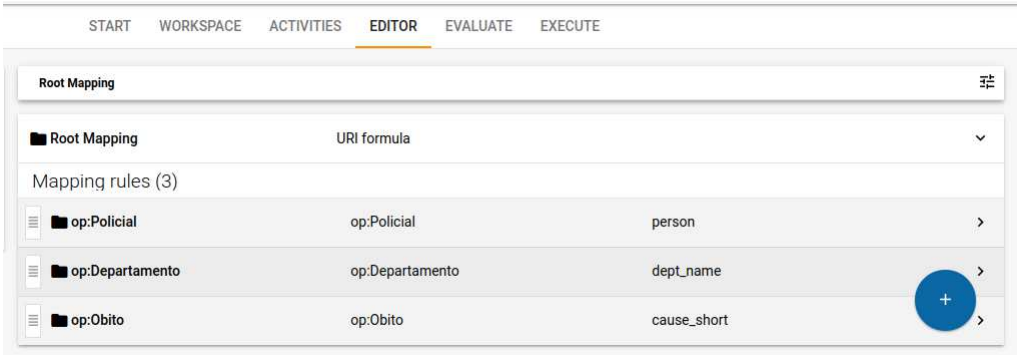
Figura 19 – Mapeamento do GCOP pelo LP-ETL



Fonte: Elaborado pelo autor

A transformação de dados no *Silk* é iniciada pelos sujeitos. A Figura 20 apresenta as classes *Policial*, *Departamento* e *Obito* do GCOP. Os predicados e os literais estão mapeados "dentro" de cada transformação.

Figura 20 – Mapeamento do GCOP pelo Silk



Fonte: Elaborado pelo autor

O *Morph-CSV* não possui interface gráfica para o mapeando, logo o GCOP é feito em um arquivo texto. A Figura 21 demonstra parte da estrutura hierárquica do mapeamento, que é formado por palavras chaves: "policial" e "departamento" são itens que abrem elementos da classe *Policial* e *Departamento* do GCOP, respectivamente; "sources" define a fonte dos dados; "s" é a composição do sujeito; "p" define o predicado para o sujeito; "o" é o objeto que pode ser outra classe ou um literal.

OpenRefine utiliza um formulário para o mapeamento, a Figura 22 ilustra o mapeamento do GCOP com: a coluna da esquerda que representa o sujeito; a coluna do meio é o predicado; a última coluna é o objeto.

Figura 21 – Mapeamento do GCOP pelo Morph-CSV

```
mappings:
  policial:
    sources:
      - [RegistroObitoPolicial.csv~csv]
    s: op:$(person)
    po:
      - [a, op:Policial]
      - [op:nomePolicial, $(person)]
      - p: op:fazParte
        o:
          mapping: departamento
          condition:
            function: equal
            parameters:
              - [str1, $(dept_name), s]
              - [str2, $(dept_name), o]
      - p: op:possuiObtido
        o:
          mapping: obito
          condition:
            function: equal
            parameters:
              - [str1, $(cause_short), s]
              - [str2, $(cause_short), o]
  departamento:
    sources:
      - [RegistroObitoPolicial.csv~csv]
    s: op:$(dept_name)
    po:
      - [a, op:Departamento]
```

Fonte: Elaborado pelo autor

Figura 22 – Mapeamento do GCOP pelo OpenRefine

Fonte: Elaborado pelo autor

Morph-CSV é a única ferramenta sem interface gráfica (RM.II) que auxilia no mapeamento, não pontuando. As outras ferramentas recebem 1 ponto.

O *LD-R* não possui recurso para mapeamento antes da carga (RM.III), mas possui mapeamento após o carregamento dos dados (RM.IV). Logo, o *LD-R* recebe 0 ponto no RM.III

e 1 ponto no RM.IV. Já o *Morph-CSV* possui o RM.III (recebe 1 ponto), entretanto não possui o RM.IV (não recebe ponto). As outras ferramentas possuem tanto o RM.III quanto o RM.IV, assim cada ferramenta recebe 1 ponto em cada recurso.

O recurso para processamento em lote (RP.I) não está presente no *HADatAc*, no *LD-R* e no *Morph-CSV*, dessa forma cada ferramenta não pontua. Conforme a Figura 23, as ferramentas apresentam formas diferentes para chamar o processo: o *Karma* (a) e o *Silk* (c) utilizam linha de comando; o *OpenRefine* (b) utiliza *script* de execução; o *LP-ETL* inicia a tarefa com a chamada de URL configurado no projeto. Dessa forma, *LP-ETL*; *Karma*; *OpenRefine*; *Silk* recebem 1 ponto. Todas as ferramentas possuem processamento online, pontuando no 1 ponto no RP.II.

Figura 23 – Chamada de processamento em lote

(a) Karma

```
mvn exec:java -Dexec.mainClass="edu.isi.karma.rdf.OfflineRdfGenerator"
-Dexec.args="
--sourcetype JSON
--filepath \"~/dataset.json\"
--modelfilepath \"~/model-dataset.n3\"
--sourcename dataset
--outputfile dataset-rdf.n3
--JSONOutputFile dataset-rdf.json" -Dexec.classpathScope=compile
```

(b) OpenRefine

```
require 'refine'
prj = Refine.new('project_name' => 'cleanup', 'file_name' => 'dataset.csv')
prj.apply_operations('operations.json')
puts prj.export_rows('csv')
```

(c) Silk

```
java -DconfigFile=file.xml -DlinkSpec=InterlinkID -Dthreads=1 -DlogQueries=true -
    ↪ Dreload=true -jar silk.jar
```

Fonte: Elaborado pelo autor

O *HADatAc* e o *LD-R* não permitem filtrar os registros (RR.I) do arquivo CSV, não pontuando no RR.I. Já as restantes das ferramentas recebem 1 ponto. O *karma* possui suporte unicamente para o *OpenRDF*, ao passo que o *HADatAc* possui suporte somente para o *Blaze-Graph*. Portanto, ambas as ferramentas recebem 1 ponto no RR.II. O *Morph-CVS*, como OBDA, funciona como um mediador (Seção 2.5), ou seja, as triplas são oriundas da ferramenta como uma visão (view), dispensando um *triplestore*. As restantes das ferramentas possuem suporte a múltiplos *triplestores*. Em suma, *LD-R*; *LP-ETL*; *Morph-CSV*; *OpenRefine*; *Silk* recebem 2 ponto no RR.II.

O *LD-R* não possui nenhum recurso para serialização - logo não pontua no RS.I; RS.II; RS.III. O *HADatAc* não serializa LD (RS.II), recebendo 1 ponto no RS.I e RS.III, mas não no RS.II. Todas as outras ferramentas recebem 1 ponto no RS.I; RS.II e RS.III.

O arcabouço comparativo é representado pela Tabela 12. O *LD-R* possui o resultado menos favorável, uma vez que possui pouco mais da metade dos recursos do arcabouço. Além

de não possuir nenhum recurso da classe serialização. O *LP-ETL* é a ferramenta mais completa, única com todos os recursos. As ferramentas *OpenRefine* e *Silk* não possuem o RD.VI (*Named Graph*), com isso ambas ficam em segundo lugar.

Tabela 12 – Arcabouço comparativo de ferramentas para integração semântica

Classe	Recurso	HADatAc	Karma	LD-R	LP-ETL	Morph-CSV	OpenRefine	Silk
Metadados	RD.I	1	1	1	1	1	1	1
	RD.II	1	1	0	1	1	1	1
	RD.III	1	1	1	1	1	1	1
	RD.IV	1	1	1	1	1	1	1
	RD.V	0	1	0	1	1	1	1
	RD.VI	0	0	0	1	1	0	0
	RD.VII	1	1	2	2	2	2	2
	Subtotal	5	6	5	8	8	7	7
Mapeamento	RM.I	3	3	2	3	3	3	3
	RM.II	1	1	1	1	0	1	1
	RM.III	1	1	0	1	1	1	1
	RM.IV	1	1	1	1	0	1	1
	Subtotal	6	6	4	6	4	6	6
Processamento	RP.I	0	1	0	1	0	1	1
	RP.II	1	1	1	1	1	1	1
	Subtotal	1	2	1	2	1	2	2
Persistência	RR.I	0	1	0	1	1	1	1
	RR.II	1	1	2	2	2	2	2
	Subtotal	1	2	2	3	3	3	3
Serialização	RS.I	1	1	0	1	1	1	1
	RS.II	0	1	0	1	1	1	1
	RS.III	1	1	0	1	1	1	1
	Subtotal	2	3	0	3	3	3	3
	Total	15	19	12	22	19	21	21
	Cobertura	68,18%	86,36%	54,55%	100,00%	86,36%	95,45%	95,45%

Fonte: Elaborado pelo autor

5.5 Discussão

Esta seção discute os resultados e achados no desenvolver do trabalho. Também é feita síntese do objetivo principal.

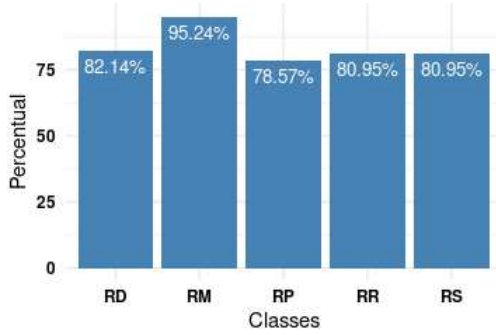
A primeira etapa na metodologia é a seleção do arquivo CSV para integração semântica e sua representação semântica em forma de grafo de conhecimento. A utilização do grafo de conhecimento permite início imediato da integração semântica. Conforme discutido na Seção 2.4.2, a modelagem para se obter o grafo se restringe ao conhecimento pontual. No entanto, o conhecimento deve evoluir e as generalizações de uma ontologia de domínio devem ser desenvolvidas para alcançar um modelo mais próximo da realidade (Seção 2.4.1). Em todos os casos o grafo de conhecimento contribuiu para a análise pragmática das ferramentas com avaliação mais próxima do mundo real.

A segunda etapa é a seleção das ferramentas com características para integração semântica de dados tabulares. As características das ferramentas e os componentes para integração semântica é discutida na Seção 2.5. A definição de critérios e a seleção é apresentada na Seção 5.2. Na pesquisa nos repositórios digitais científicas retornaram somente duas ferramentas que adotaram o modelo de esquema mediador (OBDA). Dessas duas, somente o *Morph-CSV* faz a integração semântica dos dados tabulares nativamente, pois o *Ontop* necessita de *software* de terceiros. Em suma, o *esquema mediador* é 1 em 7 das ferramentas para integração semântica ou pouco mais de 14%. As ferramentas que adotam o *armazém de dados* representam quase 86% das ferramentas.

A terceira etapa é o mapeamento pela literatura científica dos recursos das ferramentas semânticas. Os recursos obtidos no Capítulo 3 possuem fortes influências das características dos bancos de dados relacionais. Na Seção 5.3 é a seleção dos recursos e a proposta do arcabouço comparativo. Nesta proposta é apresentado dois novos recursos originais: interface gráfica para o mapeamento (RM.II); suporte ao *triplestore* (RR.II).

Na quarta e última etapa as ferramentas para integração semântica são analisadas à luz do arcabouço comparativo proposto. Os resultados identificaram as diferenças das ferramentas, além de destacar a ferramenta com todos os recursos e a ferramenta com menos recursos. Conforme a Figura 24, os recursos para mapeamento possuem maior presença nas ferramentas com 95,25%. Os recursos para metadados estão em segundo lugar com 82,14%. Em terceiro lugar estão os recursos para persistência e serialização, ambos com 80,95%. Em último lugar, com 78,57%, os recursos para processamento estão menos presentes nas ferramentas.

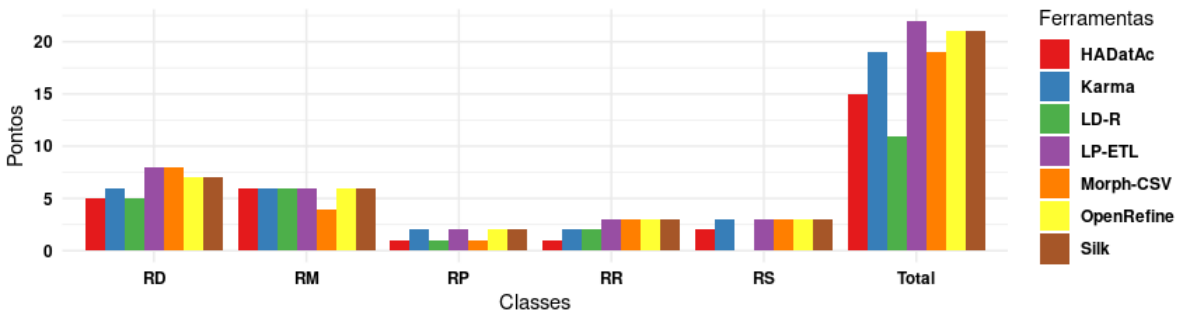
Figura 24 – Percentual de recursos implementados por classe



Fonte: Elaborado pelo autor

A Figura 25 explicita as diferenças das ferramentas por classe dos recursos. O *LP-ETL* é a ferramenta que possui cobertura total do arcabouço comparativo, ao passo que o *LD-R* possui quase metade dos recursos. O *named graph* (RD.VI) cria contexto às triplas com um nome. Esse recurso está presente somente no *LP-ETL* e no *Morph-CSV*. Criar um URI (RD.VII.2) personalizado possui a vantagem de se adaptar aos requisitos do projeto, no entanto, o *HADatAc*

Figura 25 – Resultado das ferramentas por classe dos recursos



Fonte: Elaborado pelo autor

e o *Karma* não possuem o recurso. O *Morph-CSV* é a única ferramenta que não fornece uma interface (RM.II) para auxiliar o mapeamento. Todo o mapeamento é feito em um arquivo texto, desse modo a validação do mapeamento ocorre pela tentativa e erro. Esse fato reforça a importância de uma interface para auxiliar o usuário no mapeamento. O processamento em lote (RP.I) é necessário para integração semântica que possa levar horas. As ferramentas *HADatAc*; *LD-R*; *Morph-CSV* não possuem esse recurso. Gerar o GCOP ontológico não é possível ao utilizar *LD-R*, pois seu mapeamento é *direto aumentado* (RM.I.2). Além disso, a ferramenta não possui nenhuma serialização.

Diante dos resultados apresentados, o objetivo principal do trabalho foi alcançado. O arcabouço comparativo proposto neste trabalho consegue salientar as diferenças das ferramentas para integração semântica de dados tabulares em CSV. A reta numérica positiva contribui para identificar quais ferramentas possuem mais recursos ou recursos mais avançados.

6 CONSIDERAÇÕES FINAIS

As pesquisas científicas beneficiam do conhecimento gerado com os dados semânticos. Os dados semânticos são superiores aos restantes tipos de dados, pois vão além de meramente armazenar informações ao permitir representar coisas do mundo com fidelidade e inferir conhecimento. A ciência da informação apresenta contribuições robustas para integração semântica de dados com os métodos e instrumentos.

Os projetos de integração semântica de dados necessitam utilizar ferramentas para automatizar o processo. Comparar ferramentas é um problema recorrente nos trabalhos na literatura científica. O arcabouço comparativo é a solução objetiva para comparar ferramentas. No entanto, não foi encontrado na literatura um arcabouço comparativo para integração semântica de dados tabulares em CSV.

O objetivo principal é propor um arcabouço comparativo de ferramentas para a integração semântica de dados tabulares em CSV utilizando uma reta numérica positiva para avaliação dos recursos. Os objetivos específicos são: (i) construir um grafo de conhecimento a partir de dados tabulares como linha de base para integração semântica; (ii) determinar, com base na literatura científica, quais ferramentas possuem características para integração semântica de dados tabulares; (iii) mapear insumos na literatura para construir o arcabouço comparativo para avaliação de ferramentas para integração semântica de dados; (iv) comparar as ferramentas que realizam integração semântica de dados.

O percurso metodológico foi dividido em 4 etapas. A primeira etapa foi a elevação semântica dos dados tabulares em um arquivo CSV. O resultado com base na literatura científica foi 3 possíveis métodos para se obter um grafo de conhecimento que são: mapeamento direto, mapeamento direto aumentado e mapeamento semântico. Tanto o arquivo CSV quanto os possíveis grafos de conhecimento são utilizados como linha de base na avaliação das ferramentas. A segunda etapa foi a busca e seleção de ferramentas para integração semântica nos repositórios científicos. Foram estabelecidos os seguintes critérios para seleção das ferramentas: : (i) processar arquivo CSV nativamente, ou seja, sem utilização de software de terceiros; (ii) possuir interface gráfica para o usuário final; (iii) integrar os dados semanticamente utilizando LD; (iv) ter o código fonte aberto (open source). Para cada ferramenta foram realizados testes preliminares para aferir as características. O resultado foram 7 ferramentas selecionadas e 3 ferramentas eliminadas. A terceira etapa é discussão e escolha dos recursos para a formação do arcabouço comparativo. Nessa etapa os recursos foram selecionados com base nos trabalhos do Capítulo 3. Esses recursos são aplicáveis a integração semântica de dados e possuem funcionalidades correspondentes para manipulação de arquivos CSV. Na quarta etapa é realizada a avaliação das ferramentas. As ferramentas são instaladas e a integração semântica de dados é realizada em cada uma delas. O

arcabouço comparativo é construído com base no desempenho de cada ferramenta. A questão de pesquisa "quais os recursos são essenciais para a avaliação de ferramentas para integração semântica de dados tabulares em CSV em um arcabouço comparativo com resultados que possam ser medidos numericamente?" foi respondido com os resultados.

Com essa pesquisa foi possível identificar que o armazém de dados é a principal abordagem adotada pelas ferramentas para realizar a integração semântica de dados. O *LD-R* foi a única ferramenta a adotar o mapeamento direto aumentado, além de não possuir nenhum recurso para serialização. Já o *LP-ETL* possui todos os recursos do arcabouço comparativo, ao passo que *OpenRefine* e *Silk* ambas estão em segundo lugar por não possuírem recursos para *named graphs*. A reta numérica positiva do arcabouço comparativo auxilia na sumarização dos resultados, ademais contribui para gerar gráficos e outras informações analíticas.

Portanto, a questão de pesquisa foi respondida e o objetivo principal foi alcançado. O arcabouço comparativo ao passar pelo método científico demonstrou ser efetivo para destacar as ferramentas para integração semântica de dados tabulares em CSV com mais recursos.

6.1 Limitações e Trabalhos Futuros

Apesar dos resultados do arcabouço serem satisfatórios, o trabalho possui algumas limitações. As ferramentas apresentaram telas e métodos de mapeamento diversificados. As abordagens adotadas pelas ferramentas na intermediação do dado tabular para o LD apresentaram prós e contras. Neste trabalho não foi avaliado a dificuldade e o tempo demandado para realizar cada mapeamento. O arquivo CSV utilizado como caso de uso possui algumas dezenas de milhares de registros. Esse volume de dados não foi suficiente para perceber diferença nítida de performance na integração semântica. A aferição de desempenho deveria utilizar arquivos com diversos volumes de dados. A escala de volume permitiria perceber em qual ponto as ferramentas não seriam eficientes.

Para os trabalhos futuros, aplicar o arcabouço comparativo apresentado nesse trabalho nas ferramentas para integração semanticamente banco de dados relacionais. Assim, validar se o arcabouço comparativo é efetivo para outra fonte de dados. Propor classes de recursos para o contexto da fonte de dados, por exemplo, banco de dados relacional. Essas classes auxiliariam a avaliação levando em conta as especificidades de cada fonte. Uma pesquisa em um grupo focal para aferir quão intuitivo são as ferramentas. Realizar teste de performance em cada ferramenta com diversos volumes de dados em um contexto real de utilização.

REFERÊNCIAS

ADELFIO, M. D.; SAMET, H. Schema extraction for tabular data on the web. **Proc. VLDB Endow.**, VLDB Endowment, v. 6, n. 6, p. 421–432, abr. 2013. ISSN 2150-8097. Disponível em: https://doi.org/10.14778/2536336.2536343.

ALMEIDA, M. B.; BAX, M. P. Taxonomia para projetos de integração de fontes de dados baseados em ontologias. *In*: **V Enancib**. [*S.l.: s.n.*], 2003. p. 1–20.

ALMEIDA, M. B.; BAX, M. P. Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. **Ciência da Informação**, scielo, v. 32, p. 7 – 20, 12 2003. ISSN 0100-1965. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652003000300002&nrm=iso.

ANTONIOU, G.; HARMELEN, F. V. Web ontology language: Owl. *In*: **Handbook on ontologies**. [*s.l.*]: Springer, 2004. p. 67–92.

ATTARD, J. *et al.* A systematic review of open government data initiatives. **Government Information Quarterly**, v. 32, n. 4, p. 399 – 418, 2015. ISSN 0740-624X. Disponível em: http://www.sciencedirect.com/science/article/pii/S0740624X1500091X.

AUER, S. *et al.* Towards a knowledge graph for science. *In*: **Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics**. New York, NY, USA: Association for Computing Machinery, 2018. (WIMS '18). ISBN 9781450354899. Disponível em: https://doi.org/10.1145/3227609.3227689.

BAGUI, S.; BOURESSA, J. Mapping rdf and rdf-schema to the entity relationship model. **Journal of Emerging Trends in Computing and Information Sciences**, Citeseer, v. 5, n. 12, p. 953–961, 2014.

BAX, M. P. Introdução às linguagens de marcas. **Ciência da Informação**, scielo, v. 30, p. 32 – 38, 04 2001. ISSN 0100-1965. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652001000100005&nrm=iso.

BERNABÉ-DÍAZ, J. A. *et al.* Efficient, semantics-rich transformation and integration of large datasets. **Expert Systems with Applications**, Elsevier, v. 133, p. 198–214, 2019.

BERNERS-LEE, T. Semantic web road map. **Design Issues for the World Wide Web**, W3C, v. 2008, n. September 1998, p. 1–10, 1998. Disponível em: http://www.w3.org/DesignIssues/Semantic.html.

BERNERS-LEE, T. **Uniform Resource Identifier (URI): Generic Syntax**. 2019. Disponível em: https://tools.ietf.org/html/rfc3986. Acesso em: 31 jul. 2019.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. **Scientific american**, JSTOR, v. 284, n. 5, p. 34–43, 2001.

Biffl, S. *et al.* Efficient monitoring of multi-disciplinary engineering constraints with semantic data integration in the multi-model dashboard process. *In*: **Proceedings of the 2014 IEEE Emerging Technology and Factory Automation (ETFA)**. [*S.l.: s.n.*], 2014. p. 1–10.

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

BIZER, C. *et al.* Linked data on the web (ldow2008). *In*: **Proceedings of the 17th International Conference on World Wide Web**. New York, NY, USA: Association for Computing Machinery, 2008. (WWW '08), p. 1265–1266. ISBN 9781605580852. Disponível em: https://doi.org/10.1145/1367497.1367760.

BORGIDA, A. *et al.* Classic: A structural data model for objects. *In*: ACM. **ACM Sigmod record**. [*s.l.*], 1989. v. 18, n. 2, p. 58–67.

BRAY, T. *et al.* **Extensible Markup Language (XML) 1.0 (Fifth Edition)**. 2008. W3C Recommendation. Disponível em: http://www.w3.org/TR/REC-xml/. Acesso em: 10 ago. 2020.

BRENNAN, R.; FEENEY, K. C.; GAVIN, O. Publishing social sciences datasets as linked data: a political violence case study. *In*: **Exploration, Navigation and Retrieval of Information in Cultural Heritage workshop (ENRICH 2013), Dublin, Ireland**. [*S.l.: s.n.*], 2013.

CALVANESE, D. *et al.* Ontop: Answering sparql queries over relational databases. **Semantic Web**, IOS Press, v. 8, n. 3, p. 471–487, 2017.

CARROLL, J. J. *et al.* Named graphs, provenance and trust. *In*: **Proceedings of the 14th International Conference on World Wide Web**. New York, NY, USA: Association for Computing Machinery, 2005. (WWW '05), p. 613–622. ISBN 1595930469. Disponível em: https://doi.org/10.1145/1060745.1060835.

CHAVES-FRAGA, D. *et al.* Enhancing obda query translation over tabular data with morph-csv. **arXiv.org**, Cornell University Library, arXiv.org, Ithaca, 2020. ISSN 2331-8422. Disponível em: http://search.proquest.com/docview/2346217296/.

CHEN, L. *et al.* Blank nodes in rdf. **JSW**, v. 7, n. 9, p. 1993–1999, 2012.

DAVIES, S. *et al.* User interface design considerations for linked data authoring environments. *In*: BIZER, C. *et al.* (Ed.). **Proceedings of the WWW2010 Workshop on Linked Data on the Web, LDOW 2010, Raleigh, USA, April 27, 2010**. CEUR-WS.org, 2010. (CEUR Workshop Proceedings, v. 628). Disponível em: http://ceur-ws.org/Vol-628/ldow2010_paper17.pdf.

DIMOU, A. *et al.* What factors influence the design of a linked data generation algorithm? *In*: BERNERS-LEE, T. *et al.* (Ed.). **Proceedings of the 11th Workshop on Linked Data on the Web**. [*s.n.*], 2018. Disponível em: http://events.linkeddata.org/ldow2018/papers/LDOW2018_paper_12.pdf.

DING, L. *et al.* Reflections on provenance ontology encodings. *In*: SPRINGER. **International Provenance and Annotation Workshop**. [*s.l.*], 2010. p. 198–205.

DOAN, A.; HALEVY, A.; IVES, Z. **Principles of data integration**. [*s.l.*]: Elsevier, 2012.

EHRLINGER, L.; WÖSS, W. Towards a definition of knowledge graphs. **SEMANTiCS (Posters, Demos, SuCCESS)**, v. 48, p. 1–4, 2016.

EMMONS, I. *et al.* Rdf literal data types in practice. *In*: **The 7th International Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS 2011)**. [*S.l.: s.n.*], 2011. v. 1.

FLORIAN, B.; MARTIN, K. **Linked Open Data: The Essentials - A Quick Start Guide for Decision Makers**. [*s.l.*]: edition mono/monochrom, Vienna, Austria, 2012.

GALVÃO, M. C. B.; RICARTE, I. L. M. O prontuário eletrônico do paciente no século xxi: contribuições necessárias da ciência da informação. **InCID: Revista de Ciência da Informação e Documentação**, v. 2, n. 2, p. 77–100, 2011.

GANGEMI, A. *et al.* Framester: A wide coverage linguistic linked data hub. *In*: BLOMQVIST, E. *et al.* (Ed.). **Knowledge Engineering and Knowledge Management**. Cham: Springer International Publishing, 2016. p. 239–254. ISBN 978-3-319-49004-5.

GARDNER, S. P. Ontologies and semantic data integration. **Drug Discovery Today**, v. 10, n. 14, p. 1001 – 1007, 2005. ISSN 1359-6446. Disponível em: http://www.sciencedirect.com/science/article/pii/S135964460503504X.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2008.

GRUBER, T. R. A translation approach to portable ontology specifications. **Knowledge acquisition**, Academic Press Ltd., GBR, v. 5, n. 2, p. 199–220, jun. 1993. ISSN 1042-8143. Disponível em: https://doi.org/10.1006/knac.1993.1008.

GUARINO, N.; OBERLE, D.; STAAB, S. What is an ontology? *In*: ____. **Handbook on Ontologies**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. p. 1–17. ISBN 978-3-540-92673-3. Disponível em: https://doi.org/10.1007/978-3-540-92673-3_0.

HAMMER, M.; MCLEOD, D. The semantic data model: a modelling mechanism for data base applications. *In*: ACM. **Proceedings of the 1978 ACM SIGMOD international conference on management of data**. [*s.l.*], 1978. p. 26–36.

HARTIG, O. Provenance information in the web of data. **LDOW**, v. 538, 2009.

HERT, M.; REIF, G.; GALL, H. C. A comparison of rdb-to-rdf mapping languages. *In*: ACM. **Proceedings of the 7th International Conference on Semantic Systems**. [*s.l.*], 2011. p. 25–32.

HEYVAERT, P. *et al.* RMLEditor: a graph-based mapping editor for Linked Data mappings. *In*: SACK, H. *et al.* (Ed.). **The Semantic Web: Latest Advances and New Domains (ESWC 2016)**. Springer, 2016. (Lecture Notes in Computer Science, v. 9678), p. 709–723. ISBN 978-3-319-34129-3. Disponível em: http://dx.doi.org/10.1007/978-3-319-34129-3_43.

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

IZZA, S. Integration of industrial information systems: from syntactic to semantic integration approaches. **Enterprise Information Systems**, Taylor & Francis, v. 3, n. 1, p. 1–57, 2009. Disponível em: https://doi.org/10.1080/17517570802521163.

IZZA, S.; VINCENT, L.; BURLAT, P. Exploiting semantic web services in achieving flexible application integration in the microelectronics field. **Computers in Industry**, v. 59, n. 7, p. 722 – 740, 2008. ISSN 0166-3615. Enterprise Integration and Interoperability in Manufacturing Systems. Disponível em: http://www.sciencedirect.com/science/article/pii/S0166361508000468.

JUNIOR, A. C. *et al.* An evaluation of uplift mapping languages. **International Journal of Web Information Systems**, Emerald Publishing Limited, v. 13, n. 4, p. 405–424, 2017.

JUNIOR, A. C.; DEBRUYNE, C.; O'SULLIVAN, D. Juma uplift: using a block metaphor for representing uplift mappings. *In*: IEEE. **2018 IEEE 12th International Conference on Semantic Computing (ICSC)**. [*s.l.*], 2018. p. 211–218.

KEPEKLIAN, G.; BIHANIC, L.; TRONCY, R. Datalift: A platform for integrating big and linked data. *In*: **Proceedings of the International Conference on Big Data from Space, Frascati, Italy**. [*S.l.: s.n.*], 2014. p. 370–373.

KHALILI, A. Linked data reactor: a framework for building reactive linked data applications. *In*: **LIME/SemDev@ ESWC**. [*S.l.: s.n.*], 2016.

KLÍMEK, J.; ŠKODA, P.; NEČASKÝ, M. Linkedpipes etl: Evolved linked data preparation. *In*: SACK, H. *et al.* (Ed.). **The Semantic Web**. Cham: Springer International Publishing, 2016. p. 95–100. ISBN 978-3-319-47602-5.

KNOBLOCK, C. A. *et al.* Interactively mapping data sources into the semantic web. *In*: CEUR-WS. ORG. **Proceedings of the First International Conference on Linked Science-Volume 783**. [*s.l.*], 2011. p. 13–24.

KUSUMASARI, T. F. *et al.* Data profiling for data quality improvement with openrefine. *In*: IEEE. **2016 International Conference on Information Technology Systems and Innovation (ICITSI)**. [*s.l.*], 2016. p. 1–6.

LAKATOS, E. M.; MARCONI, M. de A. **Fundamentos de metodologia científica**. 6. ed. 5. reimp. São Paulo: Atlas, 2007. ISBN 9788522440153.

LASSILA, O.; SWICK, R. R. **Resource Description Framework (RDF) Model and Syntax Specification**. [*s.l.*], 1999. Disponível em: http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

LI, X. *et al.* Personal knowledge graph population from user utterances in conversational understanding. *In*: IEEE. **2014 IEEE Spoken Language Technology Workshop (SLT)**. [*s.l.*], 2014. p. 224–229.

LOPEZ, V. *et al.* Data access linking and integration with dali: Building a safety net for an ocean of city data. *In*: ARENAS, M. *et al.* (Ed.). **The Semantic Web - ISWC 2015**. Cham: Springer International Publishing, 2015. p. 186–202. ISBN 978-3-319-25010-6.

LUDÄSCHER, B. *et al.* Managing scientific data: From data integration to scientific workflows. **Geoinformatics: Data to knowledge**, Geological Society of America, v. 397, p. 109, 2006.

MCBRIDE, B. The resource description framework (rdf) and its vocabulary description language rdfs. *In*: **Handbook on ontologies**. [*s.l.*]: Springer, 2004. p. 51–65.

MCCUSKER, J. P.; LUCIANO, J. S.; MCGUINNESS, D. L. Towards an ontology for conceptual modeling. *In*: **ICBO**. [*S.l.: s.n.*], 2011.

MICHEL, F.; MONTAGNAT, J.; ZUCKER, C. F. **A survey of RDB to RDF translation approaches and tools**. [*s.l.*], 2014. ISRN I3S/RR 2013-04-FR 24 pages. Disponível em: https://hal.archives-ouvertes.fr/hal-00903568.

MILOSLAVSKAYA, N.; TOLSTOY, A. Big data, fast data and data lake concepts. **Procedia Computer Science**, Elsevier, v. 88, n. 300-305, p. 63, 2016.

PAULHEIM, H. Knowledge graph refinement: A survey of approaches and evaluation methods. **Semantic web**, IOS Press, v. 8, n. 3, p. 489–508, 2017.

PINHEIRO, P. *et al.* Hadatac: A framework for scientific data integration using ontologies. *In*: **International Semantic Web Conference (P&D/Industry/BlueSky)**. [*S.l.: s.n.*], 2018.

POWERS, S. **Practical RDF: solving problems with the resource description framework**. [*s.l.*]: "O'Reilly Media, Inc.", 2003.

RASHID, S. M. *et al.* The semantic data dictionary–an approach for describing and annotating data. **Data Intelligence**, MIT Press, p. 443–486, 2020.

SCHULTZ, A. *et al.* Ldif-linked data integration framework. *In*: CEUR-WS. ORG. **Proceedings of the Second International Conference on Consuming Linked Data-Volume 782**. [*s.l.*], 2011. p. 125–130.

SEGARAN, T.; EVANS, C.; TAYLOR, J. **Programming the Semantic Web: Build Flexible Applications with Graph Data**. [*s.l.*]: "O'Reilly Media, Inc.", 2009.

SEQUEDA, J.; PRIYATNA, F.; VILLAZóN-TERRAZAS, B. Relational database to rdf mapping patterns. *In*: **Proceedings of the 3rd International Conference on Ontology Patterns - Volume 929**. Aachen, DEU: CEUR-WS.org, 2012. (WOP'12), p. 97–108.

SHAFRANOVICH, Y. **Common Format and MIME Type for Comma-Separated Values (CSV) Files**. 2019. Disponível em: https://tools.ietf.org/html/rfc4180. Acesso em: 15 jul. 2019.

STEIN, B.; MORRISON, A. The enterprise data lake: Better integration and deeper analytics. **PwC Technology Forecast: Rethinking integration**, v. 1, n. 1-9, p. 18, 2014.

STONE, D. *et al.* **User Interface Design and Evaluation**. Elsevier Science, 2005. (Interactive Technologies). ISBN 9780080520322. Disponível em: https://books.google.com.br/books?id=VvSoyqPBPbMC.

TARAPANOFF, K.; JÚNIOR, R. H. d. A.; CORMIER, P. M. J. Sociedade da informação e inteligência em unidades de informação. **Ciência da informação**, SciELO Brasil, v. 29, n. 3, p. 91–100, 2000.

THOMAS, H.; O'SULLIVAN, D.; BRENNAN, R. Ontology mapping representations: A pragmatic evaluation. *In*: **SEKE**. [*S.l.: s.n.*], 2009. p. 228–232.

USCHOLD, M.; KING, M. Towards a methodology for building ontologies. *In*: **In Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with IJCAI-95**. [*S.l.: s.n.*], 1995.

VOLZ, J. *et al.* Silk-a link discovery framework for the web of data. **LDOW**, Citeseer, v. 538, p. 53, 2009.

W3C Recommendation. **OWL 2 Web Ontology Language Document Overview (Second Edition)**. 2019. Disponível em: https://www.w3.org/TR/owl2-overview/. Acesso em: 31 jul. 2019.

W3C Recommendation. **OWL web ontology language overview**. 2019. Disponível em: http://www.w3.org/TR/2004/REC-owl-ref-20040210. Acesso em: 31 jul. 2019.

W3C Recommendation. **OWL web ontology language overview**. 2019. Disponível em: https://www.w3.org/TR/2004/REC-owl-features-20040210. Acesso em: 31 jul. 2019.

W3C Recommendation. **RDF 1.1 N-Triples - A line-based syntax for an RDF graph**. 2020. Disponível em: https://www.w3.org/TR/n-triples/. Acesso em: 31 jul. 2020.

W3C Recommendation. **RDF 1.1 Turtle - Terse RDF Triple Language**. 2020. Disponível em: https://www.w3.org/TR/turtle. Acesso em: 31 jul. 2020.

Wikipedia contributors. **Semantic Web Stack — Wikipedia, The Free Encyclopedia**. 2019. Disponível em: https://en.wikipedia.org/w/index.php?title=Semantic_Web_Stack&oldid=891678688. Acesso em: 20 abr. 2019.

WU, J. *et al.* Personalizing actions in context for risk management using semantic web technologies. *In*: SPRINGER. **International Semantic Web Conference**. [*s.l.*], 2017. p. 367–383.

XU, Y. *et al.* Component-based mediation services for the integration of medical applications. **Artif. Intell. Med.**, Elsevier Science Publishers Ltd., GBR, v. 27, n. 3, p. 283–304, mar. 2003. ISSN 0933-3657. Disponível em: https://doi.org/10.1016/S0933-3657(03)00007-1.