Ligações Semânticas Utilizando Predicados SKOS

Ricardo Ávila[1], Salomão Santos[1], David Araújo[1], Vânia Vidal[1], José Macêdo[1]

[1]Universidade Federal do Ceará (UFC) – Fortaleza – Brasil

{ricardoavila,salomaosantos,araujodavid,vvidal,jose.macedo}@lia.ufc.br

Abstract. *Glossaries play a central role in creating better semantic relationship between data sources and facilitating understanding within a domain such as oil and gas Exploration and Production (E&P). This work proposes a type of semantic enrichment between glossaries, using SKOS vocabulary and SPARQL queries. The results were satisfactory, generating predicates skos:exactMatch, skos:broader, skos:narrower and skos:related, which suggests that automatic methods can be used to improve a quality of terminological mappings.*

Resumo. *Glossários desempenham um papel central para a melhoria das relações semânticas entre fontes de dados heterogêneas e para uma melhor compreensão das definições no domínio da Extração e Produção de Petróleo (E&P). Este trabalho propõe o enriquecimento semântico entre glossários utilizando o vocabulário SKOS e as consultas SPARQL. Os resultados foram satisfatórios, gerando predicados skos:exactMatch, skos:broader, skos:narrower e skos:related, o que sugere que os métodos automáticos podem ser usados para melhorar a qualidade dos mapeamentos terminológicos.*

1. Introdução

Serviços interativos como o *Linked Data Mashups* (LDM) são ofertados na *Web* combinando o conteúdo de diferentes fontes de dados em um novo serviço [Rahm et al. 2007]. Desenvolver essa conciliação por meio da construção de um Esquema Global, responsável por representar semanticamente duas ou mais fontes de dados que se deseja integrar é uma tarefa complexa que implica utilizar uma grande quantidade de código específico para acessar cada fonte de dados ou serviço, acarretando alto custo de desenvolvimento e manutenção.

A iniciativa *Linked Data* (LD) apresenta-se como uma solução atraente para o trato dessa problemática, provendo mecanismos para a publicação, a recuperação e a integração de dados distribuídos na *Web* de Dados [Bizer et al. 2007]. A publicação e a conexão de dados de forma estruturada na *Web* permitem a interligação de recursos entre diferentes fontes de dados, possibilitando, assim, a conexão dessas fontes em um espaço global único, possibilitando a inferência e um melhor uso dos dados [Lopes et al. 2016].

A combinação dessas técnicas com tecnologias semânticas modernas, com a utilização de abordagens de modelagem flexíveis e extensíveis, pode permitir a integração de dados anteriormente distintos, provendo facilidades de pesquisa e produzindo novos recursos e serviços de dados neutros em plataforma para a colaboração e a disseminação do conhecimento.

Existem diversas arquiteturas de aplicações de LDM que dependem amplamente de seu uso [Heath and Bizer 2011]. Quanto à integração de dados utilizando LDM, exis-

tem duas abordagens: a abordagem materializada, em que os dados são coletados, persistidos e consultados em uma base de dados centralizada, e a abordagem virtual que possibilita consultas em um conjunto de fontes de dados fixos. Nossa abordagem visa a efetuar a fusão dos glossários, criando uma base de conhecimento materializada no domínio do petróleo.

O *mashup*[1] de glossários no domínio do petróleo visa à integração de vários glossários de termos relacionados à Extração e à Produção de Petróleo (E&P). Essa integração permitirá uma melhor compreensão das relações entre as definições no domínio de produção e a extração de petróleo. As principais contribuições deste trabalho são: *(i)* definir heurísticas de geração de *links* semânticos e *(ii)* facilitar a integração semântica de ontologias no domínio do petróleo.

Para o aprofundamento desta pesquisa, utilizamos *Simple Knowledge Organization System* (SKOS) [Miles et al. 2005] e outros vocabulários para a construção da Ontologia de Domínio. SKOS permite o mapeamento das relações semânticas entre termos, utilizando métodos de raciocínio semânticos para sugerir ligações de uma terminologia para outra, por exemplo, SKOS possibilita a utilização de regras lógicas, como *subsumption* e *transitivity*, para inferir ligações não assertivas com base em ligações assertivas e relações hierárquicas entre termos dentro das terminologias.

Normalmente, se A possui *skos:exactMatch* B, e C *is-a* A, então se pode inferir que C possui *skos:broader* com B. Como SKOS necessita do esforço de especialistas e de severas atividades manuais, propomos o uso de consultas SPARQL [Prud'hommeaux and Seaborne 2008] baseado em regras simples, que identificará automaticamente os mapeamentos entre termos utilizando os predicados SKOS. Na linguagem de consulta SPARQL, o comando CONSTRUCT permite que, a partir de um resultado de pesquisa, seja construído um conjunto de triplas. Desse modo, os novos *links* semânticos serão materializados conforme as regras propostas, com a geração semiautomática de predicados *skos:exactMatch*, *skos:narrower*, *skos:broader* e *skos:related*.

Este artigo está organizado da seguinte forma: Na Seção 2, apresentamos o processo de geração de *Linked Data Mashup*. Na Seção 3, detalhamos o ambiente desenvolvido para a execução de *Linked Data Mashup*. Na Seção 4, denotamos a metodologia proposta. Na seção 5, exibimos os experimentos e os resultados obtidos. Na Seção 6, discutimos os trabalhos relacionados. Por fim, na Seção 7, são delineadas as conclusões e os trabalhos futuros.

2. Processo de Geração de LDM

O processo de criação de um LDM incide inicialmente na modelagem de uma ontologia de domínio e integração semântica [Tran et al. 2014]. A modelagem da ontologia de domínio e a integração semântica dispõe das seguintes etapas: *(i)* Elaboração do modelo conceitual da aplicação em uma Ontologia de Domínio (OD); *(ii)* Descrição de cada fonte de dados por meio de sua respectiva Ontologia Fonte (OF), descrevendo os dados que serão exportados; *(iii)* Mapeamento das correspondências entre a OD e as OFs e *(iv)* os *Links* no formato padrão *Resource Description Framework* (RDF)[2] descobertos entre as Fontes de Dados (FD) distintas determinam as ligações entre as OFs.

[1] Aplicação web que usa conteúdo de mais de uma fonte para criar um novo serviço completo.
[2] https://www.w3.org/RDF/

Tabela 1. Glossários de Termos no Domínio do Petróleo

Glossário	*Link*
Schlumberger Oilfield Glossary	http://www.glossary.oilfield.slb.com
Glossário da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP)	http://www.anp.gov.br/?id=582
Bureau of Safety and Environmental Enforcement (BSEE)	http://www.bsee.gov/Glossary-of-Terms
PetroWiki - SPE's E&P Glossary	http://petrowiki.org/Category:Glossary
Wikipédia	https://wikipedia.org

Para validar os conceitos propostos, selecionamos os glossários apresentados na Tabela 1 com base em *(i)* busca de repositórios de termos relacionados à Extração e à Produção de Petróleo (E&P), preocupando-nos em selecionar organizações que atuam nesse domínio e *(ii)* fontes de dados disponibilizados em *sites* de acesso ao público ou sem nenhuma restrição de direitos autorais para o uso dos termos para fins de pesquisa. Optamos por partir dessas fontes por considerarmos esses glossários produzidos criteriosamente, com orientação de profissionais que atuam e os criaram para o uso constante e a correta apresentação das informações.

Após a seleção dos glossários, utilizamos a OD e as OFs do *mashup* de glossários, representados nas Figuras 1 e 2, respectivamente. O passo a passo para a captura dos termos nas FDs, a transformação em RDF e o armazenamento são apresentados na Seção 3.

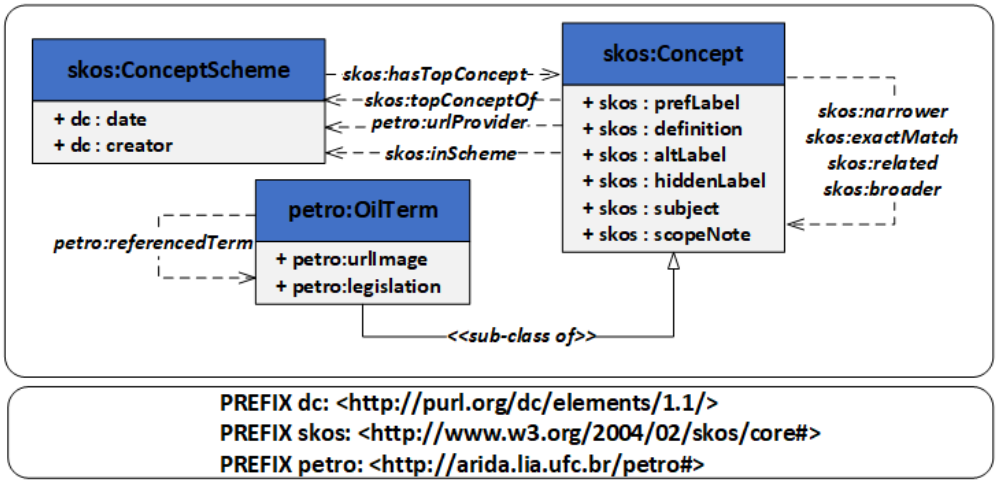

Figura 1. Ontologia de Domínio (OD) do *Mashup* de Glossários

3. Ambiente de Execução de LDM

A partir da seleção dos glossários e da implantação do *mashup*, necessitamos de alguns passos para a viabilização do ambiente. Destarte, utilizamos uma arquitetura visando a unificar as informações das diversas fontes de dados sobre os termos da área de Extração e Produção de Petróleo (E&P).

Os componentes dessa arquitetura possuem o objetivo de coletar os termos das diversas fontes de dados, assim como tratar as informações coletadas para adequá-las a

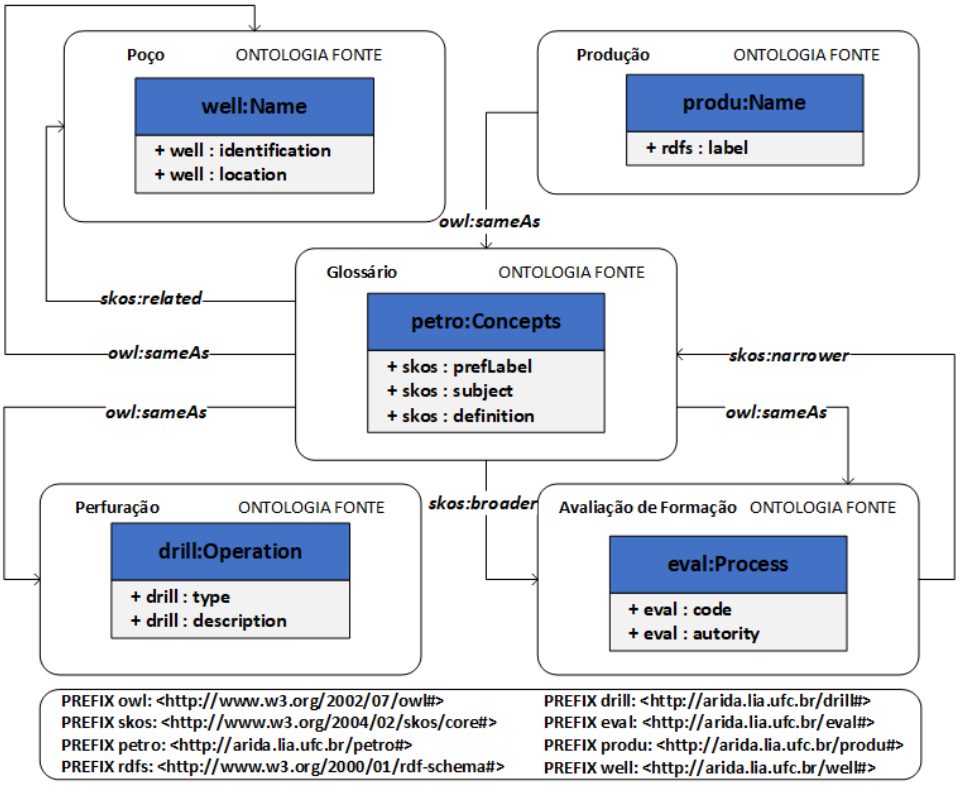

Figura 2. Ontologias Fontes (OFs) do *Mashup* de Glossários

uma mesma estrutura comum, a ontologia de domínio do *mashup* de glossários. As fontes de dados representam os diversos glossários apresentados na Tabela 1.

A partir de cada glossário, um *crawler* é desenvolvido para colher as informações disponíveis. Um *web crawler* é um programa ou *script* automatizado que navega em páginas da *Internet* de maneira metódica e automatizada, permitindo extrair tipos específicos de informação contidos na páginas. Cada *crawler* na arquitetura é responsável, então, por mapear a estrutura dos termos de uma página de glossário para a modelagem da base de documentos; após isso, a coleta é executada, e os termos extraídos são persistidos na base.

O componente base de documentos é responsável por agrupar ao máximo as definições dos diversos glossários para os mesmos termos. Esse processo permite uma inferência inicial entre as estruturas utilizadas pelos diferentes glossários ao definir um mesmo termo. Com isso, o processo de gerar triplas ficou mais rápido e completo. A modelagem definida na base para a persistência das definições foi concebida após a análise da estrutura da informação disponibilizada nas diferentes fontes de glossários, e cada uma delas contém termos relacionados ao domínio e à questão. Portanto, uma união das diferentes estruturas permite a maximização das informações sobre um mesmo termo. Essa base de documentos utiliza o MongoDB[3], e, como a estrutura de armazenamento é baseada em documentos JSON[4], isso facilita a heterogeneidade e a atualização de novos atributos identificados a partir de novos glossários. A modelagem proposta para a estrutura do armazenamento em MongoDB é apresentada na Figura 3.

[3]https://www.mongodb.com
[4]http://www.json.org

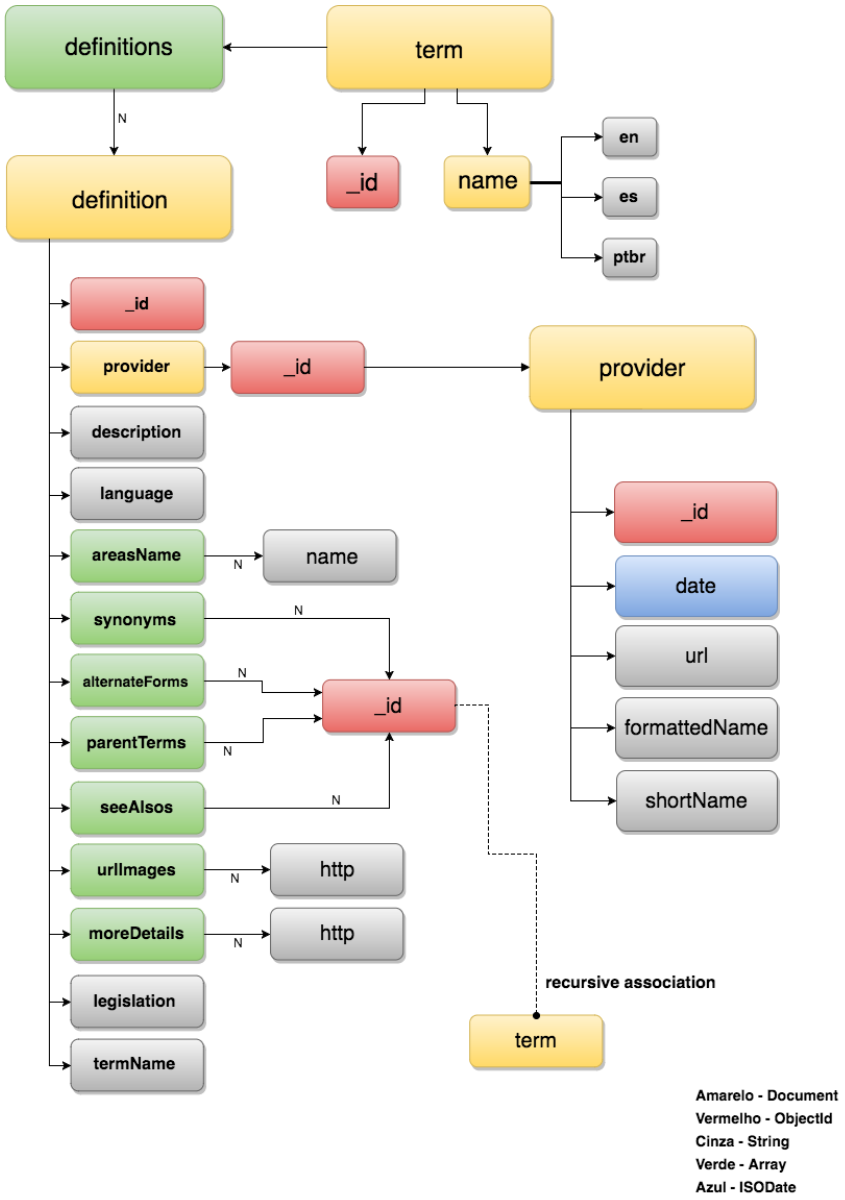

Figura 3. Modelo de Dados Armazenados em MongoDB

O conceito inicial por trás dessa modelagem é agrupar ao máximo as definições dos diversos glossários para os mesmos termos. Esse processo irá futuramente auxiliar o processo de enriquecimento das ontologias. A estrutura utilizada permite a realização de inferências entre definições de diferentes glossários para um mesmo termo. É importante destacar que essa modelagem foi concebida após a análise da estrutura de informação disponibilizadas nos cinco primeiros glossários utilizados, conforme apresentado na Tabela 1. Os diversos atributos apresentados nessa modelagem são a união das diferentes informações extraídas nesses glossários. Destarte, nem todos os termos possuirão todos os atributos preenchidos, dependendo da definição apresentada em cada glossário, sendo algumas mais completas do que outras.

A modelagem proposta facilita a geração das triplas, pois possibilita a estruturação das definições de diferentes glossários em diferentes idiomas para um mesmo termo. Esse agrupamento permite o entendimento dos dados de forma mais clara, visto que não se

encontram mais apenas como textos nas páginas Web dos glossários - que estruturam os dados de forma heterogênea. Com a identificação mais clara dos atributos comuns dos termos, temos um aumento do potencial de geração de *links* semânticos entre eles (i.e.: *skos:exactMatch*, *skos:related*, *skos:broader* e *skos:narrower*).

Os dois últimos componentes são responsáveis pelos processos de triplificação e amazenamento das triplas. O triplificador é um *script* que percorre todos os termos da base de documentos, transformando-os em instâncias triplificadas da ontologia de domínio do *mashup* de glossários. O triplificador foi implementado em *Python* e gera as triplas no padrão N-Triples[5]. Por último, as triplas são armazenadas em uma base RDF ou salvas em arquivos.

4. Metodologia Proposta

Para a geração semiautomática/automática dos predicados SKOS, desenvolvemos consultas SPARQL com o comando CONSTRUCT para validar e avaliar o desempenho do método proposto, utilizando como base o predicado *skos:prefLabel* dos glossários apresentados na Tabela 1. A Figura 4 apresenta o detalhamento visual das consultas desenvolvidas.

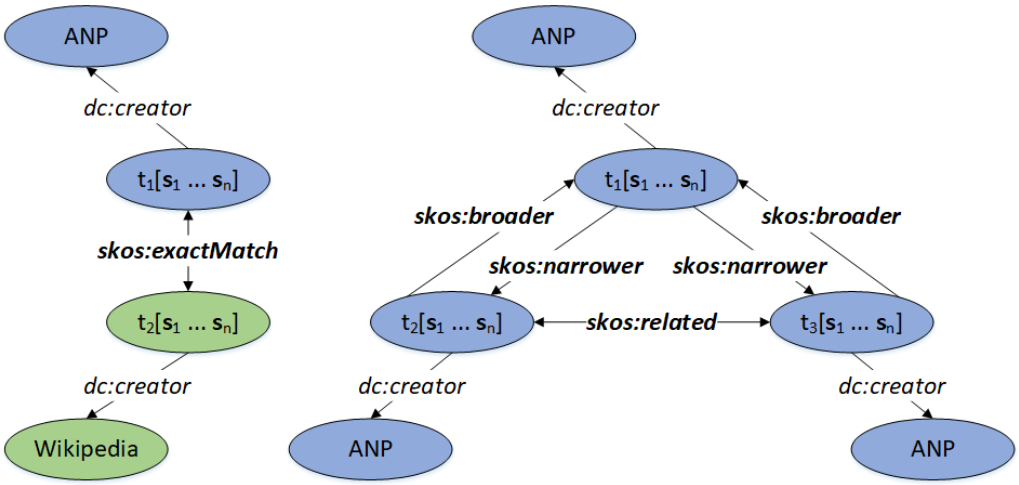

Figura 4. Descoberta de predicados SKOS

As regras das consultas para geração de predicados SKOS seguem os seguintes passos: *(i)* sendo **t**=termo e **s**=*string*, definimos que t_1[**s_1 ... s_n**] e t_2[**s_1 ... s_n**] possui ligação *skos:exactMatch* se t_1 for igual a t_2 e o predicado *dc:creator* de t_1 e t_2 forem diferentes; *(ii)* caso t_1[**s_1**] seja igual a t_2[**s_1** ... s_n] e *dc:creator* de t_1 e t_2 sejam iguais, definimos que t_1 é *skos:narrower* de t_2 e t_2 é *skos:broader* de t_1; *(iii)* finalmente, definimos que haverá ligação reflexiva *skos:related* entre t_2 e t_3 quando t_2[**s_1** ... s_n] e t_3[**s_1** ... s_n] possuírem ligações semânticas *skos:narrower* e *skos:broader* com t_1. A consulta SPARQL para a geração dos links semânticos *skos:narrower* e *skos:broader* é apresentada na Lista 1.

Para o alinhamento do predicado *skos:exactMatch* entre os glossários, utilizamos a ferramenta Silk [Volz et al. 2009]. Durante as experimentações preliminares, os resultados apontaram para o uso do algoritmo Levenshtein que obteve melhor resposta em relação a outros algoritmos como o Jaro-Winkler e o Jaccard, por exemplo. Definimos o

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

limite para a métrica de alinhamento em 0.95, ou seja, dois termos são considerados equivalentes caso a similaridade do conjunto de caracteres que os compõe seja maior ou igual a 0.95. A adequação de diferentes métricas de distância de *strings* tem sido amplamente discutida na literatura [Cohen et al. 2003]. Feito o alinhamento do *skos:exactMatch*, seguimos com os experimentos para os demais predicados.

Esses tipos de relacionamentos são representados de acordo com as seguintes definições: **ET** (*Exact Terms*), definido entre dois termos t_i e t_j, desde que sejam consideradas a mesma palavra. **ET** é simétrico, isto é, t_i **ET** $t_j \Rightarrow t_j$ **ET** t_i. **BT** (*Broader Terms*) definido entre dois termos t_i e t_j, desde que t_i possua um significado mais geral do que t_j. **BT** não é simétrico. O oposto de **BT** é **NT** (*Narrower Terms*): t_i **NT** $t_j \Rightarrow t_j$ **BT** t_i. **RT** (*Related Terms*) definido entre dois termos t_i e t_j, que são geralmente utilizados em conjunto no mesmo contexto. **RT** é simétrico: t_i **RT** $t_j \Rightarrow t_j$ **RT** t_i. A descoberta desses relacionamentos terminológicos, a partir de esquemas de fonte, é uma atividade semiautomática proposta neste trabalho.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

CONSTRUCT

{
   ?s1 <http://www.w3.org/2004/02/skos/core#narrower> ?s2 .
   ?s2 <http://www.w3.org/2004/02/skos/core#broader> ?s1 .
}

WHERE {

  ?s1 skos:prefLabel  ?l1 .
  ?s2 skos:prefLabel  ?l2 .

  FILTER (lang(?l1) = 'en' &&
        !CONTAINS(?l1, "␣") &&
        lang(?l2) = 'en' &&
        !SAMETERM(?l1, ?l2))

  BIND(CONCAT("␣",STR(?l1)) AS ?label_1)

  FILTER(STRENDS(?l2,?label_1))

}
```

Lista 1. Consulta SPARQL para geração de *links skos:narrower* e *skos:broader*

O predicado *skos:related* permite a representação de *links* associativos (não hierárquicos), como a relação entre um tipo de evento e uma categoria de entidades que normalmente participam dele. Outro uso para o *skos:related* é a ligação entre duas categorias, em que uma é mais geral, e a outra é mais específica. O *skos:related* também

pode ser usado para representar *links* parte-inteiro que não são classificados como relacionamentos hierárquicos. A consulta SPARQL para a geração dos *links* semânticos *skos:related* é apresentada na Lista 2.

As regras utilizadas pela consulta SPARQL para gerar predicados SKOS seguem uma consistência formal, com um grau satisfatório de sucesso. Elas foram definidas com o uso de um método iterativo e a inclusão das seguintes etapas: (*i*) definição intuitiva de regras a partir da aplicação de uma amostra de 100 predicados *skos:prefLabel*; (*ii*) formalização das regras em formato algorítmico; (*iii*) teste da consulta SPARQL com CONSTRUCT em uma amostra maior; (*iv*) avaliação dos resultados; e, (*v*) refinamento das regras, se for o caso.

Foram necessárias cinco interações para obtermos a versão atual da consulta SPARQL. A consulta SPARQL com o comando CONSTRUCT gera automaticamente quatro propriedades de mapeamento no SKOS: *skos:exactMatch*, *skos:broader*, *skos:narrower* e *skos:related*.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

CONSTRUCT

{
   ?f1 <http://www.w3.org/2004/02/skos/core#related> ?f2 .
}

WHERE {

  ?p skos:narrower  ?f1 .
  ?p skos:narrower  ?f2 .
  ?p skos:prefLabel ?label1 .
  ?f1 skos:prefLabel ?label2 .
  ?f2 skos:prefLabel ?label3 .

  FILTER (!SAMETERM(?f1, ?f2) &&
        lang(?label1) = 'en' &&
        lang(?label2) = 'en' &&
        lang(?label3) = 'en')

}
```

Lista 2. Consulta SPARQL para geração de *links skos:related*

5. Resultados

Com a utilização da consulta SPARQL proposta, fizemos a coleta de 5373 termos do *Schlumberger Oilfield Glossary*, tendo como base para as comparações o predicado *skos:preflabel*. Os resultados obtidos e a proporção dos predicados SKOS gerados pela consulta estão descritos nas Tabelas 2 e 3.

Tabela 2. Número de predicados SKOS gerados pela consulta SPARQL

Predicado SKOS	exactMatch	related	broader	narrower	Total
Quantidade de *Links* gerados	2054	17487	1512	1512	22565
Amostra para avaliação	103	87	76	76	341
Validação pelo especialista	101	76	70	71	318
95% intervalo de confiança	$98,63^{\pm 4,37}\%$	$85,87^{\pm 1,13}\%$	$72,47^{\pm 3,53}\%$	$72,42^{\pm 3,58}\%$	82,35

Tabela 3. Exemplos de predicados SKOS gerados pela consulta SPARQL

Schlumberger Oilfield	PetroWiki - SPE's E&P	*Link* SKOS gerado	Avaliação
porosity	porosity	exactMatch	✔
	moldic porosity	narrower	✔
	diagenetic porosity	broader	✔
vugular porosity	vug	~~related~~	sameAs
	fracture porosity	related	✔
	wet clay porosity	broader	✔
wet clay porosity	porosity	broader	✔
	electrical double layer	~~related~~	petro:referencedTerm
	isolated porosity	related	✔
	clay	✘	broader
	smectite clay	✘	related

Os resultados foram avaliados com uma amostra de 341 exemplos de predicados SKOS gerados pela consulta SPARQL. Todos os *links* elaborados foram avaliados por um geólogo especialista no domínio da Extração e Produção de Petróleo (E&P) para comprovar a eficácia do modelo proposto. Vale ressaltar que a avaliação ocorreu com a utilização das ligações geradas entre os glossários *Schlumberger Oilfield* e *PetroWiki - SPE's E&P*.

O predicado *skos:exactMatch* obteve 98,63% de acerto, o que consideramos satisfatório, uma vez que utilizamos somente *stemming* e a remoção de *stopwords* como técnicas de pré-processamento textual para melhorar os resultados das comparações entre os termos [Avila and Soares 2012].

Já para os *links* gerados para o predicado *skos:related*, obtivemos 85,87% de acerto, contudo, de acordo com a Tabela 3, outros predicados deveriam ter sido gerados. O *link skos:related* somente será gerado após a confirmação da existência de ligações *skos:broader* e *skos:narrower* para o mesmo termo pai. Em alguns casos, ocorrerão exceções que precisam ser tratadas. No caso do termo **vugular porosity**, deveria ser gerado um *link owl:sameAs* ligando-o à **vug**. Faz-se necessário o refinamento da consulta SPARQL para tratar esses casos que não dependem da similaridade dos termos, mas sim do entendimento prévio do domínio em que estão sendo geradas as ligações semânticas.

Finalmente, para os predicados *skos:broader* e *skos:narrower*, obtivemos, respectivamente, 72,47% e 72,42%. O baixo desempenho desses resultados está intrinsecamente ligado à dependência que a consulta SPARQL tem com a comparação das *strings* que compõem cada termo. Dessa forma, os *links* somente serão gerados se ocorrer o *matching* entre os termos comparados.

Avaliando conjuntamente os predicados SKOS gerados, a consulta SPARQL ob-

teve 82,35% de ligações SKOS geradas corretamente. Levando-se em consideração que o predicado *skos:exactMatch* é a ligação semântica mais importante para o alinhamento terminológico, alcançando 98,63% de acerto nos experimentos, consideramos a consulta SPARQL eficaz para o mapeamento das correspondências entre a OD e as OFs. O desempenho da consulta SPARQL é menos convincente para as outras propriedades do SKOS, com resultados acima de 72%, o que consideramos eficaz.

Para trabalhos futuros, utilizaremos a proporção de cada tipo de ligação semântica gerada como base para o desenvolvimento de uma métrica de similaridade que apresente o nível de cobertura entre diferentes vocabulários controlados, dentro do domínio. Dessa forma, espera-se diminuir a dependência de um especialista na validação dos mapeamentos.

6. Trabalhos Relacionados

[Zapilko et al. 2012] detalha o uso de SKOS e o processo de adaptação de classes e propriedades para criar uma representação semanticamente completa. De acordo com os autores, essa nova modelagem permitiu a descoberta de ligações de equivalência e termos compostos com ambiguidades.

[van Ossenbruggen et al. 2011] apresenta as principais limitações das ferramentas atuais utilizadas para o alinhamento de instâncias/conceitos propondo uma abordagem alternativa em que utilizaram dois casos de uso para o mapeamento de ligações. Demonstra como a plataforma desenvolvida por eles obteve melhores resultados, em que, de acordo com os autores, os ganhos obtidos foram significativos. Entre as técnicas utilizadas destaca-se o uso de aprendizado de máquina.

Na pesquisa de [Ge and Chen 2010], foi desenvolvido um sistema baseado em ontologias para gerenciar dados de exploração de petróleo que abordam as questões de integração de dados e compartilhamento de informações. De acordo com os autores, a abordagem garante a validade dos dados de petróleo que irá apoiar o processo de descoberta de conhecimento de petróleo.

[Kazi and Kurian 2014] propõem uma metodologia de enriquecimento de ontologias, extraindo padrões de bases de conhecimento por meio de novas inferências derivadas do próprio domínio e utilizando algoritmos de aprendizado de máquina como redes neurais ou árvores de decisão. O conteúdo extraído passa por processos de mineração de dados e pela validação de um especialista. A ontologia construída auxilia na construção de um sistema especialista.

O trabalho de [d'Aquin et al. 2012] define um novo método de descoberta de conhecimento, combinando *(i)* as técnicas de mineração de dados para fazer emergir modelos implícitos de dados e *(ii)* o uso de padrões de engenharia de ontologias para capturar esses modelos de forma reutilizável. Os resultados atingidos apontam para a redução de tempo na preparação de dados e na interpretação de resultados, nas atividades de consulta de especialistas e na construção de ontologias e nas ligações semânticas.

[Miranda et al. 2016] apresentam estratégias alternativas para armazenar e acessar ontologias, com o objetivo de apoiar os processos de compartilhamento, o reuso, o planejamento, a avaliação, a personalização e a adaptação de conhecimentos em cenários relacionados à aprendizagem. Os resultados dos experimentos permitiram que os autores definissem uma estrutura capaz de suportar, do ponto de vista metodológico e tecnológico,

o uso de ontologias no contexto de um sistema educacional baseado na Web Semântica.

Os trabalhos aqui citados apresentaram diferentes maneiras de aplicar a ontologia para a descoberta de padrões e a correta categorização de termos e conceitos. Ainda não existe uma forma única para trabalhar com conceitos em contextos heterogêneos. Em cada caso, deve trabalhar-se no sentido de compreender o domínio que está sendo explorado e buscar os modelos mais apropriados para mapear/desenvolver sua respectiva definição, identificando os padrões que melhor se adaptam àquele contexto. Nosso trabalho apresentou resultados satisfatórios para a descoberta de ligações semânticas que podem auxiliar no enriquecimento de ontologias, favorecendo a descoberta de conhecimento.

7. Conclusão e Trabalhos Futuros

Demonstramos neste artigo uma consulta SPARQL com CONSTRUCT, conforme os resultados obtidos, eficaz para a geração semiautomática de predicados *skos:exactMatch*, sugerindo que a metodologia proposta pode ser utilizada para melhorar a qualidade dos mapeamentos terminológicos entre a OD e as OFs. A consulta SPARQL foi testada com a utilização de glossários de termos no domínio do petróleo, que pode, também em princípio, ser utilizada para outras terminologias e fontes de mapeamento.

O desempenho da metodologia proposta foi menos eficaz nos demais predicados *skos:broader*, *skos:narrower* e *skos:related*. Esses resultados ocorreram devido ao fato de as ligações serem geradas por meio de comparações de *strings*. Por exemplo, o termo **wet clay porosity** deve ter um *link skos:broader* com **clay**, porém a consulta SPARQL não mapeou essa ligação semântica. Outro fato importante é a impossibilidade de gerar *links* quando não houver alinhamento terminológico entre as *strings* comparadas. Dentre as possíveis soluções, podemos utilizar o sinônimo dos termos, aprendizado de máquina e/ou mapeamento das classes e subclasses do domínio.

De modo geral, a consulta SPARQL proposta depende principalmente da qualidade das terminologias comparadas. O uso de terminologias com princípios mais sistemáticos e relações hierárquicas mais transparentes pode facilitar os mapeamentos terminológicos e a geração dos *links* semânticos. As linguagens de representação de conhecimento formal, como a Linguagem de Ontologia da Web (OWL), podem ajudar nessa tarefa.

Os resultados apresentados foram para um único glossário, especificamente no domínio de Exploração e Produção de Petróleo (E&P). Daremos continuidade à pesquisa, aplicando o *framework* proposto para a integração de glossários em outros domínios. Entendemos que ainda necessitamos melhorar os resultados, principalmente em relação à geração dos *links* semânticos em casos de sinônimos, hiperônimos e hipônimos. Trataremos da resolução dessas entidades heterogêneas em trabalhos futuros.

Referências

Avila, R. and Soares, J. M. (2012). Concepção de ferramenta de apoio à correção de questões dissertativas com base na adaptação de algoritmos de comparação e busca textual combinados com técnicas de pré-processamento de textos. In *RENOTE Revista Novas Tecnologias na Educação*, volume 10.

Bizer, C., Cyganiak, R., and Gauß, T. (2007). The rdf book mashup: from web apis to a web of data. In *ESWC'07 Workshop on Scripting for the Semantic Web*, volume 1.

Cohen, W. W., Ravikumar, P., and Fienberg, S. E. (2003). A comparison of string distance metrics for name-matching tasks. In *Proceedings of the IJCAI-2003 Workshop on*, pages 73–78, Acapulco, Mexico.

d'Aquin, M., Kronberger, G., and Suárez-Figueroa, M. C. (2012). Combining data mining and ontology engineering to enrich ontologies and linked data. In *KNOW@LOD*, volume 868 of *CEUR Workshop Proceedings*, pages 19–24. CEUR-WS.org.

Ge, J. and Chen, Z. (2010). Constructing ontology-based petroleum exploration database for knowledge discovery. *Trans Tech Publications, Switzerland*, 20-23:975–980.

Heath, T. and Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 1st edition.

Kazi, A. and Kurian, D. (2014). An ontology based approach to data mining. In *International Journal of Engineering Development and Research (IJEDR)*, volume 2.

Lopes, G., Vidal, V., and Oliveira, M., editors (2016). *Construção de Linked Data Mashup para Integração de Dados da Saúde Pública*. SBC - XXXI Simpósio Brasileiro de Banco de Dados.

Miles, A., Matthews, B., Wilson, M., and Brickley, D. (2005). Skos core: Simple knowledge organisation for the web. In *Proceedings of the 2005 International Conference on Dublin Core and Metadata Applications: Vocabularies in Practice*, DCMI '05, pages 1:1–1:9.

Miranda, S., Orciuoli, F., and Sampson, D. G. (2016). A skos-based framework for subject ontologies to improve learning experiences. *Comput. Hum. Behav.*, 61(C):609–621.

Prud'hommeaux, E. and Seaborne, A. (2008). SPARQL Query Language for RDF. W3C Recommendation.

Rahm, A. T. D. A. E., Thor, D., and Aumueller, E. (2007). Data integration support for mashups. In *Workshops at the Twenty-Second AAAI Conference on Artificial Intelligence*.

Tran, T. N., Truong, D. K., Hoang, H. H., and Le, T. M. (2014). *Linked Data Mashups: A Review on Technologies, Applications and Challenges*, pages 253–262. Springer International Publishing, Cham.

van Ossenbruggen, J., Hildebrand, M., and de Boer, V. (2011). Interactive vocabulary alignment. In Gradmann, S., Borri, F., Meghini, C., and Schuldt, H., editors, *TPDL*, volume 6966 of *Lecture Notes in Computer Science*, pages 296–307. Springer.

Volz, J., Bizer, C., Gaedke, M., and Kobilarov, G. (2009). Silk - a link discovery framework for the web of data. In Bizer, C., Heath, T., Berners-Lee, T., and Idehen, K., editors, *LDOW*, volume 538 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Zapilko, B., Schaible, J., Mayr, P., and Mathiak, B. (2012). Thesoz: A skos representation of the thesaurus for the social sciences. *CoRR*, abs/1209.5850.