Framework Baseado em Ontologias para Publicação e Integração Semântica de Glossários

Ricardo Ávila[1], David Araujo[1], Gabriel Lopes[2], Vânia Vidal[1], José Macedo[1]

[1]Departamento de Computação – Universidade Federal do Ceará (UFC)

[2]Instituto Federal do Ceará (IFCE)

{ricardoavila, araujodavid, vvidal, jose.macedo}@lia.ufc.br,
gabriellopes9102@gmail.com

Resumo. *Glossários desempenham um papel central para a melhoria dos serviços de busca semântica e alinhamento de ontologias de fontes de dados heterogêneas. Esse trabalho lida com o problema de interoperabilidade semântica entre glossários e apresenta um framework baseado em ontologias para integração semântica de glossários. O framework proposto é utilizado para construção de um mashup que tem o objetivo de integrar terminologias de diferentes glossários no domínio de petróleo.*

Abstract. *Glossaries play a central role in improving semantic search services and aligning ontologies from heterogeneous data sources. This work deals with the semantic interoperability problem between glossaries and presents an framework based on ontologies for semantic integration of glossaries. The proposed framework is used to construct a mashup that aims to integrate terminologies of different glossaries in the petroleum domain.*

1. Introdução

A integração de informações de fontes heterogêneas é um tema antigo de pesquisa, com trabalhos datando há mais de 30 anos [Batini *et al.*, 1986], o qual apresenta problemáticas que perpetuam ainda hoje. Uma dessas problemáticas ao integrar dados é a conciliação semântica das informações em fontes heterogêneas [Hull, 1997].

Segundo Lenzerini (2002), essa conciliação é realizada por meio da construção de um Esquema Global, responsável por representar semanticamente as duas ou mais fontes de dados que se deseja integrar. Para isso, os dados devem ser analisados, entendidos e, então, mapeados para esse esquema global. Tal tarefa pode tornar-se inviável em bancos de dados relacionais, que ainda representam a maior parcela dos dados disponíveis[1], uma vez que a semântica das informações neles contidas está limitada à nomenclatura de suas tabelas e ao conteúdo destas.

A iniciativa *Linked Data* (LD) apresenta-se como uma solução atraente para o trato dessa problemática, uma vez que os dados deixam de ser representados por tabelas, com pouca ou nenhuma semântica, e passam a ser representadas por triplas RDF[2] descritas por uma ontologia. O *linked data*, então, promove a publicação de dados,

[1] DB-ENGINES. DB-Engines Ranking. Acessado em 17/02/2017. Disponível em: <http://db-engines.com/en/ranking>.
[2] https://www.w3.org/RDF

anteriormente isolados, como grafos RDF interligados. Com os dados mais fáceis de serem descobertos e entendidos e com o uso de padrões W3C[3], como SPARQL[4], o *linked data* está impulsionando uma mudança de paradigma na construção de aplicações de *mashup*[5] (integração), *i.e.* aplicações *web* que utilizam dados de mais de uma fonte de dados para prover um novo serviço. Um *Linked Data Mashup* (LDM) é uma aplicação *web* que oferece novas funcionalidades, combinando, agregando e transformando informações disponíveis em fontes *linked data* na *Web* de dados. Existem diversos exemplos de casos de sucesso de aplicações semânticas construídas com o uso de LDM. Um exemplo famoso é o BBC *Music*[6], que integra dados de duas fontes *linked data*, *DBPedia*[7] e *MusicBrainz*[8].

Apesar da mudança de paradigma em integração de dados que o *linked data* proporciona, a construção de um LDM ainda é uma tarefa desafiadora. Segundo Vidal (2015), existem quatro desafios principais para criação de *mashups* em *linked data*: (*i*) seleção das fontes *linked data* relevantes para a aplicação; (*ii*) extração e tradução de fontes de dados distintas para uma ontologia comum; (*iii*) identificação de *links* que denotam a similaridade entre instâncias em fontes distintas e, finalmente, (*iv*) combinação e fusão de múltiplas representações de um mesmo objeto do mundo real em uma única representação.

Nesse contexto, dados heterogêneos na *web* podem acarretar diversas problemáticas. Por exemplo, para uma determinada empresa ou um grupo de pesquisa desenvolver uma base de conhecimento sobre determinado assunto, devem ser realizadas consultas, possivelmente em glossários na *web*, sobre os diversos conceitos a serem inseridos nessa base de conhecimento. Se a definição de tais conceitos divergirem nos diversos glossários, deve ser realizada uma conciliação semântica em tais glossários. No Brasil, uma área em que essa problemática é bem aparente é a de Exploração e Produção de Petróleo (E&P). Nessa área, as definições dos termos relacionados a petróleo, *e.g.* poço de petróleo e porosidade, divergem nos diversos glossários existentes na *web*. Uma forma de lidar com essa problemática é por meio da criação de *links* semânticos entre os termos, o que não é uma tarefa trivial.

Diante disso, este trabalho apresenta um *framework* baseado em ontologias para construção de *mashup* de glossários. A integração semântica das terminologias de diferentes glossários é essencial para amenizar o problema de interoperabilidade entre diferentes glossários de um mesmo domínio de aplicação [Isaac *et al.*, 2009]. Nossa proposta é realizar a integração semântica das terminologias de diferentes glossários, essencial para amenizar o problema de interoperabilidade entre diferentes glossários de um mesmo domínio de aplicação. Para tanto, também é proposta uma abordagem para a geração de *links* semânticos entre termos em glossários distintos. Além disso, é apresentado um estudo de caso em que o *framework* proposto é utilizado para construir um *mashup* de glossários a ser utilizado por empresas de E&P.

O *framework* proposto é utilizado na construção de um *mashup* que integra

[3] https://www.w3.org/
[4] http://www.w3.org/TR/rdf-sparql-query/
[5] Um *mashup* é um site personalizado ou uma aplicação *web* que usa conteúdo de mais de uma fonte para criar um novo serviço completo.
[6] http://www.bbc.co.uk/music
[7] http://wiki.dbpedia.org/
[8] https://musicbrainz.org/

vários glossários de termos relacionados à E&P, cujas motivações são: (*i*) enriquecer as ontologias no domínio de E&P; (*ii*) melhorar as buscas semânticas e (*iii*) facilitar a integração semântica de ontologias no domínio de E&P.

O restante do artigo está organizado como segue. A Seção 2 apresenta os conceitos para o desenvolvimento da ontologia de mashup de glossários. Na Seção 3, detalhamos o processo de publicação dos glossários como *linked data*. Na Seção 4, denotamos as etapas para a geração de *links* entre os glossários, bem como os experimentos realizados e os seus respectivos resultados obtidos na validação da metodologia proposta para a geração de links entre os glossários. Na Seção 6, discutimos os trabalhos relacionados. Por fim, na Seção 7, são delineadas as conclusões e os trabalhos futuros.

2. Ontologia de *Mashup* de Glossários

Mashups de dados interligados (LDM) são serviços interativos disponibilizados na Web que mesclam o conteúdo de diferentes fontes de dados interligados em um novo serviço [Rahm *et al.* 2007]. O processo de concepção de um LDM dá-se início com a modelagem da ontologia de domínio [Tran *et al.* 2014].

A ontologia proposta neste trabalho reusa conceitos do vocabulário SKOS (*Simple Knowledge Organization System*) [Miles *et al*. 2005]. A comunidade de Web Semântica desenvolveu o SKOS, que é baseado em representações formais de vocabulários declarados por meio de diretivas RDF, para representar diferentes sistemas de organização do conhecimento, como *thesaurus,* esquemas de classificação, glossários e taxonomias, bem como para compartilhá-los em um ambiente distribuído.

Seguindo as recomendações do SKOS, um sistema de organização do conhecimento pode ser visto como um esquema conceitual que inclui um conjunto de conceitos. O SKOS considera um conceito (identificado pela classe *skos:concept*) como a unidade mais elementar. Um conceito pode ser conectado a qualquer *string*, independentemente da linguagem natural, determinando o seu rótulo preferido (mesmo que possua outra linguagem natural), podendo, ainda, possuir infinitas descrições alternativas. Por meio das propriedades *skos:prefLabel* e *skos:altLabel*, as descrições preferidas e as suas alternativas podem ser ligadas aos conceitos. Além disso, uma ou mais notações (*skos:notation*), incluindo uma sequência de caracteres em qualquer linguagem natural, podem ser atribuídas a um conceito SKOS para identificá-lo no campo de aplicação de outro esquema conceitual. A Figura 1 mostra as principais classes e relacionamentos da ontologia proposta utilizadas na representação de *mashup* de glossários.

Além das propriedades que descrevem os conceitos, o SKOS inclui as seguintes propriedades para afirmar relacionamentos semânticos entre conceitos: hierárquicas, associativas e de equivalência, como definido no Quadro 1.

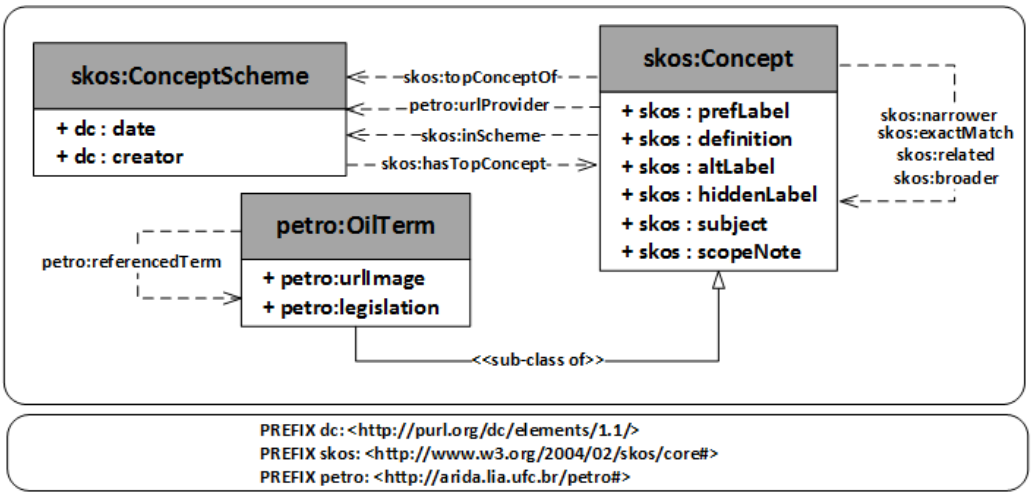

Figura 1. Ontologia de Domínio do *Mashup* de Glossários

Quadro 1. Categorias de Relações no SKOS

Categoria	Propriedades	Definição
Relações Hierárquicas	*skos:broader*	Relações hierárquicas entre conceitos, indicando que determinado conceito possui um significado mais amplo.
	skos:narrower	Relações hierárquicas entre conceitos, indicando que determinado conceito possui um significado mais específico.
Relações Associativas	*skos:related*	Relações associativas entre conceitos .
Relações de Equivalências	*skos:exactMatch*	Relações de equivalências entre conceitos que possuem alto grau de correspondência e podem ser utilizados indistintamente em uma ampla gama de aplicações.
	skos:closeMatch	Relações de equivalências entre conceitos que podem ser considerados como similares em contexto previamente determinado.
	skos:broadMatch	Relações de equivalências, considerando a estrutura hierárquica de um conceito que possui um significado mais amplo.
	skos:narrowMatch	Relações de equivalências, considerando a estrutura hierárquica de um conceito que possui um significado mais específico.
	skos:relatedMatch	Relações de equivalências, considerando as estruturas associativas existentes entre conceitos.

No contexto dos estudos da Web Semântica, inferência refere-se à possibilidade de deduzir uma afirmação por meio de outras afirmações, permitindo que processos automáticos indiquem novos conjuntos de regras e novas relações que podem ser consumidas e desenvolvidas.

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.

Usualmente, se X possui *skos:exactMatch* Y, e Z *is-a* X, então pode-se inferir que Z possui *skos:broader* com Y. Uma vez que SKOS depende do empenho de especialistas e de exaustivas atividades manuais, propomos o uso de um algoritmo, baseado em regras simples, que classificará automaticamente os mapeamentos entre termos utilizando os predicados SKOS.

Outros tipos de inferências úteis podem ser adicionados ao adotar o padrão SKOS:

- Tipificação, com base no *range* e no *domain* das propriedades e/ou classes derivadas.
- Superpropriedades, com base na extensão das propriedades.
- Relacionamentos transitivos, como *skos:broader* e *skos:narrower*.
- Propriedades inversas.

3. Publicação dos Glossários como *Linked Data*

Para a construção do *Mashup* de E&P, usamos os glossários listados na Tabela 1. Tendo como critérios a (*i*) escolha de glossários de termos do petróleo de organizações que atuam no domínio do petróleo e (*ii*) fontes de dados ofertados sem nenhuma restrição de direitos autorais para fins de pesquisa.

Tabela 1. Glossário 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

A partir da identificação dos glossários a serem utilizados, foi concebida uma arquitetura que tem como objetivo integrar as informações dos glossários de termos da área de E&P. A Figura 2 apresenta a arquitetura proposta, que é dividida em cinco componentes. Esses componentes interligados possuem como objetivo a coleta dos termos dos glossários e o tratamento dessas informações em uma estrutura comum.

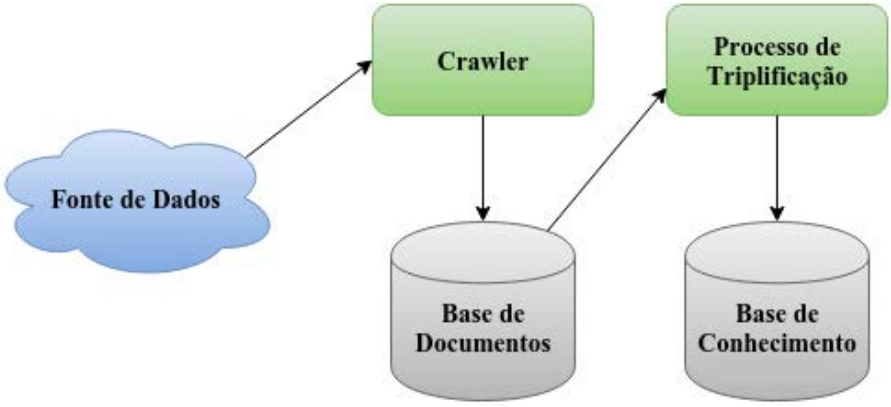

Figura 2: Arquitetura do Processo de Coleta e Tratamento das Fontes de Dados

O *crawler*[9] é responsável por mapear a estrutura dos termos de uma fonte de dados para a modelagem da base de documentos, após isso, a coleta é executada e os

[9] *Crawler* é um rastreador da Web que coleta dados sistematicamente na Internet, normalmente para fins de indexação de páginas.

termos extraídos e persistidos na base. A relação entre uma fonte de dados e um *crawler* é de um para um, ou seja, cada fonte de dados nova é necessária à implementação de um *crawler* específico pra mesma. Esses *crawlers* são responsáveis por mapear as informações necessárias das páginas, extraí-las e armazená-las no MongoDB[10] em uma mesma modelagem.

A base de documentos utiliza o MongoDB para a primeira etapa de armazenamento dos dados. O processo de triplificação é responsável por ler os documentos no MongoDB e transformá-los em um grafo RDF utilizando a modelagem da Ontologia de Domínio do *Mashup* de Glossários, detalhada na Seção 2. Por último, esse grafo RDF está pronto para ser publicado.

Esse processo de triplificação realiza uma leitura da base de documentos do MongoDB e realiza os passos de extração, transformação e carga (ETL) para as definições de cada glossário por vez. Para isso, optou-se pelo desenvolvimento de uma rotina em linguagem de programação Python para executar esse processo. Essa decisão tem como base a alta manutenibilidade desse tipo de *script*, assim como uma baixa curva de aprendizado e melhoria de tal lógica.

4. Geração de *Links* entre os Glossários

Produzir ligações entre fontes heterogêneas usando linguagens de programação convencionais é um trabalho complicado que acarreta o uso de grande quantidade de código específico para consumir cada fonte de dados ou serviço, resultando em alto custo de desenvolvimento e manutenção.

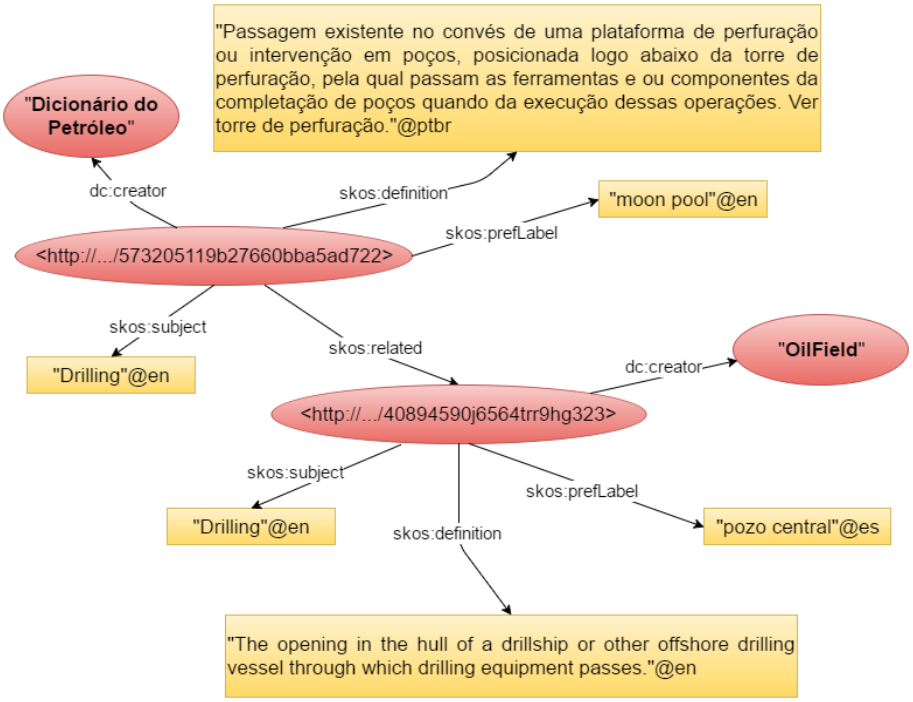

Figura 3. Exemplo de *link* Semântico na Ontologia de Glossários de E&P

As ligações ontológicas têm por objetivo determinar as relações semânticas entre

[10] https://www.mongodb.org

os elementos de diferentes sistemas de conhecimento como, ontologias e glossários. O conjunto de relações semânticas geralmente compreende equivalência de conceitos, gerando ligações de conceitos hierárquicos (*skos:broader* | *skos:narrower*) e ligações de relacionamento (*skos:related*).

Esses tipos de relacionamentos, conforme o Quadro 1, 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** e 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.
- **NT** (*Narrower Terms*) é o oposto de **BT**: 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** e 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 das atividades semiautomáticas propostas nesse trabalho.

Um exemplo de ligação semântica é apresentado na Figura 3. O termo "*moon pool"* teve o *link* semântico *skos:related* gerado após a comparação das definições. Para identificar os padrões entre as definições de cada termo, devemos executar anteriormente etapas de pré-processamento textual e de conversão dos textos para um único padrão de linguagem. De modo que a maioria das bases de glossários utilizada nesse projeto está disponível na língua inglesa, foi definido o uso desta como padrão para a conversão dos textos em uma linguagem única para a comparação dos textos.

4.1 Processo Semiautomático para Geração de *Links*

O processo de geração de *links* entre glossários pode ser visto como a identificação de termos, conceitos e relações hierárquicas que são aproximadamente equivalentes [Rahm e Bernstein, 2001]. O problema, portanto, é definir o significado de equivalência entre conceitos.

A modelagem da ontologia de domínio e 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*) *Links* RDF descobertos entre as Fontes de Dados (FD) distintas determinam as ligações entre as OFs. A Figura 4 apresenta as OFs do *Mashup* de Glossários de E&P.

Para a geração automática de predicados SKOS, implementamos um algoritmo para validar e avaliar o desempenho do método proposto, utilizando como base o predicado *skos:prefLabel* dos glossários apresentados na Tabela 1. A Figura 5 ilustra o algoritmo desenvolvido.

As regras do algoritmo gerador de predicados SKOS seguem os seguintes passos: (*i*) sendo **t** = termo e **s** = *string*, definimos que $\mathbf{t}_1[\mathbf{s}_1 \ldots \mathbf{s}_n]$ e $\mathbf{t}_2[\mathbf{s}_1 \ldots \mathbf{s}_n]$ possui ligação *skos:exactMatch* se $\mathbf{t}_1$ for igual a $\mathbf{t}_2$ e o predicado *dc:creator* de $\mathbf{t}_1$ e $\mathbf{t}_2$ forem diferentes; (*ii*) caso $\mathbf{t}_1[\mathbf{s}_1]$ seja igual a $\mathbf{t}_2[\mathbf{s}_1 \ldots \mathbf{s}_n]$ e *dc:creator* de $\mathbf{t}_1$ e $\mathbf{t}_2$ sejam iguais, definimos que

$\mathbf{t}_1$ é *skos:narrower* de $\mathbf{t}_2$ e $\mathbf{t}_2$ é *skos:broader* de $\mathbf{t}_1$; (*iii*) finalmente, definimos que haverá ligação reflexiva *skos:related* entre $\mathbf{t}_2$ e $\mathbf{t}_3$ quando $\mathbf{t}_2[\mathbf{s}_1 \ldots \mathbf{s}_n]$ e $\mathbf{t}_3[\mathbf{s}_1 \ldots \mathbf{s}_n]$ possuírem ligações semânticas *skos:narrower* e *skos:broader* com $\mathbf{t}_1$. A descoberta desses relacionamentos terminológicos a partir de esquemas de fonte é uma das atividades semiautomáticas propostas neste trabalho.

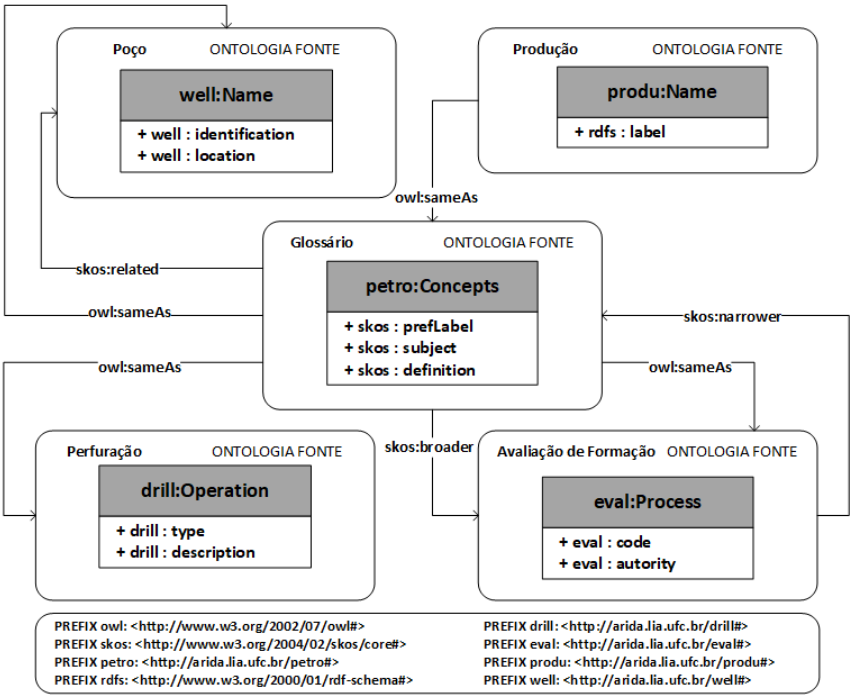

Figura 4. Ontologias Fontes do Mashup de Glossários de E&P

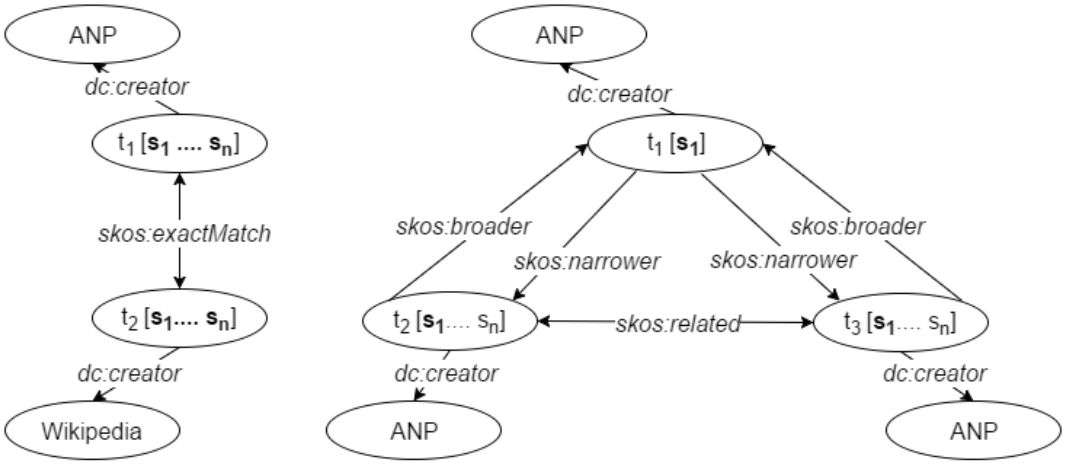

Figura 5. Descoberta de predicados SKOS

As regras definidas no algoritmo para geração de predicados SKOS seguem uma consistência formal, com um grau satisfatório de sucesso. Elas foram definidas usando um método iterativo, incluindo as seguintes etapas: (*i*) definição intuitiva de regras utilizando uma amostra de 100 predicados *skos:prefLabel*; (*ii*) formalização das regras em formato algorítmico; (*iii*) teste do algoritmo em uma amostra maior; (*iv*) avaliação

dos resultados; e, (*v*) refinamento das regras, se fosse o caso. Foram necessárias cinco iterações para obtermos a versão atual do algoritmo. O algoritmo gera automaticamente quatro propriedades de mapeamento no SKOS: *skos:exactMatch*, *skos:broader*, *skos:narrower* e *skos:related*.

4.2. Resultados da Geração de *Links*

Aplicando o algoritmo proposto, realizamos a coleta de 5.373 conceitos do *Schlumberger Oilfield Glossary* utilizando como base para as comparações o predicado *skos:preflabel*. Os resultados obtidos e a proporção dos predicados SKOS gerados pelo algoritmo são apresentados nas Tabelas 2 e 3.

Para validar os resultados, foram selecionados de 341 exemplos de predicados SKOS processados pelo algoritmo. Todos os *links* gerados pelo algoritmo foram analisados por um especialista com o intuito de comprovar a eficácia do modelo proposto.

Tabela 2. Predicados SKOS gerados

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

A geração dos predicados *skos:exactMatch* atingiram 98,63% de taxa de acerto, de acordo com a avaliação do especialista, o que consideramos aceitável, uma vez que utilizamos apenas as técnicas de pré-processamento textual *stemming* e a remoção de *stopwords* para melhorar o resultados dos alinhamentos entre os termos [Avila e Soares, 2012].

De modo semelhante, após validação do especialista, os *links* gerados para o predicado *skos:related*, obtivemos 85,87% de acerto. A Tabela 3 apresenta alguns exemplos de outros predicados que deveriam ter sido gerados. De acordo com as regras do algoritmo desenvolvido, o *link* do predicado *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 necessitam de tratamento. Especificamente no caso do termo ***vugular porosity***, teria que ser gerado um *link* com o predicado *owl:sameAs* ligando-o a ***vug***. Faz-se necessário o refinamento do algoritmo para tratar esses casos que não dependem da similaridade dos termos, mas sim do entendimento prévio do domínio que está sendo utilizado para a geração de ligações semânticas.

Por último, para os predicados *skos:broader* e *skos:narrower*, obtivemos, respectivamente, 72,47% e 72,42%. O desempenho desses resultados está diretamente ligado à dependência que o algoritmo tem com a comparação das *strings* que compõem cada termo. Dessa forma, os *links* somente serão gerados se ocorrer o alinhamento (*matching*) entre os termos comparados. Esse tipo de limitação será tratado em trabalhos futuros, utilizando, por exemplo, o emprego da *Wordnet* [Fellbaum, 1998].

Tabela 3. Avaliação dos links SKOS

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

Avaliando conjuntamente os predicados SKOS gerados, o algoritmo obteve 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 o algoritmo eficaz para o mapeamento das correspondências entre a OD e as OFs. O desempenho do algoritmo é menos convincente para as outras propriedades do SKOS, com resultados acima de 72%, o que consideramos eficaz.

5. Trabalhos Relacionados

A proposta de [Kazi e Kurian, 2014], baseia-se em uma metodologia de enriquecimento de ontologias extraindo padrões de bases conhecimento por meio de novas inferências derivadas do próprio domínio, 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.

No trabalho de [d'Aquin *et al.*, 2012] é definido um novo método de descoberta de conhecimento combinando (*i*) técnicas de mineração de dados para fazer emergir modelos implícitos de dados e (*ii*) o uso padrões de engenharia de ontologias para capturar esses modelos de forma reutilizável. Os resultados obtidos apontam a redução de tempo na preparação de dados e interpretação de resultados, nas atividades de consulta de especialistas e na construção de ontologias e nas ligações semânticas.

O entendimento de conceitos por parte dos humanos e a inferência e descoberta de conhecimento por meio de raciocínio automático por parte das máquinas são apresentados em [Michael *et al.*, 2001]. A definição de um conceito (termo) dentro de uma ontologia permite que pessoas compreendam a própria ontologia, permitindo que máquinas consigam inferir conhecimento sobre o domínio, facilitando a ligação entre ontologias heterogêneas e, consequentemente, a inferência de novos conhecimentos.

Na pesquisa de [Ge e 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 petróleo.

[Biasiotti *et al.*, 2011] propõe um *framework* para o controle da interoperabilidade entre *thesaurus*, buscando garantir uma validação cruzada entre as coleções de termos e suas diferentes línguas. Em particular, foram apresentadas as normas para o desenvolvimento de um estudo de caso utilizado no mapeamento de cinco *thesauri* da União Europeia. Os experimentos apontaram bons resultados de alinhamento entre os *thesauri*, utilizando o algoritmo de similaridade *Levenshtein*, porém, no contexto léxico, o desempenho do algoritmo não foi representativo.

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-se trabalhar no sentido de compreender o domínio que está sendo explorado e buscar os modelos mais apropriados para mapear/desenvolver a sua respectiva definição, identificando os padrões que melhor se adaptam àquele contexto.

6. Conclusão

Demonstramos neste artigo um algoritmo, 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. O algoritmo foi testado utilizando glossários de termos no domínio do petróleo, mas pode, em princípio, ser utilizado 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 às 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 o algoritmo 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, o algoritmo proposto depende principalmente da qualidade das terminologias comparadas. O uso de terminologias com princípios mais sistemáticos e relações hierárquicas mais transparentes podem 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 a *framework* proposta 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. e 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.

Batini, C.; Lenzerini, M.; Navathe, S. B. *A comparative analysis of methodologies for database schema integration*. ACM Comput. Surv., ACM, New York, NY, USA, v. 18, n. 4, p. 323–364, dez. 1986.

Biasiotti , M.A., Faro, S., e Francesconi, E. (2011). *Thesaurus Mapping for Promoting Semantic Interoperability of European Public Services*. eChallenges e-2011 Conference Proceedings, pages 1-10.

d'Aquin, M., Kronberger, G., e Suarez-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.

Fellbaum, C. (1998). *WordNet – An Electronic Lexical Database.* MIT Press.

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

Hull, R. *Managing semantic heterogeneity in databases: A theoretical prospective*. In: Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems. New York, NY, USA: ACM, p. 51–61.

Isaac, A.; Wang, S.; Zinn, C.; Matthezing, H.; van der Meij, L. e Schlobach, S. (2009). *Evaluating Thesaurus Alignments for Semantic Interoperability in the Library Domain*. IEEE Intelligent Systems Special Issue on AI and Cultural Heritage.

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

Lenzerini, M. *Data Integration: A Theoretical Perspective*. In: Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of database Systems. New York, NY, USA: ACM, 2002. (PODS '02), p. 233–246.

Michael, J., Mejino Jr., J., e Rosse, C. (2001). *The Role of Definitions in Biomedical Concept Representation*. Proceedings Annual Symposium. *AMIA*, pages 463–467.

Miles, A., Matthews, B., Wilson, M., e Brickley, D. (2005). *SKOS Core: Simple Knowledge Organization 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.

Rahm, E. e Bernstein, P. (2001). *A Survey of Approaches to Automatic Schema Matching*. The International Journal on VLDB, vol. 10, no. 4, pp. 334–350.

Rahm, E., Thor, D., e 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., e Le, T. M. (2014). *Linked Data Mashups: A Review on Technologies*. Applications and Challenges, pages 253–262. Springer International Publishing, Cham.

Vidal, V. M. P. *et al*. *Advanced Information Systems Engineering*: 27th international Conference, caise 2015, Stockholm, Sweden, June 8-12, 2015, Proceedings. In: Springer International Publishing, 2015. cap. Specification and Incremental Maintenance of Linked Data Mashup Views, p. 214–229.