Uma Arquitetura Baseada em Ontologias para Composição Semântica de Workflows

Luan Fonseca Garcia[1], **Jean-François Rainaud**[2], **Mara Abel**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS – Brasil

[2]IFPEN - IFP Energies Nouvelles
Rueil Malmaison - França

{luan.garcia,marabel}@inf.ufrgs.br,jean-francois.rainaud@ifpen.fr

Abstract. *The E&P industry generates a big amount of data every day. These data are used by the involved professionals in various workflows. Due to the background difference of these professionals and the great amount of data involved, semantic interoperability between these data is a big challenge. In this work, we propose an ontology-based framework for semantic workflow composition in the Geology domain. We combine together an ontology network, the notion of ontology-based data access and the concept of data lake to propose a solution for the data interoperability problem during workflow execution.*

Resumo. *A indústria de E&P gera diariamente uma grande quantidade de dados. Estes dados são utilizados pelos profissionais envolvidos em diversos workflows. Devido a diferença de formação destes profissionais e a grande quantidade de dados envolvida, a interoperabilidade semântica destes dados é um grande desafio. Neste trabalho, propomos uma arquitetura baseada em ontologias para composição semântica de workflows para o domínio da Geologia. Combinamos uma rede de ontologias, a noção de acesso a dados baseado em ontologias e o conceito de data lake para propor uma solução ao problema de interoperabilidade de dados durante a execução de workflows.*

1. Introdução

A cadeia de exploração e produção de petróleo (E&P) gera diariamente uma enorme quantidade de dados provenientes de variadas fontes. Entre eles, dados sísmicos, dados de poço, dados de perfuração, entre outros. Estes dados são utilizados em variados *workflows* que são ou realizados pelos profissionais envolvidos ou automaticamente, por sistemas. A indústria petrolífera depende do uso eficiente destes dados para a construção de modelos computacionais e para reduzir incerteza e risco nas tomadas de decisões [Werlang et al. 2014].

Segundo Edison [Edison et al. 2011], o entendimento da indústria de E&P é de que a interoperabilidade de dados é crucial para conectar tecnologias e inovações com as pessoas e processos em sua cadeia. Wache [Wache et al. 2001] define Interoperabilidade como o problema de integrar sistemas computacionais heterogêneos e distribuídos. Para isto, é necessário que, além de possuir acesso total aos dados, estes dados possam ser interpretados e processados sem nenhuma perda do significado pretendido dos mesmos.

Este trabalho investiga o tema de interoperabilidade semântica de dados na cadeia de exploração e produção de petróleo, faz parte de um estudo de doutorado e tem previsão de defesa para fevereiro de 2020. Propomos neste artigo uma arquitetura para descrição e composição semântica de *workflows* no domínio da Geologia. Esta arquitetura é baseada em uma Rede de Ontologias e no conceito de Acesso a Dados Baseado em Ontologias (OBDA).

Workflow no contexto deste trabalho tem sua origem na disciplina de Modelagem de Processos de Negócios (BPM). Um *workflow* é uma abstração de um processo de negócios e compreende uma série de passos (conhecidos como tarefas ou atividades), dependências entre tarefas, regras de fluxo e participantes. Uma tarefa pode representar uma tarefa humana ou um sistema computacional [Cardoso and Sheth 2003].

Uma rede de ontologias é uma coleção de ontologias que estão conectadas entre si através de relacionamentos como alinhamentos, dependências, versionamentos, entre outras relações [Suárez-Figueroa et al. 2012]. Grandes ontologias monolíticas são de difícil compreensão e possuem um alto custo de manutenção. Utilizar uma rede de ontologias em face a uma grande ontologia monolítica permite modularizar domínios distintos em ontologias menores e mais específicas, o que possibilita o reuso de ontologias já existentes, também facilitando a manutenção do conhecimento.

No paradigma de Acesso a Dados Baseado em Ontologias, uma ontologia funciona como um esquema de alto nível sobre uma fonte de dados e fornece um vocabulário para consultas de usuários. O sistema reescreve estas consultas de usuários para a linguagem da fonte de dados original e realiza esta consulta [Civili et al. 2013].

As contribuições deste trabalho são a definição de uma rede de ontologias que suporta a formalização de conhecimento geológico de forma integrada com tarefas e processos realizados no domínio, e uma arquitetura que suporta a integração deste conhecimento com o armazenamento de suas instâncias em um banco de dados relacional aliado a um *data lake* através da noção de OBDA.

2. Trabalhos Relacionados

Investigamos trabalhos que buscassem a descrição de *workflows* com o suporte de ontologias. A seguir, faremos uma descrição dos principais trabalhos e uma breve comparação com a arquitetura que estamos propondo.

Em [Cardoso and Sheth 2003], os autores utilizam uma ontologia para descrever semanticamente tarefas e interfaces de *web services*. O foco do trabalho é descobrir e apresentar ao usuário *web services* já existentes. Eles compõem tarefas de usuário com *web services* para definir *workflows* a serem executados. Este trabalho difere do nosso pois os autores não utilizam ontologias para definir o conhecimento do domínio em que as tarefas são realizadas, não contempla tarefas que não sejam automáticas e também não se preocupam com o armazenamento dos dados.

Em [Dang et al. 2008], os autores propõem uma arquitetura baseada em ontologia para *workflows* no domínio da medicina. Eles extraíram os principais conceitos do domínio das agências *Siemens Medical Solution* e *Agency for Healthcare Research*, e desenvolveram uma grande ontologia contendo cinco visões diferentes. O usuário da arquitetura proposta pode compor *workflows* através de tarefas já existentes na base de

dados. O sistema gera arquivos *BPEL* para serem executados por alguma orquestrador de *workflow*. Nosso trabalho difere na forma como os dados são acessados e armazenados. Além disso propomos uma rede de ontologias que é alinhada por uma ontologia de topo, enquanto o trabalho de Dang e colegas utiliza uma ontologia única.

Em [Belaid et al. 2010], os autores propõem composição de *workflows* de *web services*. Eles utilizam uma ontologia de domínio para descrever dados geológicos e uma ontologia de serviços para descrever *web services* e então mapeiam os conceitos geológicos com os respectivos conceitos na ontologia de serviços como entradas e saídas. O sistema gera arquivos *BPEL* com a composição dos *web services* em workflows. O trabalho de Belaid e colegas difere do nosso por não suportar a descrição de tarefas que não sejam execuções de *web services* e não fornecer meio de armazenar os dados que serão utilizados durante a execução destes *workflows* ou que serão gerados após sua execução.

3. Conceitos e Tecnologias Utilizadas

As ontologias desenvolvidas ou reutilizadas neste trabalho estão alinhadas com a ontologia de topo Basic Formal Ontology (BFO). A BFO é uma ontologia de topo desenvolvida para auxiliar na integração de dados para a pesquisa científica. Ela foi construída deliberadamente pequena e seus conceitos são independentes de domínio. A BFO auxilia na interoperabilidade entre vários domínios fornecendo uma estrutura comum de alto nível, o que propicia a informação existente em diferentes ontologias a fazer parte de uma arquitetura comum para categorização e raciocínio [Arp et al. 2015].

Para nos beneficiarmos do conceito de ODBA utilizamos o sistema *open-source* Ontop. O sistema Ontop transforma um banco de dados relacional em um grafo RDF virtual, e então mapeia os conceitos e relações da ontologia ao banco de dados. Este grafo virtual pode ser então consultado através da linguagem SPARQL, que será traduzida para consultas SQL sobre o banco de dados de forma transparente ao usuário [Calvanese et al. 2017].

Como forma de armazenar arquivos e recursos que são utilizados durante a cadeia de E&P combinamos a estrutura formal dada pelas ontologias aliadas ao banco de dados relacional com um repositório de dados do tipo *data lake*. Um *data lake* é um repositório massivo de dados não estruturado baseado em tecnologias de baixo custo que incrementam a captura, refino, arquivamento e exploração de dados brutos dentro de uma empresa [Fang 2015].

4. Arquitetura Proposta

A arquitetura proposta neste artigo pode ser dividida em duas grandes camadas. Uma Camada de Conhecimento, responsável por armazenar e gerenciar o conhecimento, e outra Camada de Aplicação, responsável por gerenciar buscas, visualizações e composição semântica de *workflows* sobre este conhecimento.

A camada de conhecimento possui dois módulos. Uma rede de ontologias para descrever conhecimento de domínios da Geologia e de Modelagem de Processos de Negócio e uma camada contendo um banco de dados relacional e um *data lake*. O mapeamento entre a rede de ontologias e o banco de dados ocorre através do sistema Ontop. A figura 1 apresenta um esquema da camada de conhecimento.

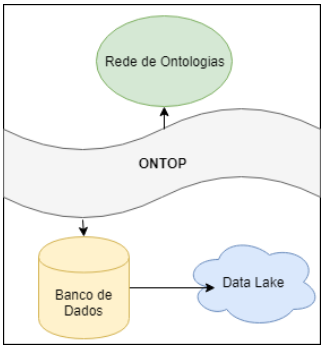

Figura 1. Arquitetura do Módulo de Conhecimento.

A rede de ontologias que propomos possui dois ramos de conhecimento distintos: um para descrever conhecimento relacionado a Geologia, e outro para descrever conhecimento relacionado a Modelagem de Processos de Negócio. Como base para alinhar estes dois ramos utilizamos a ontologia de topo BFO. Para cada ramo, uma ontologia core descreve conceitos genéricos que são compartilhados entre seus subdomínios, e ontologias de domínio específicas descrevem o conhecimento de domínio que não é compartilhado por todas as outras ontologias. A figura 2 apresenta um exemplo da hierarquia existente entre as ontologias contidas na rede.

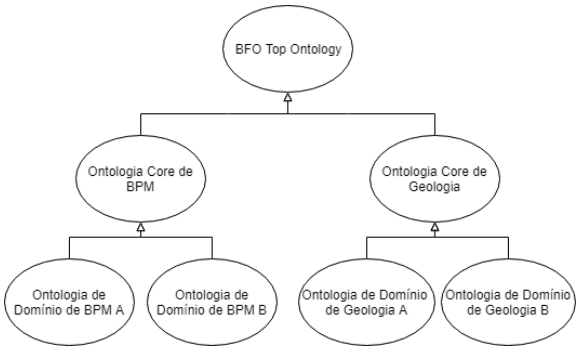

Figura 2. Hierarquia entre ontologias contidas na rede.

Para o ramo da Geologia, a ontologia core define termos genéricos como Objeto Geológico, Unidade Geológica e Processo Geológico e as relações existentes entre eles, enquanto as ontologias de domínio especializam estes conceitos genéricos, como por exemplo o conceito de Unidade Litológica (volume de rocha distinguível por suas características litológicas).

Para o ramo de BPM, a ontologia core define conceitos genéricos como Tarefa e *Workflow*, enquanto as ontologias de domínio especializam conceitos como Interpretar Textura ou Descrever Contatos, tarefas mais específicas. A ontologia core também contém relações que permitem compor novos *workflows* semânticos a partir de tarefas que estão definidas nas ontologias de domínio, como por exemplo relações de fluxo, que definem a sequência em que as tarefas devem ser executadas, e relações de entrada e saída.

O módulo de armazenamento de dados é composto por um banco de dados relacional e um *data lake*. O banco de dados armazena instâncias dos conceitos existentes nas ontologias. Estas instâncias podem ser referências a objetos da realidade, como um

poço de petróleo em específico ou uma unidade litológica específica, ou referências a arquivos, como por exemplo um arquivo *XML* ou um arquivo com dados sísmicos, e que estão armazenados no *data lake*. Além disso, o banco também armazena instâncias de tarefas e *workflows* específicos e dados relacionados a eles, como o usuário que realizou a tarefa, quais as instâncias de dados de entrada utilizados ou quais as instâncias de dados resultantes ao fim do *workflow*.

A conexão entre o módulo de armazenamento de dados e a rede de ontologias é realizada através do sistema Ontop. Os mapeamentos são realizados com a linguagem nativa de mapeamento do *framework* e relacionam classes e relações das ontologias com visões do banco de dados por SQL. As ontologias junto com os mapeamentos resultam em uma grafo RDF virtual que pode ser consultado através da linguagem SPARQL.

A camada de aplicação é uma interface responsável por gerenciar consultas do usuário sobre a camada de conhecimento e compor *workflows*. Consultas podem ser no nível de conceitos (Universais), cujo resultado são definições ontológicas, ou no nível de instâncias (Particulares), cujo resultado são instâncias do banco ou arquivos existentes no *data lake*. Por exemplo, o usuário pode consultar *o que é uma Unidade Litológica*, e neste caso o resultado deve ser a definição ontológica de Unidade Litológica, ou o usuário pode consultar *quais Unidades Litológicas existem*, e neste caso o resultado deve ser todas as instâncias de Unidade Litológica específicas existentes no banco de dados.

A composição de *workflows* se dá através da associação de tarefas e *workflows* existentes com as relações definidas nas ontologias de BPM. A figura 3 apresenta um exemplo fragmentado de como a tarefa de Interpretação Textural pode ser definida na ontologia. A tarefa possui como entrada uma Unidade Litológica e tem como saída a Granulometria desta Unidade. Ou seja, para esta tarefa o usuário deve analisar uma Objeto Geológico (entidade 3D que existe no mundo real) e definir o valor de sua qualidade Granulometria (tamanho dos grãos de rocha nesta unidade). Para instanciar esta tarefa devemos possuir uma instância de Unidade Litológica no banco (que referencia uma Unidade do mundo real). Ao término desta tarefa deve ser adicionado ao banco o valor da qualidade Granulometria para essa instância de Unidade Litológica específica e informações sobre a realização desta tarefa, como usuário, valores de entrada, e valores de saída.

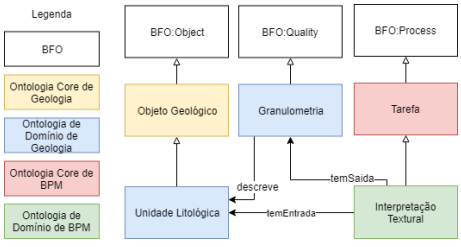

Figura 3. Exemplo de definição para a tarefa de Interpretação Textural.

5. Conclusão

Neste artigo, propomos uma arquitetura baseada em ontologias para composição semântica de *workflows*. A arquitetura utilizada uma rede de ontologias para definir formalmente conceitos do domínio e como relacioná-los com tarefas e *workflows* que os utilizam.

A arquitetura proposta permite representar conhecimento de domínio, descrever semanticamente as tarefas e *workflows*, mapeá-los para os conceitos de domínio como entradas e saídas, compor novos *workflows* a partir das tarefas e *workflows* já existentes, armazenar informações e arquivos que serão de fato utilizados nos *workflows* a serem executados e rastrear os dados que foram utilizados durante a instanciação de um *workflow*.

Este trabalho está em andamento e possui previsão de defesa da tese de doutorado para fevereiro de 2020. A previsão para término dos mapeamentos entre a rede de ontologias e o banco de dados é para o final de 2018 e a validação dos resultados está prevista para o primeiro semestre de 2019.

Referências

Arp, R., Smith, B., and Spear, A. D. (2015). *Building ontologies with basic formal ontology*. Mit Press.

Belaid, N., Ameur, Y. A., Jean, S., and Rainaud, J.-F. (2010). Toward a semantic management of geological modeling workflows. In *KEOD*, pages 282–287.

Calvanese, D., Cogrel, B., Komla-Ebri, S., Kontchakov, R., Lanti, D., Rezk, M., Rodriguez-Muro, M., and Xiao, G. (2017). Ontop: Answering sparql queries over relational databases. *Semantic Web*, 8(3):471–487.

Cardoso, J. and Sheth, A. (2003). Semantic e-workflow composition. *Journal of intelligent information systems*, 21(3):191–225.

Civili, C., Console, M., De Giacomo, G., Lembo, D., Lenzerini, M., Lepore, L., Mancini, R., Poggi, A., Rosati, R., Ruzzi, M., et al. (2013). Mastro studio: managing ontology-based data access applications. *Proceedings of the VLDB Endowment*, 6(12):1314–1317.

Dang, J., Hedayati, A., Hampel, K., and Toklu, C. (2008). An ontological knowledge framework for adaptive medical workflow. *Journal of biomedical informatics*, 41(5):829–836.

Edison, L. S., Brantley, J. D., and Edwards, S. (2011). The value of smarter oil and gas fields. *IBM Center for Applied Insights*.

Fang, H. (2015). Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. In *Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), 2015 IEEE International Conference on*, pages 820–824. IEEE.

Suárez-Figueroa, M. C., Gómez-Pérez, A., Motta, E., and Gangemi, A. (2012). *Ontology engineering in a networked world*. Springer Science & Business Media.

Wache, H., Voegele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., and Hübner, S. (2001). Ontology-based integration of information-a survey of existing approaches. In *IJCAI-01 workshop: ontologies and information sharing*, volume 2001, pages 108–117. Citeseer.

Werlang, R., Abel, M., Perrin, M., Carbonera, J. L., and Fiorini, S. R. (2014). Ontological foundations for petroleum application modeling. In *18th International Conference on Petroleum Data, Integration and Data Management*.