Uma Abordagem para Publicação de Dados de Proveniência de *Workflows* Científicos na *Web* Semântica

Rachel Castro[1], Renan Souza[1], Vitor Silva[1], Kary Ocaña[2], Daniel de Oliveira[3], Marta Mattoso[1]

[1]Universidade Federal do Rio de Janeiro (UFRJ) – Rio de Janeiro – RJ, Brasil
[2]Laboratório Nacional de Computação Científica (LNCC) – Petrópolis – RJ, Brasil
[3]Universidade Federal Fluminense (UFF) – Niterói – RJ, Brasil

rachel@poli.ufrj.br, {renanfs,silva,marta}@cos.ufrj.br,
karyann@lncc.br, danielcmo@ic.uff.br

Resumo. *A cada ano, cresce o volume de dados de proveniência coletados por meio da execução de simulações computacionais modeladas como workflows científicos. Tais dados auxiliam a reprodutibilidade das simulações e oferecem maior confiabilidade aos resultados. Apesar de essas simulações produzirem dados de proveniência que são enriquecidos com dados de domínio, eles são comumente armazenados em bancos de dados privados aos quais um número restrito de cientistas tem acesso. Isso reduz a capacidade analítica e a possibilidade de se inferir conhecimento a partir dos dados. No entanto, as tecnologias de Web Semântica (WS) conseguem facilitar o acesso público a dados na Web de forma estruturada, padronizada, aberta e interoperável. Neste artigo, propomos a adoção de tais tecnologias para publicar dados de proveniência de workflows na WS, além de propor uma nova ontologia. A exemplificação da abordagem é baseada em um estudo de caso real na área de bioinformática.*

1. Introdução

Workflows científicos são comumente usados para apoiar o desenvolvimento e a execução de simulações computacionais complexas, que podem manipular grandes volumes de dados. Dados de proveniência, que contemplam informações tanto sobre a estrutura do *workflow*, quanto sobre a maneira pela qual os dados são derivados, conseguem garantir a reprodução de tais simulações [Freire *et al*. 2008]. Assim, o cientista é capaz de analisar o fluxo de dados envolvido no processamento do *workflow*, desde o início até o final da execução da simulação computacional.

Os dados de proveniência costumam ser armazenados de forma que apenas um conjunto restrito de cientistas tem acesso às informações. Tal abordagem pode ser interessante do ponto de vista da privacidade dos dados, porém limita a capacidade analítica. A publicação, por outro lado, é muito vantajosa tanto para cientistas que produziram os dados, que terão suas descobertas validadas, quanto para cientistas que analisam os dados produzidos por terceiros, que poderão integrá-los em suas próprias pesquisas. Entretanto, considerando dados que possam ser publicados, três fatores dificultam a publicação: (i) uso de sistemas de gerência de banco de dados (SGBD) privados; (ii) uso de arquivos não-estruturados e com capacidade limitada de consultas; e (iii) criação própria da organização dos dados, sem seguir um padrão comum.

Na *Web* Semântica (WS), são utilizados formatos de dados abertos recomendados pelo W3C, como OWL e RDF, para definir a representação do domínio de forma estruturada. Ademais, SGBD semânticos possibilitam consultas estruturadas, e favorecem a publicação na *Web* e facilitam a interoperabilidade. Assim, o uso de tecnologias da WS se torna uma alternativa atraente para enriquecer trocas e análises de dados resultantes de simulações computacionais, possibilitando integração de dados de diferentes áreas do conhecimento. Dessa forma, o objetivo deste trabalho é apresentar uma estratégia para publicar de dados de proveniência de *workflows* na WS. Para tanto, seguimos uma metodologia dividida em quatro etapas [Souza *et al.* 2014]: Análise do Domínio, Engenharia da Ontologia, Triplificação, e Publicação.

O restante deste artigo é organizado da seguinte forma. A Seção 2 apresenta a análise do domínio realizada e a ontologia que propomos, chamada de PROV-O-Wf. A Seção 3 descreve a aplicação proposta para instanciar dados no formato RDF seguindo a ontologia PROV-O-Wf e a Seção 4 apresenta um estudo de caso. Finalmente, a Seção 5 conclui este artigo, apresenta trabalhos relacionados e aponta os trabalhos futuros.

2. Análise do Domínio e Engenharia da Ontologia

A Análise do Domínio é a primeira etapa da metodologia usada para publicação de dados na WS [Souza *et al.* 2014]. Para este trabalho, o foco dessa etapa é estudar os principais conceitos de simulações computacionais modeladas como *workflows*. Nessas simulações, as aplicações são complexas, manipulam grandes volumes de dados e requerem execução em ambiente de Processamento de Alto Desempenho (PAD), como *clusters* ou nuvens. Tais aplicações são encadeadas e os dados produzidos por uma aplicação são consumidos por outra, formando um *workflow*. Os Sistemas de Gerência de *Workflows* Científicos (SGWfC) paralelos são capazes de apoiar a execução desses *workflows* em ambientes de PAD. Uma característica comum desses SGWfC é coletar proveniência dos dados que fluem pelo fluxo associado ao *workflow*. Quanto mais fina a granularidade dos dados, maior o potencial analítico das consultas na base de proveniência. Além disso, é possível associar dados sobre a execução a dados do domínio das aplicações sendo gerenciadas pelo SGWfC, de modo a prover análises ainda mais ricas [Costa *et al*., 2013]. Dadas essas características, definimos que o domínio deste trabalho são dados de proveniência de *workflows* gerados por um SGWfC capaz de coletar dados em granularidade fina.

Definido o domínio, iniciamos a Engenharia da Ontologia, na qual desenvolvemos a ontologia para o domínio. Entretanto, antes de começarmos a construí-la do princípio, o conceito de reutilização de ontologias é muito importante nesta etapa, pois existem entidades que estão presentes em diversos domínios. Assim, diminuímos o esforço para modelar conceitos comuns e aumentamos o enfoque em modelar conceitos específicos. Além disso, a reutilização de um modelo já existente permite que um padrão seja difundido, o que facilita a interoperabilidade entre os dados na WS e também permite que dados com potencial para serem ligados (do inglês, *Linked Data*) sejam identificados [Souza *et al.* 2014]. Por isso, a seguir, discutimos a reutilização de ontologias existentes relacionadas à proveniência de dados e *workflows*.

A ontologia IntelLeo [Jovanovic *et al*. 2012], apesar de modelar *workflows* e apoiar execução paralela, não menciona dados de proveniência, dificultando a sua utilização no nosso domínio. Procurou-se também por ontologias de SGWfC existentes,

porém só foi possível encontrar a ontologia do sistema Taverna, chamada de MyGrid [Wolstencroft *et al*. 2008], a qual foca apenas em *workflows* da área da bioinformática e a modelagem não utiliza padrões do W3C. Como procuramos por ontologias independentes de domínios científicos específicos e acreditamos que padrões definidos pelo W3C devam ser seguidos na WS, decidimos por não reutilizá-la. Por outro lado, a ontologia PROV-O segue o modelo de dados de proveniência PROV-DM [Moreau *et al*. 2013], que é uma recomendação do W3C. Já a ontologia *Open Provenance Model for Workflows* (OPMW) descreve proveniência de *workflows* e é baseada no *Open Provenance Model*, um padrão reconhecido [Moreau *et al*., 2008]. Após o surgimento da PROV-O, a OPMW foi modificada, de modo a estender a PROV-O [Garijo *et al*. 2014]. Como a PROV-O é uma recomendação do W3C e a OPMW segue padrões reconhecidos e, além disso, ambas descrevem proveniência independente de domínio científico específico, decidimos reutilizá-las.

A PROV-O foi usada como um metamodelo para a ontologia proposta, chamada de PROV-O-Wf. Por ser uma referência para a modelagem de proveniência, as três principais classes da PROV-O, bem como suas representações visuais, serviram de base para a PROV-O-Wf: entidade (representadas por elipses), atividade (retângulos) e agente (pentágonos). Na taxonomia, todas as classes da PROV-O-Wf herdam de uma dessas classes da PROV-O. Além da PROV-O, nossa ontologia reaproveita conceitos do modelo de dados para proveniência retrospectiva, conhecido como PROV-Wf [Costa *et al*., 2013], que utiliza o PROV-DM como padrão, facilitando o mapeamento para a PROV-O. Ademais, também reusamos diretamente diversas classes da ontologia OPMW e acrescentamos alguns conceitos do nosso domínio, como mostramos a seguir.

A PROV-O-Wf é constituída de doze classes, divididas em duas partes principais, as que modelam a estrutura do *workflow* e as que modelam sua execução (Figura 1). A classe *WorkflowTemplate* da OPMW é usada para agrupar as especificações do *workflow*. Todas as execuções de um *workflow* devem estar associadas às suas especificações. A classe *WorkflowTemplateProcess* representa as atividades definidas para um *workflow*. A classe *TemplateRelation* foi acrescentada para expressar os dados que serão consumidos por uma atividade, podendo ser definida por um conjunto de campos que são representados pela classe *ParameterVariable* da OPMW. Em relação à execução, a classe *WorkflowExecutionAccount* armazena todas as execuções dos *workflows* e a *WorkflowExecutionProcess* descreve as atividades executadas. Já a *ExecutionTask* foi criada para abranger SGWfC que apoiam execuções paralelas, representando a menor unidade paralelizável do *workflow*. Por último, a classe *WorkflowExecutionArtifact* representa um recurso gerado ou consumido, podendo ser um arquivo, representado pela classe *File*.

Além de SGWfC paralelos, a ontologia PROV-O-Wf possibilita que os principais agentes de um *workflow* sejam identificados. A classe *Program* representa os programas invocados por cada atividade e a classe *Machine* identifica qual máquina do ambiente de PAD executou cada instância de *ExecutionTask*. A classe *Scientist* representa o cientista responsável por executar o *workflow*, agente de extrema importância na atribuição de responsabilidades. Quanto às propriedades de objeto, procurou-se priorizar as que já são definidas na OPMW e na PROV-O. No entanto, a propriedade *dependsOn* foi criada para representar a dependência de dados entre

atividades de um *workflow*, pois nenhuma propriedade que expressasse essa relação foi encontrada nas ontologias reutilizadas.

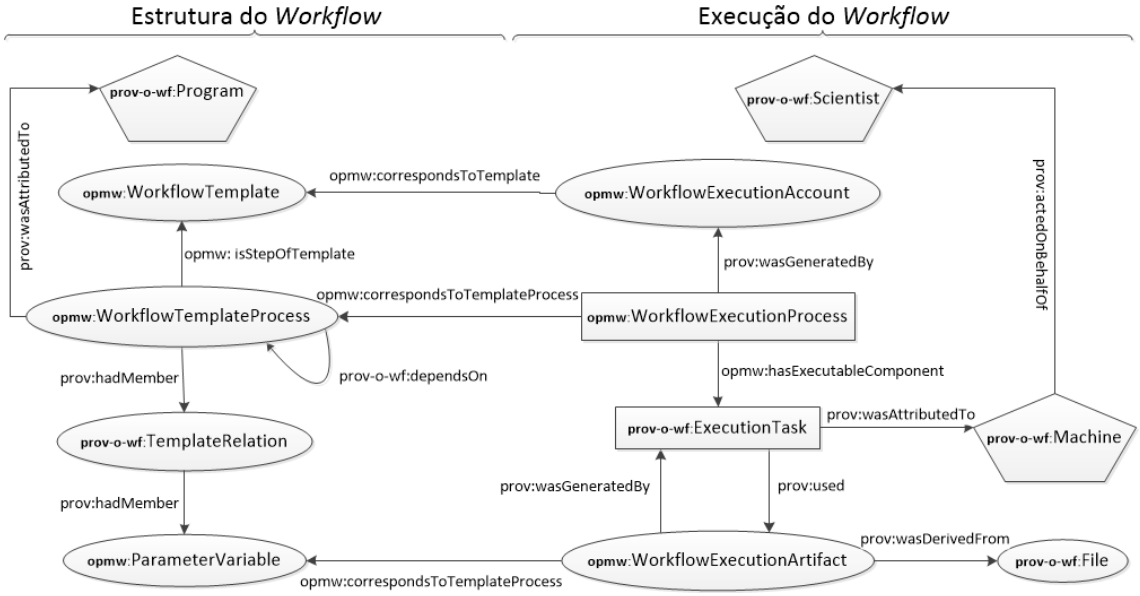

Figura 1. A Ontologia PROV-O-Wf

3. Triplificação e Publicação

Uma vez modelada a ontologia PROV-O-Wf, inicia-se a triplificação, isto é, geração das triplas RDF seguindo a ontologia e inserção em um repositório de triplas. Nesta etapa, considerou-se que os dados de proveniência são armazenados em um banco de dados relacional, como em diversos SGWfC, *e.g.*, Swift/T [Wozniak *et al.* 2013], Panda [Ikeda *et al.* 2013] e SciCumulus [Oliveira *et al.* 2010]. Assim, para gerar as triplas RDF a partir da ontologia, um processo de Extração, Transformação e Carga (ETC) é necessário para extrair os dados da base de proveniência, triplificá-los e carrega-los no repositório de triplas, o qual deve idealmente ser acessível publicamente.

Para facilitar o processo de ETC, foi construída uma interface *Web*, possibilitando a interação entre os dados disponíveis para publicação e o cientista que quer publicar seus resultados na WS. Essa interface permite que o usuário escolha quais dados de um *workflow* serão publicados, tornando viável que informações sigilosas, por exemplo, não sejam publicadas. A interface construída também disponibiliza a visualização de quais dados foram selecionados para publicação. Após finalizar o processo para publicar os dados, uma figura é gerada dinamicamente, contendo os dados instanciados seguindo a ontologia, bem como as classes e suas propriedades de objetos. Dessa forma, a ferramenta possibilita que o cientista visualize a estrutura dos dados publicados, facilitando a análise e a construção, por exemplo, de consultas SPARQL.

4. Estudo de Caso

Para exemplificar o uso da abordagem proposta, utilizamos como estudo de caso os dados de proveniência do *workflow* SciEvol [Ocaña *et al*., 2012] executado com o SGWfC paralelo SciCumulus [Oliveira *et al.* 2010]. O SciEvol é um *workflow* real da área da bioinformática, usado para a análise de Evolução Computacional Molecular. O SciCumulus armazena dados de proveniência em um banco de dados relacional privado.

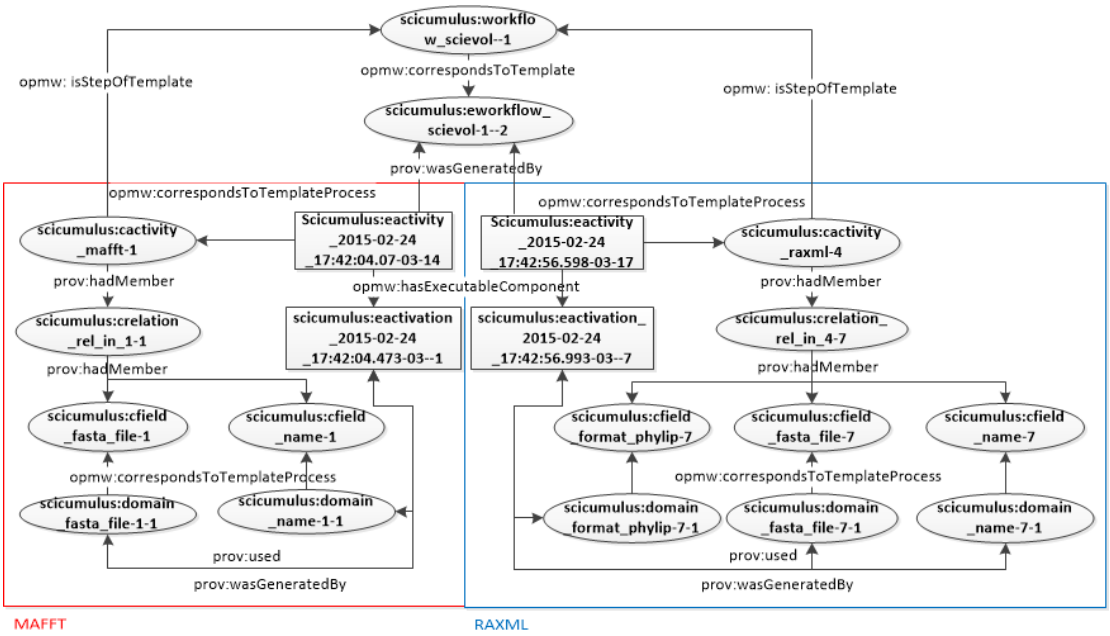

Figura 2. Instância da Ontologia para duas atividades do SciEvol

As treze atividades constituintes do SciEvol possuem dados armazenados em um conjunto de tabelas da base de proveniência do SciCumulus[1]. Para cada uma dessas tabelas, o cientista pode selecionar quais dados serão publicados, garantindo a privacidade dos mesmos. Além das classes, é possível selecionar as propriedades de dados relacionadas a cada classe. Finalmente, é inicializado o processo de ETC para instanciar as triplas no repositório que, neste estudo de caso, foi o Open Link Virtuoso. A Figura 2 apresenta a visualização da instanciação de duas atividades do SciEvol (MAFFT e RAXML) utilizando a ontologia PROV-O-Wf.

5. Conclusão, Trabalhos Relacionados e Trabalhos Futuros

Neste trabalho, propomos uma abordagem para publicação de dados de proveniência de *workflows* científicos na *Web* Semântica. Para tal, desenvolvemos uma ontologia que reutiliza ontologias bem difundidas; considera dados de proveniência em uma granularidade fina; é independente tanto de domínios científicos quanto de SGWfC específicos; e ainda abrange os SGWfC paralelos que manipulam grandes volumes de dados e executam em ambiente de PAD. Adicionalmente, criamos uma interface amigável que facilita a interação do usuário com o sistema e ainda permite que determinados dados gerados na execução da simulação científica sejam escolhidos para publicação na WS. Portanto, acreditamos que este trabalho é inovador devido à sua importância na expansão de um padrão a ser usado para interoperabilidade de dados na WS e contribui com uma área relativamente pouco explorada, principalmente quando consideramos dados de *workflows* executados em ambientes de PAD.

Em relação aos trabalhos relacionados, existem alguns sobre publicação de dados de proveniência de *workflows* usando conceitos da WS. Ding *et al*. [2011] propõem uma abordagem semelhante à apresentada neste artigo. No entanto, não foca em simulações computacionais modeladas como *workflows* e a ontologia é baseada em um padrão de representação de proveniência antigo, que não possui todos os descritores

[1] Modelo de dados do SciCumulus: http://cos.ufrj.br/~silva/relational-database-schema-SCC.png

necessários. Já Chen *et al.* [2006] propõem uma arquitetura para modelagem, coleta de dados de proveniência, porém, assim como o trabalho anterior, não foca em simulações computacionais modeladas como *workflows*, se tornando genérica demais para o cenário tratado neste artigo. Além desses, não foram encontrados trabalhos recentes de SGWfC paralelos que apoiem a publicação de dados de proveniência na WS.

Como trabalhos futuros, apesar de a WS permitir análises integradas entre dados publicados, ainda não exploramos o potencial de interligação entre os dados do estudo de caso publicados e os já existentes publicamente na WS. Assim, desenvolver a ligação dos dados produzidos neste trabalho a dados já publicados na nuvem de *Linked Open Data* (LOD) permitiria explorar ainda mais o potencial das tecnologias da WS.

Agradecimentos

Agradecemos à FAPERJ, ao CNPq e à CAPES por parcialmente apoiar este trabalho.

Referências

Chen, L., Yang, X., Tao, F., (2006), "A Semantic Web Service Based Approach for Augmented Provenance". In: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, p. 594–600, Washington, DC, USA.

Costa, F., Silva, V., Oliveira, D., Ocaña, K., Ogasawara, E., Dias, J. and Mattoso, M. (2013) "Capturing and Querying Workflow Runtime Provenance with PROV: a Practical Approach", In: Proceeding of the Joint EDBT/ICDT 2013 Workshops, New York, USA.

Ding, L., Michaelis, J., McCusker, J., McGuinness, D. L., (2011), "Linked provenance data: A semantic Web-based approach to interoperable workflow traces", Future Generation Computer Systems, v. 27, n. 6, p. 797–805.

Freire, J., Koop, D., Santos, E. and Silva, C. T. (2008) "Provenance for Computational Tasks: A Survey", Computing in Science Engineering, v. 10, n. 3, p. 11-21.

Garijo, D. and Gil, Y. (2014) "The OPMW-PROV Ontology", http://www.opmw.org/model/OPMW/.

Ikeda, R., Das Sarma, A. and Widom, J. (2013) "Logical provenance in data-oriented workflows?", In: Proceedings of the 2013 IEEE International Conference on Data Engineering, p. 877-888.

Jovanovic, J., Siadaty, M., Lages, B., Spors, K. (2012) IntelLEO Workflow Ontology. Disponível em: http://www.intelleo.eu/ontologies/workflow/spec/

Moreau, L., Freire, J., Futrelle, J., McGrath, R. E., Myers, J., Paulson, P. (2008), "The Open Provenance Model: An Overview", In: Freire, J., Koop, D., Moreau, L. (eds), Provenance and Annotation of Data and Processes, Springer Berlin Heidelberg, p. 323–326.

Moreau, L., Missier, P. and Belhajjame, B. (2013) "The PROV Data Model and Abstract Syntax Notation".

Ocaña, K., Oliveira, D., Horta, F., Dias, J., Ogasawara, E and Mattoso, M. (2012) "Exploring molecular evolution reconstruction using a parallel cloud-based scientific workflow", In: Proceedings of the 2012 Brazilian Symposium on Bioinformatics, Berlin, Heidelberg.

Oliveira, D., Ogasawara, E., Baião, F. and Mattoso, M. (2010) "SciCumulus: A Lightweight Cloud Middleware to Explore Many Task Computing Paradigm in Scientific Workflows", In: 3rd International Conference on Cloud Computing, p. 378–385, Washington, DC, USA.

Souza, R., Cottrell, L., White, B., Campos, M. L. and Mattoso, M. (2014) "Linked Open Data Publication Strategies: Application in Networking Performance Measurement Data", In: 2nd ASE International Conference on Big Data Science and Computing, Stanford, CA, USA.

Wolstencroft, K., Alper, P., Hull, D., Wroe, C., Lord, P. W., Stevens, R. D. and Goble, C. A., (2007), "The myGrid ontology: bioinformatics service discovery", Int. J. Bioinformatics Res. Appl., v. 3, n. 3, p. 303–325.

Wozniak, J. M., Armstrong, T. G., Wilde, M., Katz, D. S., Lusk, E. and Foster, I. T., (2013), "Swift/T: Large-Scale Application Composition via Distributed-Memory Dataflow Processing". In: Proceedings of the 13th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), p. 95–102.