Aspectos da Arquitetura e Implementação de um Sistema de Proveniência no Contexto de Workflows Científicos

Claudson Oliveira
Bacharelado em
Ciência da Computação
Univ. Fed. de Juiz de Fora
Juiz de Fora – MG
claudson.oliveira@ice.ufjf.br

Wander Gaspar
Mestrado em
Modelagem Computacional
Univ. Fed. de Juiz de Fora
Juiz de Fora – MG
wander@cesjf.br

Regina Braga
Departamento da
Ciência da Computação
Univ. Fed. de Juiz de Fora
Juiz de Fora – MG
regina.braga@ufjf.edu.br

Jairo Souza
Departamento da
Ciência da Computação
Univ. Fed. de Juiz de Fora
Juiz de Fora – MG
jairo.souza@ufjf.edu.br

ABSTRACT
This paper presents the architecture and implementation of a data provenance system in the field of scientific experiments processed in collaborative research environments. The text highlights the contribution of basic scientific research to the project and tangible benefits for undergraduate students.

Categories and Subject Descriptors
H.4 [**Information Systems Applications**]: General

General Terms
Documentation, Experimentation

Keywords
Workflows Científicos, Proveniência de Dados, Ontologias

1. INTRODUÇÃO
Este trabalho apresenta aspectos da arquitetura e implementação do *Scientific Workflow Provenance System* (SWfPS)[1], um sistema de proveniência de dados no domínio de experimentos científicos processados através de simulações computacionais em ambientes de pesquisa colaborativos interconectados em grade. O texto destaca também a contribuição da iniciação científica para o projeto e os benefícios tangíveis para os alunos de graduação.

[1]O SWfPS encontra-se em desenvolvimento no âmbito do Mestrado em Modelagem Computacional da Universidade Federal de Juiz de Fora (UFJF)

A concepção do SWfPS baseia-se nos requisitos fundamentais: (1) arquitetura independente dos mecanismos de controle de fluxo e formatos de dados utilizados em Sistemas de Gerenciamento de Workflows Científicos (SGWfC); (2) aplicabilidade direcionada a ambientes de execução heterogêneos e dispersos em grades computacionais; e (3) uso de metodologias e tecnologias relevantes no contexto atual no campo da proveniência de dados.

2. WORKFLOWS E PROVENIÊNCIA
Um workflow científico representa um paradigma para a modelagem e gestão de experimentos científicos complexos, cuja implementação computacional facilita a abstração e permite a composição estruturada de programas como uma sequência de atividades [5]. Nesse cenário, um SGWfC refere-se a um conjunto de ferramentas computacionais desenvolvidas para tornar a automação do processo científico mais eficiente e mais produtivo [1].

O problema da proveniência de dados por sua vez, consiste na descrição das origens de um item de dado e do processo pelo qual foi produzido [3]. A proveniência dos dados auxilia a compor uma visão da qualidade, da validade e da atualidade a cerca dos resultados obtidos. No contexto de experimentos científicos modelados computacionalmente, tal visão permite agregar valor de forma significativa no processo de análise dos resultados obtidos pelos cientistas. Porém, para se obter os benefícios advindos a partir da proveniência é necessário modelar, capturar e armazenar as informações para posterior utilização. Moureau e colaboradores [8] propõem um modelo de proveniência padrão, denominado *Open Provenance Model* (OPM), cujo intento é definir uma representação genérica e abrangente para o problema. O objetivo do grupo de pesquisadores é projetar e disponibilizar um padrão de fato para a interoperabilidade dos dados de proveniência em um contexto amplo.

3. ARQUITETURA DO SWFPS
O *Scientific Workflow Provenance System* tem por objetivo capturar e gerenciar os dados de proveniência com base no

padrão OPM em um ambiente de pesquisa científica colaborativo e distribuído. Nesse cenário, o SWfPS deve tratar as informações de proveniência no nível de um experimento científico como um todo. Isso significa que a captura e gestão dos dados devem ser independentes de qualquer tecnologia que provê suporte à execução de workflows. Assim, o SWfPS deve ser capaz de gerir a proveniência de workflows construídos a partir de SGWfCs com suporte à execução de serviços web, tais como Taverna, Kepler e Vistrails. Nesse contexto, o SWfPS fica responsável por instrumentalizar cada processo componente do experimento científico de forma a coletar e persistir os dados de proveniência.

Para a coleta dos dados de proveniência, o SWfPS requer a adaptação dos subworkflows que compõem o experimento. Esse procedimento, denominado instrumentalização, é realizado através do encapsulamento dos serviços web componentes do workflow. Esse mecanismo tem por objetivo capturar as informações de proveniência relevantes para o padrão OPM (dados de entrada e de saída, processos envolvidos, tempo de execução de cada processo, etc.) e enviar esses metadados para um repositório localizado no mesmo nó da grade computacional onde o subworkflow é processado. Essa estratégia para a persistência dos metadados constitui uma solução interessante em um ambiente de pesquisa colaborativo e distribuído, por permitir o aproveitamento dos recursos computacionais disponíveis em cada nó da grade além de contribuir para a otimização da performance da persistência dos dados coletados.

Uma vez que o SWfPS é responsável pelo processo de gerência de proveniência, torna-se possível uma maior homogeneidade no formato e na granularidade das informações armazenadas. Essas características são obtidas a partir de uma adequada instrumentalização dos subworkflows componentes do experimento científico. Acrescenta-se também que em um ambiente colaborativo existe a possibilidade de um subworkflow — composto por um ou mais serviços web proprietários — ser adaptado de forma a permitir a disponibilização e captura de um maior número de metadados de proveniência em um determinado escopo e selecionados de acordo com o objetivo do estudo dos pesquisadores. É importante considerar que o próprio processo de configuração do workflow representa um conjunto de dados de proveniência e deve ser persistido pelo SWfPS.

Um modelo em alto nível de abstração do mecanismo de funcionamento do SWfPS considerando-se um cenário típico de aplicação é apresentado na Figura 1.

3.1 Open Provenance Model

O modelo de proveniência adotado pelo SWfPS baseia-se no padrão *Open Provenance Model*, conforme especificação formulada na versão 1.1 [7]. O OPM fundamenta-se em três pilares: (1) permitir a interoperabilidade entre sistemas de proveniência; (2) representar as informações de proveniência a partir de um modelo independente de tecnologia; (3) permitir aos desenvolvedores construir ferramentas capazes de operacionalizar o uso do modelo conceitual OPM.

Em essência, OPM permite construir um grafo dirigido que expressa os relacionamentos de dependência que originaram um dado específico. O modelo é capaz de representar como

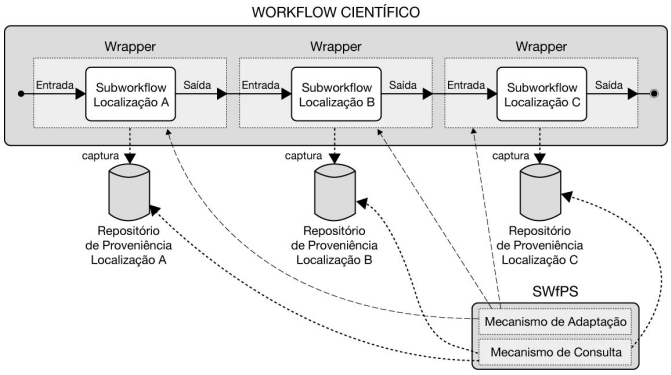

Figure 1: Mecanismo típico de funcionamento do SWfPS

as informações chegaram em um dado momento a um determinado estado e com um conjunto específico de características. OPM possui três entidades principais: (1) artefato – um pedaço de estado imutável, que pode ter uma representação física em um objeto do mundo real ou uma representação digital na computação; (2) processo – ação ou conjunto de ações realizadas em artefatos ou causadas por artefatos que resultam em novos artefatos; (3) agente – entidade contextual que age como um processo catalizador, habilitando, controlando e afetando sua execução.

3.2 Service-Oriented Architecture

O emprego da metodologia *Service-Oriented Architecture* na implementação de workflows científicos processados a partir de SGWfCs apresenta diversos aspectos positivos, como fraco acoplamento entre serviços web, abstração e autonomia, reusabilidade e interoperabilidade [6]. Os mecanismos de instrumentalização providos pelo SWfPS para a captura de proveniência são implementados como serviços web. Esses mecanismos devem ser capazes de capturar de forma consistente as entidades (artefato, processo e agente) e os relacionamentos entre as entidades, conforme definido no OPM. Além disso, outros dados relevantes no contexto do experimento podem ser coletados a partir de adaptação dos próprios serviços web constituintes originais do experimento científico. De forma concreta, a construção dos mecanismos de instrumentação é implementada a partir de web services desenvolvidos segundo o padrão *Simple Object Access Protocol* (SOAP), com o emprego de descritores *Web Services Description Language* (WSDL).

3.3 Base de Dados Relacional

O SWfPS emprega uma base de dados relacional descentralizada para a persistência das informações de proveniência. A adoção do modelo relacional representa uma alternativa conveniente para o tratamento da proveniência de dados e processos no contexto de experimentos científicos [2]. A base de dados implementa tabelas para as entidades principais do OPM — artefato, processo e agente — e para os relacionamentos entre estas entidades. Além disso, é feita a persistência de outros elementos OPM tais como papéis (*roles*), observações e restrições temporais. Concretamente, o SWfPS utiliza o banco de dados relacional Apache Derby 10.6.

3.4 Ontologias e Busca Semântica

O SWfPS deve prover mecanismos de consulta à proveniência armazenada em repositórios distribuídos. Um diferencial do SWfPS consiste em agregar recursos da web semântica na construção do sistema. O objetivo é disponibilizar um ferramental de consulta rico e abrangente, capaz de processar inferências sobre os metadados coletados durante a orquestração e execução dos experimentos científicos. Nesse contexto, inserem-se serviços web, ontologias, máquinas de inferência (*reasoners*) e a linguagem de consulta SPARQL.

O SWfPS manipula um arquivo *Web Ontology Language* (OWL) referente à ontologia dos metadados de proveniência e outro referente à representação do conhecimento no domínio do experimento científico. A ontologia de proveniência fornece um vocabulário controlado para descrever os itens específicos que compõem o padrão OPM. Por outro lado, a ontologia de domínio da aplicação permite uma maior expressividade e possibilidade de realização de inferências nas consultas aos metadados de proveniência.

O SWfPS utiliza SPARQL para a formulação das consultas aos dados de proveniência para obter uma maior expressividade a partir de buscas semânticas com a possibilidade de realização de inferências. Os resultados das consultas SPARQL devem ser convertidos para a linguagem de consulta *Structured Query Language* (SQL), com o objetivo de acessar os dados persistidos em uma base relacional. Um dos aspectos desafiadores da arquitetura consiste na tradução de consultas SPARQL em álgebra relacional e SQL.

3.5 Considerações Adicionais

O SWfPS captura e disponibiliza as informações de proveniência à medida que os dados são processados durante a execução do workflow científico. Essa abordagem, porém, tende a penalizar a performance computacional do experimento. Por exemplo, um subworkflow pode ser executado em múltiplas etapas e repetido diversas vezes. Assim, o volume de dados coletados pode ser elevado [4]. Nesse contexto, torna-se importante monitorar a execução do experimento e avaliar o impacto dos mecanismos de coleta e manipulação da proveniência.

O SWfPS deve prover recursos para o monitoramento em tempo real do processo de captura e armazenamento bem como para visualização, consulta e análise da proveniência persistida. O sistema oferece inicialmente a possibilidade de consultas em SPARQL, o que implica em um alto custo de aprendizagem da sintaxe e semântica da linguagem por parte do cientista. Assim, torna-se relevante avaliar outros mecanismos para a elaboração de consultas a partir de uma solução mais intuitiva.

4. CONTRIBUIÇÃO DA INICIAÇÃO CIENTÍFICA AO PROJETO

O presente trabalho de iniciação científica, desenvolvido no âmbito do curso de graduação em Ciência da Computação da UFJF em parceria com o Mestrado em Modelagem Computacional da universidade, consiste em implementar e refinar componentes de software do *Scientific Workflow Provenance System*. O SWfPS é escrito em Java utilizando-se o ambiente integrado de desenvolvimento Netbeans 6.8.

Considerando-se que o SWfPS encontra-se em desenvolvimento, a contribuição dos alunos de iniciação científica permite a alavancagem do projeto em múltiplas áreas: (1) construção de workflows científicos no Kepler e Taverna adaptados a partir dos mecanismos de instrumentação; (2) desenvolvimento de mecanismos de instrumentalização para serviços web disponibilizados por repositórios de dados científicos; (3) desenvolvimento de mecanismos de instrumentalização para serviços web complexos, baseados em parâmetros construídos em dialetos XML, encontrados em domínios da Bioinformática; (4) avaliação de consistência entre os dados persistidos e o modelo de proveniência OPM; (5) construção e refinamento de *queries* em SPARQL; (6) apoio no desenvolvimento do framework SWfPS, com o uso de métodos e tecnologias como Java, SVN, Glassfish, HTML, CSS, JSP, XML, RDF, OWL, SQL e SPARQL.

O projeto confere ao alunado de iniciação científica a oportunidade de trabalho e aprimoramento técnico em tópicos relevantes e atuais da Ciência da Computação, com destaque para serviços web, ontologias, web semântica, além de prática em Engenharia de Software, Banco de Dados e Programação Orientada a Objetos.

5. REFERENCES

[1] I. Altintas. Lifecycle of scientific workflows and their provenance: A usage perspective. In *SERVICES '08: Proceedings of the 2008 IEEE Congress on Services - Part I*, pages 474–475, Washington, DC, USA, 2008. IEEE Computer Society.

[2] R. S. Barga and L. A. Digiampietri. Automatic capture and efficient storage of e-science experiment provenance. *Concurr. Comput. : Pract. Exper.*, 20(5):419–429, 2008.

[3] P. Buneman, S. Khanna, and W. C. Tan. Why and where: A characterization of data provenance. In *ICDT '01: Proceedings of the 8th International Conference on Database Theory*, pages 316–330, London, UK, 2001. Springer-Verlag.

[4] J. Freire, D. Koop, E. Santos, and C. T. Silva. Provenance for computational tasks: A survey. *Computing in Science and Enginnering*, 10(3):11–21, 2008.

[5] D. Hollingsworth. Lifecycle of scientific workflows and their provenance: A usage perspective. The Workflow Reference Model TC00-1003 Issue 1.1, Workflow Management Coalition, 1995.

[6] C. Lin, S. Lu, Z. Lai, A. Chebotko, X. Fei, J. Hua, and F. Fotouhi. Service-oriented architecture for view: A visual scientific workflow management system. In *SCC '08: Proceedings of the 2008 IEEE International Conference on Services Computing*, pages 335–342, Washington, DC, USA, 2008. IEEE Computer Society.

[7] L. Moreau, B. Clifford, J. Freire, Y. Gil, P. Groth, J. Futrelle, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, Y. Simmhan, E. Stephan, and J. Bussche. The open provenance model — core specification (v1.1). Technical Report 18332, University of Southampton, 2009.

[8] L. Moreau, J. Freire, J. Futrelle, R. E. Mcgrath, J. Myers, and P. Paulson. The open provenance model: An overview. pages 323–326, 2008.