Revista Eletrônica de Sistemas de Informação
ISSN 1677-3071

v. 10, n. 2

2011

Sumário

Revista hospedada em: http://revistas.facecla.com.br/index.php/reinfo
Forma de avaliação: *double blind review*

Esta revista é (e sempre foi) eletrônica para ajudar a proteger o meio ambiente, mas, caso deseje imprimir esse artigo, saiba que ele foi editorado com uma fonte mais ecológica, a *Eco Sans*, que gasta menos tinta.

SWfPS: PROPOSIÇÃO DE UM SISTEMA DE PROVENIÊNCIA DE DADOS E PROCESSOS NO DOMÍNIO DE WORKFLOWS CIENTÍFICOS

SWFPS: PROPOSITION OF A DATA AND PROCESSES PROVENANCE SYSTEM IN THE DOMAIN OF SCIENTIFIC WORKFLOWS

(artigo submetido em outubro de 2010)

Wander Gaspar
Programa Pós-Graduação em Modelagem Computacional – Univ. Federal de Juiz de Fora (UFJF)
wandergaspar@gmail.com

Regina Braga
Departamento de Ciência da Computação – Universidade Federal de Juiz de Fora (UFJF)
regina@ufjf.edu.br

Fernanda Campos
Departamento de Ciência da Computação – Universidade Federal de Juiz de Fora (UFJF)
fernanda@ufjf.edu.br

ABSTRACT

This paper describes SWfPS, an architecture that aims to interact with Scientific Workflow Management Systems in order to capture and manipulate provenance metadata. For this purpose, SWfPS adopts an approach based on an abstract model for representing the lineage. This model, called Open Provenance Model, allows SWfPS to set up a homogeneous and interoperable infrastructure for handling provenance metadata. As a result, SWfPS is able to provide a framework for query metadata provenance generated in an e-Science scenario. Moreover, the architecture uses semantic web technology in order to process provenance queries. In this context, using ontologies and inference engines, SWfPS can make inferences about lineage and, based on those inferences, obtain important results from the extraction of information from the managed data, beyond what is explicitly registered.

Key-words: provenance; semantic web; scientific workflows; e-science

RESUMO

Este artigo descreve o SWfPS, uma arquitetura que visa a interagir com Sistemas de Gerenciamento de Workflows Científicos com o objetivo de capturar e manipular metadados de proveniência. Para esse fim, o SWfPS adota uma abordagem baseada em um modelo abstrato para a representação da proveniência de dados. Este modelo, denominado *Open Provenance Model*, permite que o SWfPS configure uma infraestrutura homogênea e interoperável para a manipulação dos metadados. Como resultado, o SWfPS é capaz de fornecer um arcabouço para a consulta à proveniência gerada em um contexto de e-Ciência. Além disso, a arquitetura utiliza a tecnologia web semântica para processar consultas aos metadados de proveniência. Neste cenário, ao utilizar ontologias e máquinas de inferência, o SWfPS é capaz de promover inferências sobre os metadados coletados e, a partir daí, obter conhecimento adicional com base na extração de informações além daquelas que estão explicitamente registradas nos metadados gerenciados.

Palavras-chave: proveniência; web semântica; workflows científicos; e-ciência

1 INTRODUÇÃO

A e-Ciência se caracteriza pela manipulação de um vasto volume de dados e utilização de recursos computacionais em larga escala, muitas vezes localizados em ambientes distribuídos. Nesse cenário, representado por alta complexidade e heterogeneidade, torna-se relevante o tratamento da proveniência de dados, que tem por objetivo descrever os dados que foram gerados ao longo da execução de um experimento científico e apresentar os processos de transformação a que foram submetidos. Assim, a proveniência auxilia a formar uma visão da qualidade, da validade e da atualidade dos dados produzidos em um ambiente de pesquisa científica.

O *SWfPS* consiste em uma arquitetura cujo objetivo é interagir com Sistemas de Gerenciamento de *Workflows* Científicos (SGWfCs) para promover a captura e a gerência dos metadados de proveniência gerados. Para esse propósito, o *SWfPS* adota uma abordagem baseada em um modelo abstrato para a representação da proveniência. Esse modelo, denominado *Open Provenance Model* (OPM), confere ao *SWfPS* a capacidade de prover uma infraestrutura homogênea e interoperável para a manipulação dos metadados de proveniência. Como resultado, o *SWfPS* permite disponibilizar um arcabouço para a consulta às informações de proveniência geradas em um cenário complexo e diversificado de e-Ciência.

Mais importante, a arquitetura faz uso de tecnologia web semântica para processar as consultas aos metadados de proveniência. Nesse contexto, a partir do emprego de ontologias e máquinas de inferências, o *SWfPS* provê recursos para efetuar deduções sobre os metadados de proveniência e obter resultados importantes ao extrair informações adicionais além daquelas que encontram-se registradas de forma explícita nas informações gerenciadas.

A concepção do modelo apresentado baseia-se em alguns requisitos considerados fundamentais ao delimitar-se o escopo do trabalho pretendido, a saber: (1) arquitetura independente dos mecanismos de controle de fluxo e formatos de dados implementados em SGWfCs; (2) aplicabilidade em uma ampla faixa de experimentos científicos, incluindo ambientes de execução heterogêneos e dispersos em grades computacionais; (3) uso de metodologias e de tecnologias relevantes no contexto atual da proveniência de dados; (4) impacto mínimo na *performance* de execução dos experimentos científicos. Um requisito adicional que o modelo procura atender consiste na adequabilidade do sistema ao usuário, em geral um pesquisador que não possui domínio de ferramentas computacionais complexas. Nesse aspecto, algumas considerações sobre questões como a automatização da captura dos metadados de proveniência e a complexidade no processo de formulação de consultas compõem o escopo do projeto.

O artigo está divido da seguinte forma. A Seção 2 apresenta aspectos importantes no âmbito da proveniência em *workflows* científicos processados através de simulação computacional e que fundamentam o presente trabalho. A Seção 3 apresenta o detalhamento do sistema de proveniência proposto. A Seção 4 discorre sobre alguns trabalhos correla-

tos. A Seção 5 conclui o trabalho ao apresentar algumas considerações adicionais sobre a implementação e ao apontar futuros desenvolvimentos.

2 PROVENIÊNCIA EM WORKFLOWS CIENTÍFICOS

O conceito de *workflow* científico surge como um paradigma para a representação e gestão de experimentos científicos complexos, cuja implementação computacional passa a ser utilizada para facilitar a abstração e permitir uma composição estruturada de programas e *scripts* como uma sequência de atividades que visa a um determinado resultado (ALTINTAS, 2008). Nesse cenário, a ampliação da complexidade na abordagem dos problemas científicos, aliada aos avanços da tecnologia como um todo e da Ciência da Computação em particular, tem resultado no desenvolvimento de SGWfCs, que envolvem um conjunto de ferramentas computacionais desenvolvidas para tornar a automação do processo científico mais eficiente e mais produtivo (MATTOSO *et al.*, 2009).

Um SGWfC permite explorar as representações de processos computacionais complexos em diversos níveis de abstração, com o objetivo de gerenciar o seu ciclo de vida e automatizar sua execução. A automação de *workflows* pode fornecer as informações necessárias para a reprodutibilidade científica e para a derivação e o compartilhamento de resultados em um ambiente de pesquisa colaborativo (GODERIS *et al.*, 2005).

O problema da proveniência de dados foi caracterizado por Buneman, Khanna e Chiew (2001). Para esses autores, a proveniência de dados, também chamada de linhagem, genealogia ou *pedigree*, consiste na descrição das origens de um item de dado e do processo pelo qual foi produzido. A proveniência dos dados auxilia a formar uma visão da qualidade, da validade e de quão recente é a informação. No escopo de *workflows* científicos, a proveniência de dados fornece informação histórica acerca dos dados manipulados a partir de suas fontes originais (SIMMHAN, PLALE e GANNON, 2005). Essa informação pregressa descreve os dados que foram gerados, apresentando os seus processos de transformação a partir de dados primários e intermediários. Nesse cenário, as informações de proveniência podem agregar valor de forma significativa no processo de gerência dos resultados obtidos computacionalmente pelos cientistas. Assim, a gestão da proveniência tem por objetivo servir de auxílio na busca de respostas a inúmeras indagações concernentes a um experimento científico, dentre as quais é possível citar: Que análises estão disponíveis? Como unificar e resumir os conhecimentos gerados? Como consultar uma base de experimentos? E muitas outras questões importantes nesse contexto (MATTOSO *et al.*, 2008).

Para se obter os benefícios advindos a partir das informações de proveniência, torna-se fundamental que tais dados sejam capturados, modelados e armazenados para posterior utilização. O gerenciamento da proveniência de dados é uma questão em aberto e que tem merecido tratamento da comunidade científica (GODERIS *et al.*, 2005; MUNROE *et al.*, 2006; SIMMHAN, PLALE e GANNON, 2006; BITON *et al.,* 2008; FREIRE

et al., 2008). Um dos problemas pesquisados refere-se à falta de concordância quanto à abrangência dos dados a serem capturados além da ausência de uma definição clara de como esse procedimento deve ser realizado (MARINHO, 2009). Moureau *et al.* (2010) propõem o modelo OPM, cujo intento é definir uma representação genérica e abrangente para o tema, além do escopo de *workflows* científicos. Existe um esforço por parte de pesquisadores envolvidos com o problema da proveniência de dados e também de desenvolvedores de sistemas de gerenciamento de *workflows* em aprimorar o OPM[1]. O objetivo é torná-lo um padrão de fato para a troca de informações entre os diversos sistemas já construídos.

É possível traçar um paralelo entre o ciclo de vida de um experimento científico e os respectivos tipos de proveniência tratados em cada um deles (MATTOSO *et al.*, 2009). Esse ciclo de vida é composto por três etapas: composição, execução e análise. A composição consiste na fase responsável pela elaboração dos *workflows* que devem fazer parte do experimento. Os dados de proveniência prospectiva estão associados a essa fase (FREIRE *et al.*, 2008). Nesse contexto, o mecanismo de coleta de proveniência deve ser capaz de capturar o encadeamento do fluxo de serviços e atividades modelados na composição do *workflow* bem como as dependências entre os dados de entrada e de saída entre os diversos processos envolvidos. Durante a fase de execução do *workflow* são coletadas informações relativas à proveniência retrospectiva (FREIRE *et al.*, 2008), que pode ser considerada um registro detalhado do histórico de execução da simulação computacional. Por último, na fase de análise, é possível analisar os resultados obtidos a partir da execução do experimento. Nesse ponto, o cientista pode promover alterações na especificação do *workflow* a partir de consultas às informações de proveniência coletadas, no sentido de refinar ou corrigir a pesquisa em desenvolvimento (CRUZ, CAMPOS e MATTOSO, 2009).

Por fim, deve-se considerar o nível de granularidade, que pode ser entendido como o grau de detalhamento dos dados coletados. Segundo Simmhan, Plale e Gannon (2005), a granularidade dos dados capturados está diretamente relacionada com a utilidade das informações de proveniência. Na literatura, é possível encontrar modelos de proveniência que tratam uma ampla faixa de granularidade (FREIRE *et al.*, 2008).

3 ARQUITETURA DO SWfPS

No domínio de *workflows* científicos, é possível distinguir dois mecanismos de captura de informações de proveniência. Alguns SGWfCs, como Vistrails[2], Taverna[3] e Redux (BARGA E DIGIAMPIETRI, 2008) empregam mecanismos internos para esse propósito. Kepler[4], Pegasus[5] e Karma[6] de-

[1] Provenance Challenge. Disponível em <http://twiki.ipaw.info/bin/view/Challenge/WebHome>.
[2] Vistrails, diponível em <http://www.vistrails.org/index.php/Main_Page>.
[3] Taverna, disponível em <http://www.taverna.org.uk/>.
[4] Kepler, disponível em <https://kepler-project.org/>.
[5] Pegasus, disponível em <http://pegasus.isi.edu/>.

legam essa responsabilidade a serviços externos, que podem ser bastante genéricos e capazes de coletar dados de proveniência em ambientes distribuídos e heterogêneos (CRUZ, CAMPOS e MATTOSO, 2009).

O sistema proposto nesse trabalho, denominado *Scientific Workflow Provenance System* (SWfPS), tem por objetivo prover o tratamento das informações de proveniência no nível de um experimento científico como um todo. Assim sendo, o que se pretende é gerenciar os dados de forma independente de qualquer tecnologia que provê suporte à execução de *workflows*. A defesa desse requisito baseia-se na crescente complexidade dos experimentos científicos (MATTOSO *et al.*, 2009). Nesse cenário, a captura dos dados de proveniência torna-se mais desafiadora nos casos em que *workflows* são executados em configurações nas quais as informações de proveniência requerem a coleta e o armazenamento a partir de fontes distintas (MARINHO, 2009).

Em um cenário típico de aplicação do SWfPS, um *workflow* científico é orquestrado de forma a encadear diversos *subworkflows* executados em grade em um ambiente computacional colaborativo a partir da invocação de serviços web. No contexto atual, em tal cenário, é possível considerar que cada *subworkflow* possa prover a gerência de proveniência de forma descentralizada, em um modelo próprio e com uma determinada granularidade, além de armazenar as informações em um formato específico. Pode-se considerar também que alguns ou mesmo todos os *subworkflows* envolvidos no encadeamento do experimento não disponham de recursos para fornecer o suporte à proveniência. O que se pretende com o SWfPS é prover um mecanismo de gerência de proveniência eficaz, capaz de gerir as informações coletadas em um ambiente computacional colaborativo porém de característica heterogênea.

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. Em um ambiente de pesquisa colaborativo e interconectado por meio de uma grade computacional, um *workflow* científico pode ser composto por diversos *subworkflows* distribuídos, onde a saída de cada um desses módulos do experimento corresponde, normalmente, à entrada do módulo seguinte.

[6] Karma, disponível em <http://www.dataandsearch.org/provenance/?q=taxonomy/term/3>.

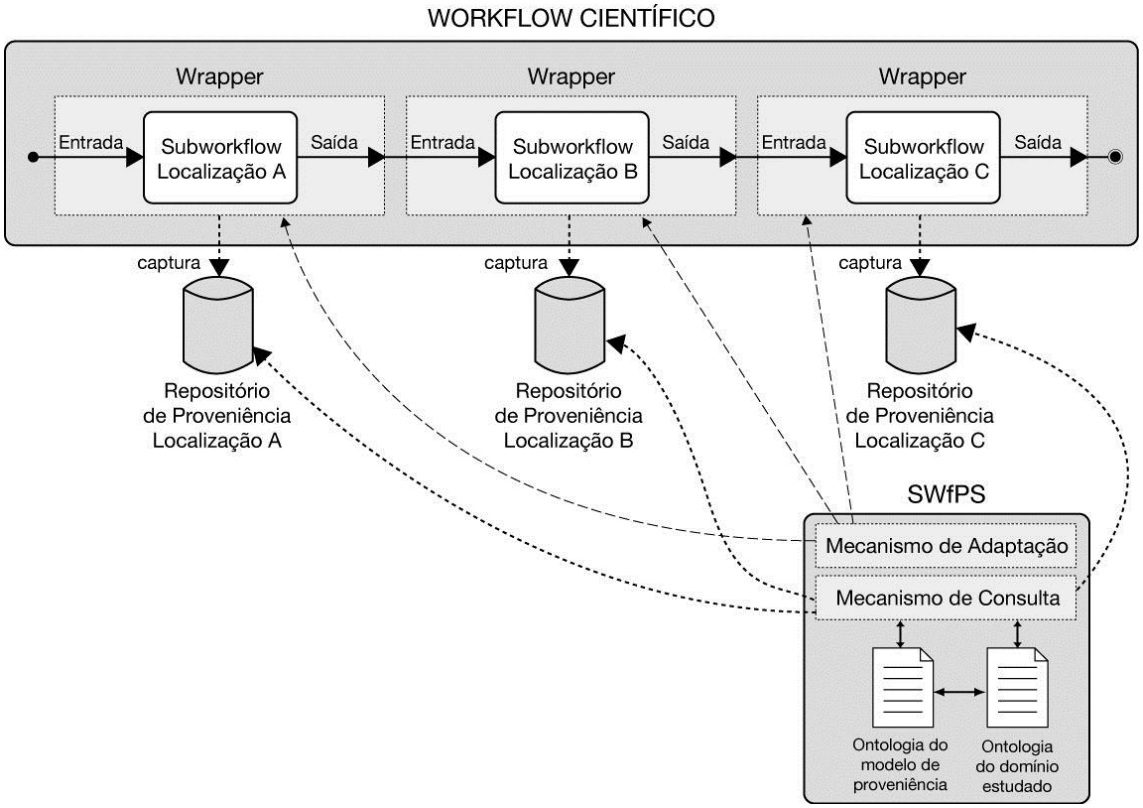

Figura 1. Mecanismo típico de funcionamento do SWfPS
Fonte: os autores.

A primeira responsabilidade do SWfPS consiste em prover e configurar um mecanismo de adaptação para cada *subworkflow* que compõe o experimento. De forma concreta, um invólucro implementado a partir de tecnologia de serviços web deve ser incorporado a cada *subworkflow*. Esse mecanismo tem por objetivo capturar as informações relevantes ao modelo OPM (dados de entrada e de saída, tempo de execução do 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 configura-se viável e constitui uma solução interessante em um ambiente de pesquisa colaborativo disperso geograficamente, porém, interconectado a partir de uma grade computacional de alto desempenho. Além disso, como o SWfPS é responsável pelo processo de gerência de proveniência, torna-se possível uma maior homogeneidade no formato e na granularidade das informações armazenadas.

Um estudo de viabilidade do modelo de arquitetura proposto deve incluir o monitoramento do desempenho de execução do experimento científico. Nesse contexto, é relevante avaliar o impacto dos mecanismos de coleta e manipulação dos metadados de proveniência durante o processo de implementação do SWfPS. Ainda sobre esse aspecto, a opção de projeto por repositórios distribuídos baseia-se na premissa de minimizar custos de performance associados à coleta de metadados de proveniência.

Nessa abordagem, o custo mais significativo pelo acesso remoto aos dados coletados deve restringir-se unicamente à fase de consulta à proveniência.

Acrescenta-se também que em um ambiente de pesquisa colaborativo, cada *subworkflow* pode ser adaptado de forma a permitir a captura de um maior número de metadados de proveniência relevantes em um determinado escopo e selecionados de acordo com o objetivo do estudo dos pesquisadores envolvidos. É importante considerar que o próprio processo de configuração do *workflow* científico se constitui em um conjunto de dados de proveniência OPM e deve ser persistido pelo SWfPS.

O sistema deve prover também mecanismos consistentes de consulta aos metadados de proveniência armazenados nos repositórios distribuídos. Nesse contexto, o diferencial do SWfPS consiste em empregar recursos da web semântica na implementação do modelo de proveniência adotado. O objetivo é disponibilizar aos pesquisadores um ferramental de consulta rico e abrangente, capaz de processar inferências sobre os metadados coletados durante a orquestração e execução do experimento científico.

A arquitetura do modelo proposto para o SWfPS compõe-se de três módulos principais e que se inter-relacionam com o objetivo de prover a implementação de diversos processos. Os módulos projetados são: (1) um gerenciador de ontologias; (2) um gerenciador de persistência de dados; e (3) um gerenciador de adaptação de *subworkflows*. As subseções seguintes procuram descrever de forma mais abrangente os módulos componentes da arquitetura proposta para o sistema.

3.1 GERENCIADOR DE ONTOLOGIAS

Um modelo de proveniência tem por objetivo especificar o conjunto de informações que são suportadas em uma abordagem de proveniência. Além de prover recursos para a representação de dados de proveniência prospectiva e retrospectiva, o modelo pode fornecer suporte a anotações. Tal classe de proveniência tem por objetivo inserir informação adicional que seja relevante para o entendimento do processo de orquestração e execução do *workflow* modelado.

Encontra-se na literatura diversos modelos que propõem soluções para a representação de proveniência (COHEN, BOULAKIA e DAVIDSON, 2006; BARGA e DIGIAMPIETRI, 2008; SCHEIDEGGER *et al.*, 2007; MOUREAU *et al.*, 2010).

O modelo de proveniência adotado pelo SWfPS baseia-se no padrão OPM, conforme especificação formulada na versão 1.1 (MOUREAU *et al.*, 2010). O desenvolvimento do padrão 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.

OPM permite caracterizar as causas que deram origem a uma infor-

mação de proveniência. Em essência, consiste em um grafo dirigido que expressa os relacionamentos de dependência que originaram um dado específico. O objetivo do modelo é ser capaz de representar como as informações chegaram em um dado momento a um determinado estado e com um conjunto específico de características. OPM fundamenta-se em três entidades principais: (1) um artefato – que pode representar desde um objeto do mundo real até um conceito abstrato em qualquer área do conhecimento; (2) um processo – ação ou conjunto de ações realizadas em artefatos ou causadas por artefatos que resultam em novos artefatos; (3) um agente – entidade contextual que age como um catalizador, habilitando, controlando e afetando a execução de um ou mais processos. Pode ser um executor do experimento, um sensor, um controlador etc.

Uma vez que o padrão OPM é independente de tecnologias de implementação, ou seja, constitui-se apenas em um modelo abstrato composto por entidades – como artefatos, processos, agentes – e pelos relacionamentos de causalidade entre estas entidades, torna-se necessário definir uma abordagem concreta para a construção de um protótipo baseado nesse modelo de proveniência, incluindo-se a escolha de requisitos de hardware e software. Segundo Golbeck e Hendler (2008), a web semântica consiste em uma abordagem natural para a proveniência. Nesse contexto, inserem-se serviços web, ontologias, máquinas de inferência e a linguagem de consulta SPARQL[7]. Todo esse ferramental tecnológico tem por objetivo permitir uma representação mais rica e consistente dos metadados de proveniência e, por consequência, disponibilizar ao pesquisador recursos mais sofisticados de consulta às informações coletadas. Também é importante considerar que, em um ambiente de execução em grade computacional, a flexibilidade da web semântica facilita a interoperabilidade ao proporcionar uma melhor integração entre os dados coletados.

O repositório openprovenance.org[8] contém esquemas do OPM em XSD (XML *schema*) e ontologias em *Web Ontology Language*[9] (OWL), além de exemplos de serialização em *Extensible Markup Language*[10] (XML) e *Resource Description Framework*[11] (RDF), que podem ser utilizados para a implementação do modelo de proveniência do SWfPS a partir do padrão OPM. A linguagem de consulta SPARQL foi padronizada pelo *World Wide Web Consortium*[12] (W3C). Considerada fundamental no contexto da tecnologia web semântica, tornou-se oficialmente recomendada pelo consórcio a partir de 2008. Alguns SGWfCs como Pegasus e VIEW (LIN *et al.,* 2008) utilizam um mecanismo de consulta à proveniência baseado na linguagem SPARQL (CRUZ *et al.,* 2009).

Nesse contexto, o emprego da tecnologia web semântica tem por objetivo prover os recursos necessários para a implementação da arqui-

[7] Disponível em http://www.w3.org/TR/rdf-sparql-query/, acesso em 18 fev 2010.
[8] Disponível em <http://openprovenance.org/>, acesso em 18 fev 2010.
[9] Disponível em < http://www.w3.org/2004/owl/>, acesso em 18 fev 2010.
[10] Disponível em < http://www.w3.org/XML/>, acesso em 18 fev 2010.
[11] Disponível em < http://www.w3.org/RDF/>, acesso em 18 fev 2010.
[12] Disponível em < http://www.w3.org/>, acesso em 18 fev 2010.

tetura do SWfPS, atendendo aos requisitos de interoperabilidade de representação do modelo de proveniência em um ambiente de pesquisa colaborativo e interconectado. Em particular, deve-se destacar o emprego da linguagem OWL, em razão do elevado poder de expressividade para representar e instanciar ontologias em ambiente web e a linguagem de consulta SPARQL, padrão do W3C para a formulação das *queries* aos metadados de proveniência armazenados em formatos da web semântica.

O gerenciador de ontologias deve possuir os seguintes componentes: (1) uma ontologia em OWL que descreve o modelo OPM; (2) uma ontologia OWL que represente o conhecimento a cerca do domínio investigado no experimento científico; e (3) um componente de software que permita a consulta em linguagem SPARQL às ontologias. A ontologia de proveniência deve fornecer um vocabulário controlado para descrever os itens específicos que compõem o padrão OPM. Deve descrever, por exemplo, a semântica de termos como serviço, mensagem, função, tempo, algoritmo etc. A ontologia do domínio da aplicação deve descrever o vocabulário próprio do tema em estudo. Em conjunto, estas ontologias têm por objetivo permitir uma maior expressividade nos resultados obtidos em consultas a partir de inferências sobre os metadados de proveniência.

3.2 GERENCIADOR DE PROVENIÊNCIA DE DADOS

Na arquitetura proposta para o SWfPS, a camada do modelo relacional é responsável pela persistência das informações de proveniências coletadas durante a execução do experimento científico. Basicamente, duas soluções podem ser usadas para tratar o problema. Em uma abordagem de armazenamento centralizado, a proveniência é mantida em um repositório local único. A maioria dos SGWfC utiliza essa solução porque facilita a gerência e a segurança, embora exponha os dados a único ponto de falha (FREIRE *et al.,* 2008). Ao contrário, em uma abordagem descentralizada, é empregada uma coleção de repositórios dispersos fisicamente e logicamente interligados por meio de uma grade computacional, onde cada base de dados pode inclusive ser gerida por um sistema de armazenamento diferente. Assim, em uma abordagem descentralizada, os mecanismos utilizados podem ser classificados como homogêneos ou heterogêneos, dependendo da uniformidade dos sistemas utilizados (CRUZ *et al.*, 2009).

Para a persistência dos dados, a maioria dos sistemas de proveniência tem adotado soluções a partir do modelo relacional, embora existam modelos que utilizem outras tecnologias, tais como arquivos ou documentos semiestruturados, estes últimos tendo XML como base (CRUZ *et al.,* 2009). Há ainda casos como o SGWfC Vistrails, que apresenta um mecanismo híbrido, que utiliza uma base de dados relacional para a proveniência retrospectiva e XML para a proveniência prospectiva (FREIRE *et al.,* 2008). Barga e Digiampietri (2008) defendem que o emprego do modelo relacional se apresenta como a alternativa mais adequada para o tratamento da proveniência no contexto de experimentos científicos e baseiam a argumentação nos quesitos confiabilidade e gerência dos dados cole-

tados de forma independente de *workflows* científicos. Alem disso, a opção por uma solução para persistência dos dados coletados a partir de um Sistema de Gerenciamento de Banco de Dados Relacional (SGBDR) deve ser capaz de prover ao SWfPS um eficiente mecanismo de armazenamento e consulta.

A infraestrutura de armazenamento proposta para o SWfPS utiliza um SGBDR disperso em um ambiente de pesquisa colaborativo interconectado por meio de uma grade computacional. Essa estratégia para a persistência das informações de proveniência pode configurar-se viável e, assim, representar uma solução interessante em cenários de cooperação tecnológica. Além disso, como o SWfPS é responsável pelo processo de gerência de proveniência, torna-se possível uma maior homogeneidade no formato e na granularidade das informações armazenadas, mesmo nos casos de experimentos constituídos por *subworkflows* concebidos a partir de ferramentas e técnicas computacionais distintas.

Entre as responsabilidades do gerenciador de persistência de dados, inclui-se o provimento de um mecanismo de mapeamento entre os metadados de proveniência coletados no padrão OPM serializados em linguagem RDF e que devem ser convertidos para tuplas que possam ser armazenados em bases de dados relacionais. Nesse sentido, Chebotko *et al.* (2007) apresentam um estudo que pode servir de base para a implementação de uma solução computacional para a questão.

Uma vez que o modelo propõe a formulação de consultas a partir de SPARQL, as ontologias do modelo de proveniência e da representação do conhecimento do domínio investigado têm por objetivo permitir uma maior expressividade nas buscas semânticas com a realização de inferências sobre os dados armazenados. Este objetivo é factível uma vez que as regras e o formalismo que regulam a combinação entre termos e relacionamentos descritos em uma ontologia permitem não só o compartilhamento e a reutilização do conhecimento, como também explicitar hipóteses e conjecturas, conforme é o interesse do SWfPS no campo da pesquisa científica.

3.3 GERENCIADOR DE ADAPTAÇÃO DE SUBWORKFLOWS

Uma importante consideração sobre a captura das informações de proveniência refere-se ao nível de coleta dos metadados, que pode ser categorizado, segundo Davidson e Freire (2008), em diversos níveis, como de *workflow*, de atividade (ou processo) e do sistema operacional subjacente.

Em um cenário típico de aplicação do SWfPS um *workflow* científico compõe-se do encadeamento de *subworkflows* implementados em centros de pesquisa dispersos geograficamente. Nesse contexto, a arquitetura proposta para o SWfPS fundamenta-se na coleta de proveniência em um nível de *subworkflow*. O gerenciador de adaptação de *subworkflows* deve ter a responsabilidade de prover e configurar um mecanismo de adaptação para cada *subworkflow* que compõe o experimento científico. Esse

mecanismo tem por objetivo capturar informações de proveniência e enviar esses metadados para um repositório localizado no mesmo nó da grade computacional onde o *subworkflow* é processado. Em um ambiente de pesquisa colaborativo, cada *subworkflow* pode ser adaptado, no sentido de prover um maior número de metadados de proveniência relevantes em um determinado escopo e selecionados de acordo com o objetivo do estudo em andamento. Acrescenta-se que o próprio processo de configuração do *workflow* científico se constitui em um conjunto de dados de proveniência a ser persistido.

Uma vez que o projeto de arquitetura do SWfPS refere-se à execução de experimentos científicos em ambientes de pesquisa colaborativos interconectados a partir de uma grade computacional, uma alternativa viável para a comunicação entre o gerenciador de adaptação e os *subworkflows* que compõem o encadeamento consiste no uso da tecnologia de serviços web. Nesse cenário, os mecanismos invólucros devem ser implementados como serviços web com o objetivo de permitir a captura dos metadados de proveniência. Esses mecanismos devem ser capazes de capturar dados como entrada e saída do *subworkflow* e tempo de execução. Além disso, outros dados relevantes no contexto do experimento podem ser coletados a partir de adaptação manual do *subworkflow*. Lin *et al.* (2008) defendem o uso da metodologia *Service-Oriented Architecture* (SOA) em aplicações envolvendo SGWfC por diversas razões, entre elas o fraco acoplamento, abstração e autonomia, reusabilidade e interoperabilidade proporcionados pelos serviços web.

4 TRABALHOS RELACIONADOS

Encontram-se na literatura alguns trabalhos relacionados a sistemas de proveniência no contexto de experimentos científicos. Barga e Digiampietri (2007) apresentam o Redux, uma proposta de representação em camadas para a proveniência em *workflows*, que abrange desde um modelo abstrato até a coleta de dados gerados durante a execução. O *Windows Workflow Foundation*[13] (WinWF) é empregado como motor para a validação do modelo. O modelo de proveniência adotado pelo Redux não se baseia no OPM, padrão para proveniência proposto por Moureau *et al.* (2010) e apoiado por importantes projetos de SGWfC, como Vistrails e Taverna. Redux utiliza um mecanismo de coleta automática e armazena os dados de proveniência em uma base de dados relacional centralizada. As consultas são formuladas em SQL.

Marinho (2009) apresenta o *ProvManager*, um sistema de proveniência independente de SGWfC e com foco em experimentos executados em ambientes distribuídos. *ProvManager* captura os dados de proveniência em nível de atividade a partir de um mecanismo automático de configuração. Uma vez que o modelo atua no nível de atividade, torna-se necessário implementar um mecanismo adaptador para cada SGWfC. O armazena-

[13] Disponível em <http://msdn.microsoft.com/en-us/netframework/aa663328.aspx>.

mento e a consulta à proveniência empregam uma solução de forma centralizada a partir de uma base de dados em Prolog. Embora Prolog possa prover técnicas de inferência sobre os dados coletados, a manipulação de um grande volume de informação em uma base não relacional pode impactar o desempenho das consultas.

SWfPS apresenta como diferencial o foco em ambientes de pesquisa colaborativos interconectados a partir de grades computacionais. Nesse contexto, emprega um modelo de armazenamento descentralizado. Além disso, o modelo baseia-se fortemente em tecnologias promissoras como o padrão OPM e a web semântica. Esses recursos podem prover, respectivamente, uma maior interoperabilidade para os dados coletados e consultas mais sofisticadas a partir de inferências.

5 CONSIDERAÇÕES ADICIONAIS DE PROJETO

A proposta do SWfPS prevê a captura e a disponibilização das informações de proveniência à medida que os dados são processados durante a execução do *workflow* científico. Essa abordagem, porém, tende a impor uma sobrecarga à execução do experimento (BITON *et al.,* 2008). Por exemplo, um *subworkflow* pode ser executado em múltiplas etapas e repetido diversas vezes. Assim, o volume de dados coletados pode ser elevado (FREIRE *et al.,* 2008). Groth *et al.* (2005) afirmam que o nível de degradação no desempenho é tanto maior quanto mais fina for a granularidade das informações coletadas.

Nesse contexto, torna-se importante monitorar o desempenho de execução do experimento científico durante o processo de implementação do SWfPS para avaliar o impacto dos mecanismos de coleta e manipulação dos metadados de proveniência. Uma alternativa para tratar o problema baseia-se no conceito de visões do usuário discutido por Biton *et al.* (2008), que consiste no uso de abstrações que permitam ao cientista definir quais informações obtidas a partir da execução de um *workflow* são relevantes e, a partir daí, estabelecer parâmetros para a coleta seletiva de dados de proveniência com base nos mecanismos adaptadores.

O SWfPS deve prover facilidades ao pesquisador para o monitoramento em tempo real do processo de captura e armazenamento bem como os recursos para visualização, consulta e análise dos dados contidos nos repositórios. Encontram-se na literatura diversos trabalhos que discutem abordagens para o problema da consulta de proveniência, entre eles Scheidegger *et al.* (2007), Golbeck e Hendler (2008), Davidson e Freire (2008) e Holland *et al.* (2008). A arquitetura do SWfPS propõe inicialmente a linguagem de consulta SPARQL para esse fim, embora considere importante avaliar outras alternativas. A adoção de SPARQL implica em um custo de aprendizagem referente à sintaxe e semântica da linguagem de consulta por parte do cientista. Portanto, torna-se importante considerar a possibilidade de implementação futura de uma interface do tipo *query-by-example* (QBE), capaz de prover um mecanismo para a elaboração de

consultas em um ambiente gráfico e mais intuitivo para o usuário, conforme apresentado pelo Vistrails (FREIRE *et al.,* 2008).

Ainda como perspectiva futura vale mencionar o estudo de adequação do SWfPS para o gerenciamento dos repositórios de metadados dispersos em uma grade computacional a partir de algum modelo de mediação, de forma a prover acesso unificado e transparente para o usuário do sistema.

REFERÊNCIAS

ALTINTAS, I. Lifecycle of scientific workflows and their provenance: a usage perspective. In: IEEE Congress on Services - Part I, SERVICES'08, Havaí, EUA. *Anais*... IEEE Computer Society, 2008.

BARGA, R.; DIGIAMPIETRI, L. Automatic capture and efficient storage of e-science experiment provenance. *Concurrency and Computation:* Practice and Experience, v. 20, n. 5, p. 419–429, 2008. doi:http://dx.doi.org/10.1002/cpe.1235

BITON, O.; COHEN-BOULAKIA, S.; DAVIDSON, S.; HARA, C. Querying and managing provenance through user views in scientific workflows. In: International Conference on Data Engineering, ICDE'08, Cancun, México. *Anais*... IEEE Computer Society, 2008.

BUNEMAN, P.; KHANNA, S.; CHIEW, W. Why and where: a characterization of data provenance. In: International Conference on Database Theory, ICDT'01, Londres, Reino Unido*. Anais*... Springer, 2001.

COHEN, S.; BOULAKIA, S.; DAVIDSON, S. Towards a model of provenance and user views in scientific workflows. In: Data Integration in the Life Sciences, DILS'06, Hinxton, Reino Unido. *Anais...* Springer, 2006.

CHEBOTKO, A.; FEI, X.; LIN, C.; LU, S.; FOTOUHI, F. storing and querying scientific workflow provenance metadata using an RDBMS. In: International Conference on e-Science and Grid Computing, E-SCIENCE'07, 3., Bangalore, India. *Anais...* IEEE Computer Society, 2007.

CRUZ, S. M. S.; CAMPOS, M. L. M.; MATTOSO, M. Towards a taxonomy of prove-nance in scientific workflow management systems. In: Congress on Services–I, SERVICES'09, Los Angeles, EUA. *Anais...* IEEE Computer Society, 2009.

DAVIDSON, S.; FREIRE, J. Provenance and scientific workflows: challenges and opportunities. In: International Conference on Management of Data, SIGMOD'08, Vancouver, Canadá. *Anais...* ACM, 2008.

FREIRE, J.; KOOP, D.; SANTOS, E.; SILVA, C. Provenance for computational tasks: a survey. *Computing in Science & Engineering*, v. 10, n. 3, p. 11–21, 2008. doi:http://dx.doi.org/10.1109/MCSE.2008.79

GODERIS, A.; SATTLER, U.; LORD, P.; GOBLE, C. Seven bottlenecks to workflow reuse and repurposing. In: International Web Semantic Conference, ISWC'05, 4., Galway, Irlanda. *Anais...* IDA Ireland, 2005.

GOLBECK, J.; HENDLER, J. A semantic web approach to the provenance challen-ge. *Concurrency and Computation:* Practice and Experience, v. 20, n. 5, p. 431–439, 2008.

GROTH, P.; MILES, S.; MOREAU, L. PReServ: provenance recording for services. 2005. Disponível em: http: //users.ecs.soton.ac.uk/lavm/papers/ Groth-AHM05.pdf. Acesso em: 07/02/2010.

HOLLAND, D.; BRAUN, U.; MACLEAN, D.; MUNISWAMY-REDDY, K.; SELTZER, M. Choosing a data model and query language for provenance. In: International Provenance and Annotation Workshop, IPAW'08, 2., Salt Lake City, EUA. *Anais...* Springer, 2008.

LIN, C.; LU, S.; LAI, Z.; CHEBOTKO, A.; FEI, X.; HUA, J.; FOTOUHI, F. Service-oriented architecture for VIEW: A Visual Scientific Workflow Management System. In: International Conference on Services Computing, SERVICES'08, Havaí, EUA. *Anais...* IEEE Computer Society, 2008.

MARINHO, A. ProvManager: uma abordagem para gerenciamento de proveniência de workflows científicos. In: Workshop de Teses e Dissertações em Engenharia de Software, XXIII SBES, 14., Fortaleza, CE. *Anais*... SBC, 2009.

MATTOSO, M.; WERNER, C.; TRAVASSOS, G.; BRAGANHOLO, V.; MURTA, L. Gerenciando experimentos científicos em larga escala. In: Seminário Integrado de Software e Hardware, SEMISH'08, 28., Belém, PA. *Anais...* SBC, 2008.

MATTOSO, M.; WERNER, C.; TRAVASSOS, G.; BRAGANHOLO, V.; MURTA, L.; OGASAWARA, E.; OLIVEIRA, F.; MARTINHO, W. Desafios no apoio à composição de experimentos científicos em larga escala. In: Seminário Integrado de Software e Hardware, SEMISH'09, 36., Bento Gonçalves, RS. *Anais...* SBC, 2009.

MOUREAU, L.; CLIFFORD, B.; FREIRE, J.; GIL, Y.; GROTH, P.; FUTRELLE, J.; KWASNIKOWSKA, N.; MILES, S.; MISSIER, P.; MYERS, J.; SIMMHAN, Y.; STEPHAN, E.; BUSSCHE, J. The open provenance model core specification v1.1. *Future Generation Computer Systems*, *in press*, 2010.

MUNROE, S.; MILES, S.; MOREAU, L.; VÁZQUEZ-SALCEDA, J. PrIMe: a software engineering methodology for developing provenance-aware applications. In: International Workshop on Software Engineering and Middleware, SEM'06, 6., Portland, EUA. *Anais...* ACM, 2006.

SCHEIDEGGER, C.; KOOP, D.; SANTOS, E.; VO, H.; CALLAHAN, S.; FREIRE, J.; SILVA, C. Tackling the provenance challenge one layer at a time. *Concurrency and Computation:* Practice and Experience, v. 20, n. 5, p. 473–483, 2007. doi:http://dx.doi.org/10.1002/cpe.1237

SIMMHAN, Y.; PLALE, B.; GANNON, D. A survey of data provenance in e-science. *ACM SIGMOD Record*, v. 34, n. 3, p. 31–36, 2005. doi:http://dx.doi.org/10.1145/1084805.1084812

SIMMHAN, Y.; PLALE, B.; GANNON, D. A framework for collecting provenance in data-centric scientific workflows. In: International Conference on Web Services, ICWS'06, Chicago, EUA. *Anais...* IEEE Computer Society, 2006.