SWfPS: Um Sistema de Proveniência de Dados e Processos no Domínio de Workflows Científicos

Wander A. Gaspar[13], Regina M. Braga[12], Fernanda A. Campos[12]

[1]Mestrado em Modelagem Computacional
[2]Núcleo de Pesquisa em Qualidade de Software
Universidade Federal de Juiz de Fora
Campus Universitário – Juiz de Fora, MG – Brasil

[3]Centro de Ensino Superior de Juiz de Fora (CES/JF)
Campus Academia – Juiz de Fora, MG – Brasil

wander@cesjf.br,{regina,fernanda}@ufjf.edu.br

Abstract*. This paper presents an architecture for a provenance considering the domain of scientific experiments performed by computer simulation. The objective of the model is to capture and manage information from collaborative research environments interconnected by a grid computing. Furthermore, the model must operate independently of any technology that provides support for workflow execution. The proposed architecture is based on standard features such as OPM, web services and ontologies.*

Resumo. *Este artigo apresenta uma proposta de arquitetura para um sistema de proveniência no domínio dos experimentos científicos executados através de simulação por computador. O objetivo do modelo projetado consiste em capturar e gerenciar as informações de proveniência em ambientes colaborativos de pesquisa interconectados através de uma grade computacional. Além disso, o modelo deve atuar de forma independente de qualquer tecnologia que provê suporte à execução de workflows. A arquitetura proposta baseia-se em recursos como o padrão OPM, serviços web e ontologias.*

1. Introdução

O objetivo desse trabalho é apresentar uma proposta de sistema de proveniência no domínio de experimentos científicos processados através de simulações em ambientes de pesquisa colaborativos dispersos geograficamente interconectados através de uma grade computacional. O tipo de sistema proposto pode ser entendido como uma ferramenta de *software* capaz de interagir com Sistemas de Gerenciamento de *Workflows* Científicos (SGWfC) com a finalidade de tratar as informações de proveniência geradas a partir dos *workflows* científicos.

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 quaisquer SGWfC; (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 experi-

mentos 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 correlatos. A Seção 5 conclui o trabalho ao apresentar algumas considerações adicionais sobre a implementação e a 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 com o objetivo de facilitar a abstração e permitir uma composição estruturada de programas e *scripts* como uma sequência de atividades que visa um determinado resultado (Hollingsworth 1995). Nesse cenário, a ampliação da complexidade na abordagem dos problemas científicos aliado aos avanços da tecnologia como um todo e da Ciência da Computação em particular têm resultado no desenvolvimento de SGWfC, que se constituem em um conjunto de ferramentas computacionais desenvolvidas para tornar a automação do processo científico mais eficiente e mais produtivo (Altintas 2008).

Um SGWfC permite explorar as representações de processos computacionais complexos em diversos níveis de abstração, com o objetivo de gerenciar o ciclo de vida e automatizar a 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 *et al.* (2001). Para o autor, 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 et al. 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, Mun-

roe et al. 2006, Simmhan et al. 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.* (2007) e colaboradores propõem um modelo de proveniência padrão, denominado *Open Provenance Model* (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 um grupo 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 *log* 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 et al. 2009).

Por fim, deve-se considerar o nível de granularidade, que pode ser entendido como o grau de detalhamento dos dados coletados. Segundo Symmhan et al. (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 SGWfC, como Vistrails[2], Taverna[3] e Redux (Barga e Digiampietri 2007) empregam mecanismos internos para esse propósito. Kepler[4], Pegasus[5] e Karma[6] delegam essa responsabilidade a serviços externos, que podem ser bastante genéricos e capazes de coletar dados de proveniência em ambientes distribuídos e heterogêneos (Cruz et al. 2009).

O sistema proposto nesse trabalho, denominado "*Scientific Workflow Provenance System"* (SWfPS), tem por objetivo prover o tratamento das informações de proveni-

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

ência 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 et al. 2009).

Em um cenário típico de aplicação do SWfPS, um *workflow* científico é orquestrado de forma a encadear diversos *subworkflows* executados em *grid* 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 através 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, na maioria dos casos, à entrada do módulo seguinte.

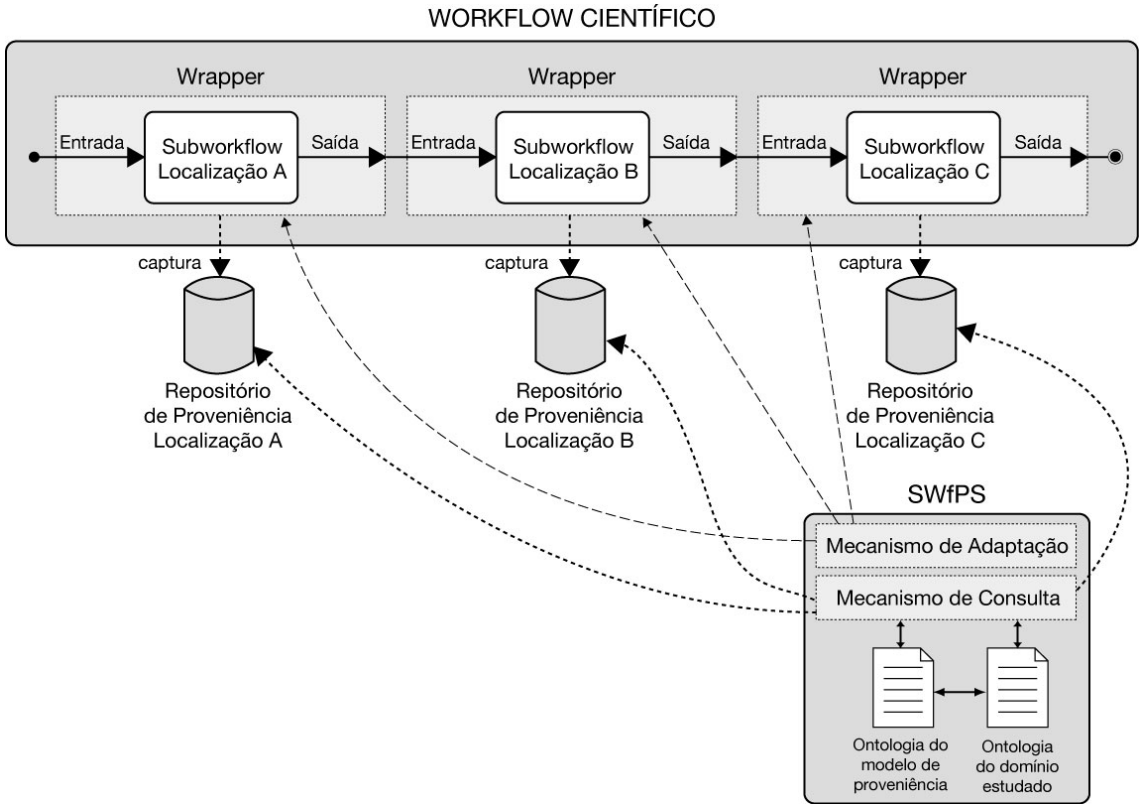

Figura 1. Mecanismo típico de funcionamento do SWfPS

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 *wrapper* 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 do *grid* 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, a fim de se obter mais semântica a partir das informações coletadas. Diversos modelos que propõem soluções para a representação de proveniên-

cia são encontrados na literatura (Moureau e Ibbotson 2006, Cohen et al. 2006, Barga e Digiampietri 2007, Scheidegger et al. 2007, Moureau et al. 2007).

O modelo de proveniência adotado pelo SWfPS baseia-se no padrão OPM, conforme especificação formulada na versão 1.1, publicada em dezembro de 2009 (Moureau et al. 2009). 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 (Moureau et al. 2007).

OPM permite caracterizar as causas que deram origem a uma informação de proveniência. Em essência, consiste em um grafo dirigido que expressa os relacionamento 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) 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.

Uma vez que o padrão *Open Provenance Model* é independente de tecnologias de implementação, torna-se necessário definir uma abordagem para a construção de um protótipo baseado nesse modelo de proveniência. Segundo Golbeck e Hendler (2007), a *web* semântica consiste em uma abordagem natural para a proveniência. Nesse contexto, inserem-se serviços *web*, ontologias, máquinas de inferência (*reasoners*) 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 SGWfC como Pegasus e VIEW (Lin et al. 2008) utilizam um mecanismo de consulta à proveniência baseado na linguagem SPARQL (Cruz et al. 2009).

---

[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.

Nesse contexto, o emprego da tecnologia *web* semântica tem por objetivo prover os recursos necessários para que a arquitetura do SWfPS atenda ao requisito de interoperabilidade de representação do modelo de proveniência em um ambiente de pesquisa colaborativo interconectado através de uma grade computacional. Em particular, OWL deve ser usada para expressar as ontologias definidas no sistema, RDF deve ser empregada para serializar os metadados de proveniência e a linguagem SPARQL deve ser usada para a formulação das consultas.

O gerenciador de ontologias deve possuir os seguintes componentes: (1) um arquivo OWL referente à ontologia de metadados de proveniência; (2) um arquivo OWL contendo a representação do conhecimento referente ao domínio investigado no experimento científico em estudo; e (3) um componente de *software* capaz de administrar e permitir a consulta em linguagem SPARQL a esses arquivos. A ontologia de proveniência deve fornecer um vocabulário controlado para descrever os itens específicos que compõem a documentação de um processo segundo o padrão OPM. Deve descrever, por exemplo, a semântica de termos como serviço, mensagem, função, tempo, algoritmo etc., conforme apresentados na Figura 2. Por outro lado, a ontologia no domínio da aplicação em estudo deve permitir uma maior expressividade e possibilidade de realização de inferências nas consultas aos metadados de proveniência.

3.2. Gerenciador de Persistência 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 através 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 semi-estruturados, 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 (2007) 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 coletados de forma independente de *workflows* científicos. Alem disso, a opção por uma solução para persistência dos dados coletados a partir de um SGBDR deve ser capaz de prover ao SWfPS um eficiente mecanismo de armazenamento e consulta.

A infraestrutura de armazenamento proposta para o SWfPS utiliza uma Sistema de Gerenciamento de Banco de Dados Relacional (SGBDR) disperso em um ambiente de pesquisa colaborativo interconectado através de um *grid* computacional. Essa estra-

tégia 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 de proveniência pelo usuário do sistema a partir da linguagem SPARQL, as ontologias de metadados de proveniência e de representação do conhecimento referente ao domínio investigado devem permitir uma maior expressividade a partir de buscas semânticas com a possibilidade de realização de inferências. Esse recurso se mostra interessante e constitui em um dos diferenciais do projeto proposto para o SWfPS. Os resultados das consultas SPARQL devem ainda ser convertidos para a linguagem de consulta *Structured Query Language* (SQL), com o objetivo de acessar os dados persistidos em uma base relacional. É possível encontrar na literatura científica diversos trabalhos que propõem a utilização de um SGBDR para armazenar e consultar dados RDF usando o SQL e linguagens de consulta SPARQL. Um dos problemas mais desafiadores de tal abordagem é a tradução de consultas SPARQL em álgebra relacional e SQL. Chebotko *et al* (2007) apresentam um algoritmo denominado SPARQLtoSQL que permite a conversão de um grafo padrão SPARQL em uma consulta SQL equivalente.

Outra atribuição do gerenciador de persistência é controlar o acesso aos repositórios dispersos no ambiente colaborativo interconectado, uma vez que as consultas de proveniência podem envolver dados de uma ou mais fontes. Futuramente, espera-se a adequação do SWfPS para permitir o gerenciamento dos repositórios dispersos de metadados a partir de algum modelo de mediação, de forma a prover acesso unificado e transparente para o usuário do sistema.

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 de 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 do *grid* computacional onde o *subwork-*

flow é processado. Em um ambiente de pesquisa colaborativo, cada *subworkflow* pode ser adaptado, no sentido de prover um maior numero 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 a 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 *web services*. Nesse cenário, os mecanismos *wrapper* 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) defende 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 (*engine*) 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.* (2007) e apoiado por importantes projetos de SGWfC, como Vistrails e Taverna (Moureau et al., 2009). 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 *et al.* (2009) apresentam 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 armazenamento 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 *grids* computacionais. Nesse contexto, emprega um modelo de armazenamento descentralizado. Além disso, o modelo baseia-se fortemente em tecnologias promissoras como o padrão OPM e ontologias. Esses recursos podem pro-

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

ver, 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 da performance 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 (*user views*) 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 a partir dos 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), Davison 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. (2008) "Lifecycle of Scientific Workflows and their Provenance: A Usage Perspective". In: SERVICES '08: Proceedings of the 2008 IEEE Congress on Services - Part I, pp. 474–475, IEEE Computer Society, doi: http://dx.doi.org/10.1109/SERVICES-1.2008.87.

Barga, R. e Digiampietri, L. (2007) "Automatic Capture and Efficient Storage of e-Science Experiment Provenance", Concurrency and Computation: Practice and Experience, v.20, n.5, p.419–429, John Wiley & Sons, doi: 10.1002/cpe.1235.

Biton, O., Cohen-Boulakia, S., Davidson, S. e Hara, C. (2008) "Querying and Managing Provenance through User Views in Scientific Workflows", In: ICDE'08: 24th International Conference on Data Engineering, IEEE Computer Society, Washington, EUA, p.1072–1081, doi: http://dx.doi.org/10.1109/ICDE.2008.4497516.

Buneman, P., Khanna, S. e Chiew, W. (2001) "Why and Where: a Characterization of Data Provenance. In: ICDT'01: 8th International Conference on Database Theory, LNCS, v.1973, p.316–330, Springer, doi: 10.1007/3-540-44503-X_20.

Cohen, S., Boulakia, S. e Davidson, S. (2006) "Towards a Model of Provenance and User Views in Scientific Workflows", Data Integration in the Life Sciences, LNCS 4075, Springer, p.264–279, doi: http://dx.doi.org/10.1007/11799511.

Chebotko, A., Fei, X., Lin, C., Lu, S. e Fotouhi, F. (2007) "Storing and Querying Scientific Workflow Provenance Metadata Using an RDBMS", In: E-SCIENCE'07: 3rd International Conference on E-Science and Grid Computing. IEEE Computer Society, Washington, EUA, doi: http://dx.doi.org/10.1109/E-SCIENCE.2007.70.

Cruz, S. M. S., Campos, M. L. M. e Mattoso, M. (2009) "Towards a Taxonomy of Provenance in Scientific Workflow Management Systems", SERVICES'09: Congress on Services–I, p.259–266, IEEE Computer Society, Washington, EUA, doi: http://dx.doi.org/10.1109/SERVICES-I.2009.18.

Davidson, S. e Freire, J. (2008) "Provenance and Scientific Workflows: Challenges and Opportunities. SIGMOD'08: International Conference on Management of Data, Vancouver, Canadá, p.1345-1350, doi: http://doi.acm.org/10.1145/1376616. 1376772.

Freire, J., Koop, D., Santos, E. e Silva, C. (2008) "Provenance for Computational Tasks: a Survey", Computing in Science & Engineering, v.10, n.3, p.11–21, doi: http://doi.ieeecomputersociety.org/10.1109/MCSE.2008.79.

Goderis, A., Sattler, U., Lord, P. e Goble, C. (2005) "Seven Bottlenecks to Workflow Reuse and Repurposing", ISWC'05: 4th International Web Semantic Conference. LNCS, v.3792, p.323–337, doi: 10.1007/11574620_25.

Golbeck, J. e Hendler, J. (2007) "A Semantic Web Approach to the Provenance Challenge", Concurrency and Computation: Practice and Experience, v.20, n.5, p.431–439.

Groth, P., Miles, S. e Moreau, L. (2005) "PReServ: Provenance Recording for Services", UK OST e-Science Second All Hands Meeting 2005, disponível em <http://users.ecs.soton.ac.uk/lavm/papers/Groth-AHM05.pdf>, acesso em 07 fev 2010.

Holland, D., Braun, U., Maclean, D., Muniswamy-Reddy, K. e Seltzer, M. (2008) "Choosing a Data Model and Query Language for Provenance", IPAW'08: Second International Provenance and Annotation Workshop, disponível em <www.eecs. harvard.edu/~kiran/pubs/ipaw08.pdf>, acesso em 16 fev 2010.

Hollingsworth, D. (1995) "Lifecycle of Scientific Workflows and their Provenance: A Usage Perspective", The Workflow Reference Model TC00-1003 Issue 1.1, Workflow Management Coalition, 1995.

Lin, C., Lu, S., Lai, Z., Chebotko, A., Fei, X., Hua, J. e Fotouhi, F. (2008) "Service-Oriented Architecture for VIEW: A Visual Scientific Workflow Management Sys-

tem", In: SERVICES'08: International Conference on Services Computing, v.1, IEEE Computer Society, Washington, EUA, doi: http://dx.doi.org/10.1109/SCC.2008.118.

Marinho, A. (2009) "ProvManager: uma Abordagem para Gerenciamento de Proveniência de Workflows Científicos", 14º Workshop de Teses e Dissertações em Engenharia de Software, In: XXIII SBES, Fortaleza, CE, Brasil, disponível em <http:// www.lbd.dcc.ufmg.br:8080/colecoes/wtes/2009/005.pdf>, acesso em 05 fev 2010.

Mattoso, M., Werner, C., Travassos, G., Braganholo, V. e Murta, L. (2008) "Gerenciando Experimentos Científicos em Larga Escala", In: SEMISH'08: XXVIII Seminário Integrado de Software e Hardware, Belém, PA, Brasil, disponível em <http://www.lbd.dcc.ufmg.br:8080/colecoes/semish/2008/009.pdf>, acesso em 09 fev 2010.

Mattoso, M., Werner, C., Travassos, G., Braganholo, V., Murta, L., Ogasawara, E., Oliveira, F. e Martinho, W. (2009) "Desafios no Apoio à Composição de Experimentos Científicos em Larga Escala", In: SEMISH'09: XXXVI Seminário Integrado de Software e Hardware, Bento Gonçalves, RS, Brasil, disponível em <www.sbc.org.br/bibliotecadigital/download.php?paper=1313>, acesso em 06 fev 2010.

Moureau, L. e Ibbotson, J. (2006) "Standardisation of Provenance Systems in Service Oriented Architectures", White Paper, University of Southampton, disponível em <http://eprints.ecs.soton.ac.uk/12198/1/WhitePaper.pdf>, acesso em 11 fev 2010.

Moureau, L., Freire, J., Futrelle, J., Mcgrath, R., Myers, J. e Paulson, P. (2007) "The Open Provenance Model", Technical Report, , University of Southampton, disponível em <http://eprints.ecs.soton.ac.uk/14979/1/opm.pdf>, acesso em 11 fev 2010.

Moureau, L., Clifford, B., Freire, J., Gil, Y., Groth, P., Futrelle, J., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Simmhan, Y., Stephan, E. e Bussche, J. (2009) "The Open Provenance Model Core Specification v1.1", Technical Report, University of Southampton, disponível em <http://eprints.ecs.soton.ac.uk/18332/1/opm.pdf>, acesso em 11 fev 2010.

Munroe, S., Miles, S., Moreau, L. e Vázquez-Salceda, J. (2006) "PrIMe: a Software Engineering Methodology for Developing Provenance-aware Applications", In: SEM'06: 6th International Workshop on Software Engineering and Middleware, p.39–46, Portland, EUA, doi: http://doi.acm.org/10.1145/1210525.1210535.

Scheidegger, C., Koop, D., Santos, E., Vo, H., Callahan, S., Freire, J. e Silva, C. (2007) "Tackling the Provenance Challenge one Layer at a Time", Concurrency and Computation: Practice and Experience, v.20, n.5, p.473–483, doi: 10.1002/cpe.1237.

Simmhan, Y., Plale, B. e Gannon, D. (2005) "A Survey of Data Provenance in e-Science", SIGMOD Record, v.34, n.3, p.31–36, doi: http://doi.acm. org/10.1145/1084805.1084812, ACM, Nova York, EUA.

Simmhan, Y., Plale, B. e Gannon, D. (2006) "A Framework for Collecting Provenance in Data-centric Scientific Workflows", In: ICWS'06: International Conference on Web Services, p.427–436, IEEE Computer Society, Washington, EUA, doi: http://dx.doi.org/10.1109/ ICWS.2006.5.