SciProvMiner: Arquitetura para Captura de Proveniência de dados de Workflows Científicos utilizando o Modelo OPM

Tatiane O. M. Alves Regina M. M. Braga
Departamento de Ciência da Computação
Universidade Federal de Juiz de Fora – Juiz de Fora, MG – Brazil
Área de Pesquisa: Engenharia de Software/Banco de Dados

tatiane.ornelas@ice.ufjf.br, regina.braga@acessa.com

Resumo. *Prover informação histórica de experimentos científicos, com o objetivo de tratar o problema da perda de conhecimento do cientista sobre o experimento tem sido o foco de diversas pesquisas. Este trabalho apresenta a arquitetura SciProvMiner, que tem como objetivo principal a coleta e consulta de proveniência prospectiva e retrospectiva de experimentos científicos fazendo uso da ontologia OPMO e de técnicas de data mining. Com a abordagem, pretende-se fornecer ao cientista funcionalidades que possam auxiliá-lo a tomada de decisão com base na proveniência de dados de workflows científicos.*

1. Introdução

Computação em larga escala tem sido amplamente utilizada como metodologia para a realização de pesquisa científica: existem vários casos de sucesso em muitos domínios, incluindo a física, bioinformática, engenharia e ciências geográficas [Wong et al. 2005]. Essas computações em larga escala, que sustentam um processo científico, são geralmente referidas como e-Science [Wong et al. 2005]. Segundo [Mattoso et al. 2008], o termo e-ciência ou e-science significa o apoio ao cientista para o desenvolvimento de ciência em larga escala utilizando infra-estrutura computacional.

Este trabalho tem como objetivo principal especificar uma arquitetura para a coleta, gerência e consulta da proveniência de dados e processos no contexto de experimentos científicos. A arquitetura proposta deve prover uma camada de interoperabilidade capaz de interagir com os SGWfC (Sistemas de Gerenciamento de Workflows Científicos) tendo como finalidade capturar as informações de proveniência geradas a partir de workflows científicos e prover ao cientista meios de consulta a esses dados que lhe traga um maior conhecimento sobre o experimento realizado, através da utilização das regras de completude e inferência definidas no modelo OPM, exploradas a partir da aplicação da ontologia OPMO e de técnicas de mineração de dados nas informações de proveniência capturadas.

O restante deste artigo é estruturado da seguinte forma: Na seção 2, são apresentados os pressupostos teóricos que sustentam este trabalho. Na seção 3 são descritos alguns trabalhos relacionados. A seção 4 apresenta a arquitetura do SciprovMiner, mostrando suas camadas e características. Finalmente, a seção 5 apresenta as considerações finais.

2. Pressupostos Teóricos

Em se tratando de experimentação científica, proveniência de dados pode ser definida como informação que auxilia a determinar a derivação histórica do produto de dado, a partir de suas fontes de origem, sendo considerado um componente essencial para permitir reprodutibilidade do resultado, compartilhamento e reuso de conhecimento pela comunidade científica [Freire et al. 2008]. Para obter benefícios de informações de proveniência, estas têm que ser capturadas, modeladas e armazenadas para futuras consultas. Gerenciamento de proveniência é uma questão em aberto que está sendo abordada por vários SGWfCs (Sistemas Gerenciadores de Workflow Científico) e pela série Provenance Challenge [Moreau et al. 2011]. Um dos problemas em aberto na área de gerência de proveniência está relacionado com quais dados de proveniência devem ser coletados e como. Segundo Freire et al. (2008) existem diversos modelos de proveniência propostos por pesquisadores na literatura. Todos estes modelos suportam alguma forma de proveniência retrospectiva, que captura os passos que foram executados, bem como as informações sobre os ambientes de execução utilizados para obter um produto de dado específico, e a maioria dos SGWfCs fornecem meios para capturar proveniência prospectiva, que captura a especificação de uma tarefa computacional (ou seja, um workflow), correspondendo aos passos que precisam ser seguidos para gerar um produto de dados ou classe de produtos de dados [Lim et al. 2010].

Porém, essa diversificação de modelos prejudica a tarefa de cientistas que manipulam vários sistemas de gerenciamento de proveniência bem como a comunicação com sistemas com os quais queiram trocar informações de proveniência [Marinho 2011]. Devido a essa preocupação em se prover interoperabilidade de sistemas por meio de troca de proveniência, há a iniciativa de promover um modelo padrão, o *Open Provenance Model* (OPM) [Moreau et al. 2011], que define uma representação genérica para dados de proveniência. No modelo OPM supõe-se que a proveniência dos objetos (digitais ou não) pode ser representada por um grafo de causalidade anotada, que é um grafo acíclico dirigido, enriquecido com anotações capturadas de outras informações relativas à execução [Moreau et al. 2011].

No modelo OPM, grafos de proveniência são composto por três tipos de nós [Moreau et al. 2011]: Artefatos: representam um dado de estado imutável, que pode ter um corpo físico em um objeto físico, ou uma representação digital em um sistema de computador; Processos: representam ações realizadas ou causadas por artefatos, e resultam em novos artefatos.; Agentes: representam entidades contextuais agindo como um catalisador de um processo, permitindo, facilitando, controlar, ou afetando sua execução.

Alguns modelos de proveniência usam a tecnologia da Web Semântica tanto para representar quanto para consultar informações de proveniência. Em Moreau et al. [2011] é definida uma ontologia OWL para capturar os conceitos do OPM na versão 1.1 e as inferências válidas neste modelo. Além disso, esta ontologia OWL especifica uma serialização RDF da versão 1.1 do modelo abstrato OPM. A essa ontologia é dado o nome de Open Provenance Model Ontology (OPMO).

3. Trabalhos Relacionados

Recentemente, diversas pesquisas têm sido feitas no intuito de solucionar desafios em proveniência de dados no domínio científico. Em Marinho (2011), o autor propõe uma arquitetura para gerência de proveniência de dados denominada ProvManager. Tanto a arquitetura do SciProvMiner quanto a do ProvManager têm como foco a captura da proveniência em workflows orquestrados a partir de recursos heterogêneos e geograficamente distribuídos baseados na invocação de serviços web. Porém o SciProvMiner provê uma infraestrutura baseada na web semântica para representação e consulta aos metadados de proveniência, o que lhe confere um maior poder de expressividade para inferência de conhecimento novo. Por outro lado, o ProvManager estabelece um mecanismo de coleta automatizada de proveniência, enquanto o SciProvMiner não trabalha esse requisito. O ProvManager coleta tanto proveniência prospectiva quanto retrospectiva. O SciProvMiner coleta também a proveniência prospectiva e retrospectiva, com o diferencial de se propor a utilização do OPM para a captura tanto da proveniência retrospectiva (inerente ao modelo) quanto da proveniência prospectiva, a partir de uma extensão do modelo OPM. A vantagem de tal abordagem é a interoperabilidade dos dados capturados, uma vez que se está utilizando um modelo padrão. Considerando a camada de mineração de dados do SciProvMiner, o ProvManager não possui.

Em Lim et al. (2010) é proposta uma extensão do modelo OPM para modelar proveniência prospectiva, além da retrospectiva já suportada no OPM nativo e um framework para coletar ambas, proveniência prospectiva e retrospectiva. O SciProvMiner também utiliza uma extensão do OPM para suportar a modelagem de proveniência prospectiva. Porém, a arquitetura proposta em Lim et al. (2010) não prove infraestrutura baseada na web semântica - RDF, OWL, ontologias, máquinas de inferência, como é feito no SciProvMiner. Como exemplo, para implementar as regras de inferência definidas pelo modelo OPM o SciProvMiner utiliza a Ontologia OPMO enquanto em Lim et al (2011) essas regras são implementadas através de Views.

Em Zeng et al. (2011) os autores apresentam um método de mineração baseado em dados de proveniência para criar e analisar workflows científicos. No SciProvMiner também utiliza-se métodos de mineração aplicados nos dados de proveniência coletados, mas o propósito é descoberta de conhecimento inesperado, não focado na criação de workflows científicos.

4. SciProvMiner – Arquitetura para captura de proveniência de dados de workflows científicos utilizando o modelo OPM

Em um contexto de e-Science, a ênfase na concepção e construção da arquitetura do SciProvMiner consiste em empregar recursos da web semântica para oferecer ao pesquisador um ambiente fundamentado na interoperabilidade para a gerência e consulta à proveniência de dados heterogêneos e distribuídos.

Especificamente, SciProvMiner estende a arquitetura SciProv proposta em Valente(2011). O SciProv propõe um modelo de proveniência de dados e processos cujo propósito consiste em interagir com os sistemas de gerenciamento de workflows científicos utilizados em um ambiente colaborativo com a finalidade de capturar e gerir as informações de linhagem geradas [Valente 2011]. A arquitetura do SciProvMiner possui as características apresentadas pelo SciProv e evolui esta arquitetura adicionando

a ela a capacidade de capturar a proveniência prospectiva e uma camada de mineração de dados com o objetivo de explorar os dados de proveniência coletados, extraindo, a partir desses dados, informações úteis e desconhecidas que venham a aumentar o poder de análise do cientista sobre o experimento realizado. Com essas modificações, o SciProvMiner fornece uma maior cobertura da proveniência coletada, possibilitando que consultas relacionadas com a especificação do workflow possam ser respondidas, e fornece ao cientista funcionalidades novas, que possam auxiliá-lo na tomada de decisão. Estas novas funcionalidades são baseadas na proveniência de dados de workflows científicos pela utilização da ontologia OPMO e técnicas de mineração de dados para poder extrair conhecimento novo e útil ao usuário com base nas regras de completude e inferência válidas no modelo OPM.

Uma representação da arquitetura do SciProvMiner com as respectivas contribuições em destaque é apresentada na Figura 1.

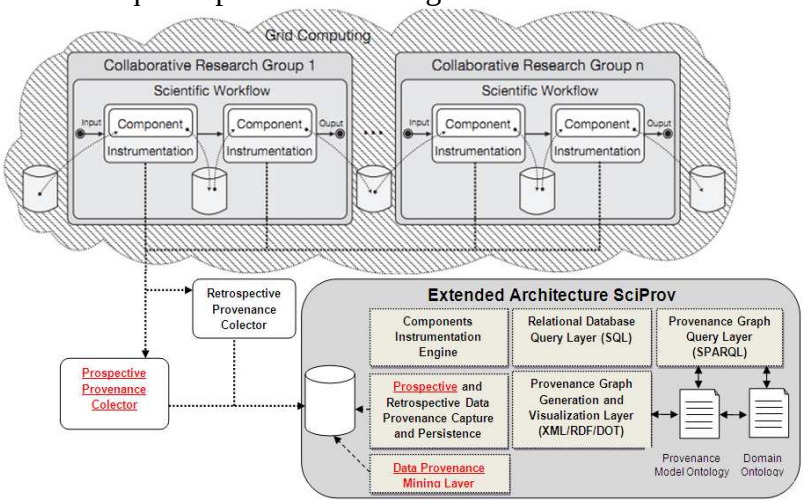

Figura 1. Arquitetura do SciProvMiner

Assim como o SciProv, o SciProvMiner apresenta uma abordagem para o gerenciamento da proveniência de dados de forma independente de SGWfCs. Desta forma, os pesquisadores podem modelar workflows científicos a partir de SGWfCs distintos (como Kepler, Taverna e Vistrails) e cuja execução requer o acesso a repositórios heterogêneos (dados relacionais, semiestruturados, etc.) e distribuídos em uma grade computacional. Neste contexto, um mecanismo de instrumentalização é implementado a partir de tecnologia de serviços web e configurado manualmente para cada componente cuja proveniência deve ser coletada (Figura 1).

A “camada de consulta ao grafo de proveniência” (Provenance Graph Query Layer) na Figura 1 tem por objetivo prover um mecanismo e uma interface para que o usuário formule consultas a partir de uma linguagem de consulta da web semântica. Essa camada está associada às ontologias do modelo OPM de proveniência e do domínio estudado. Esse recurso confere ao SciProvMiner a possibilidade de processar consultas a partir de máquinas de inferência, capazes de efetuar deduções sobre essas bases de conhecimento e obter resultados importantes ao extrair informações adicionais além daquelas que encontram-se registradas de forma explícita nos grafos de proveniência gerados.

Na camada de mineração de dados de proveniência “Data Provenance Mining

Layer” é realizada a busca por padrões desconhecidos e úteis nos dados de proveniência aplicando-se técnicas de mineração de dados para este fim. As técnicas de mineração são utilizadas sobre as regras de completude definidas do modelo OPM em Moreau et al. [2011] para tentar encontrar peças do modelo desconhecidas (artefato ou processo, dependendo da regra) que, se descobertas aumentam o conhecimento do cientista a cerca do experimento. Esse processo de mineração é detalhado na seção 4.2.

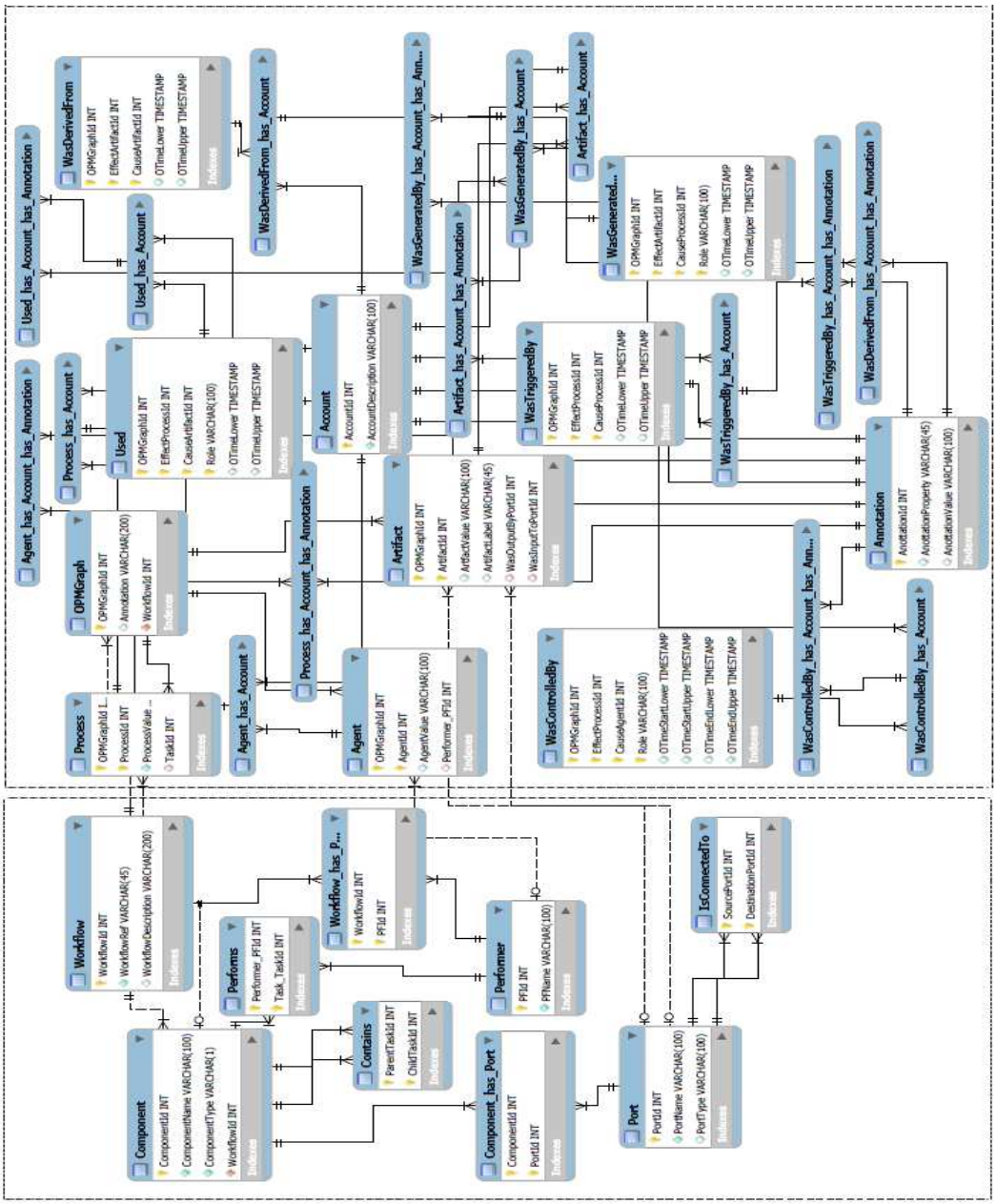

Figura 2. Modelo Relacional do SciProvMiner

4.1 Modelo de Proveniência

Considerando a importância em se obter todas as informações necessárias sobre proveniência, e considerando que o modelo OPM proposto em Moreau et al. (2011) prevê apenas a captura da proveniência retrospectiva, utilizamos a extensão do modelo

OPM proposto em Lim et al. (2011) que captura também a proveniência prospectiva, para projetar o modelo de proveniência do SciProvMiner. A Figura 2 apresenta o modelo relacional do SciProvMiner, estando a proveniência prospectiva representada no retângulo à esquerda da figura e a proveniência retrospectiva representada no retângulo à direita da figura. Com esta extensão se torna possível coletar informações da proveniência prospectiva, tais como descrição do workflow, as tarefas que fazem parte do workflow, subtarefas de uma tarefa, o agente que desempenhou uma tarefa bem como as portas de entrada e saída de uma tarefa, onde a porta de saída de uma tarefa pode ser conectada a porta de entrada de outra tarefa, caracterizando assim o fluxo de dados.

4.2 Aplicação das regras de completude definidas no modelo OPM pelo SciProvMiner

Em Moreau et al. [2011] é dito ser esperado que algoritmos inteligentes possam explorar o modelo de dados do OPM para fornecer novas funcionalidades para os usuários. Com vistas nisso, o SciProvMiner utiliza a ontologia OPMO e técnicas de mineração de dados para poder extrair conhecimento novo e útil ao usuário com base nas regras de completude e inferência válidas no modelo OPM.

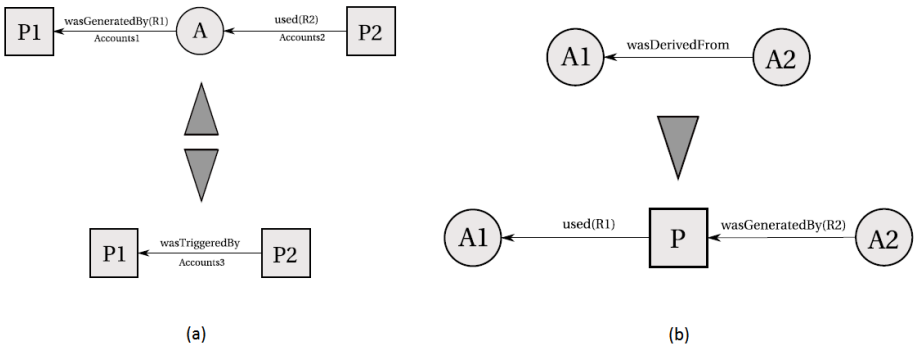

Figura 3. Completude: (a) Eliminação e Introdução de Artefato e (b) Introdução de Processo [adaptado de Moreau et al. 2011]

Regras de completude são transformações diretas, em um único passo, que definem como um subgrafo pode ser convertido em outro subgrafo. Já as inferências em múltiplos passos envolvem múltiplas transações e são definidas para permitir que o usuário encontre causas indiretas que deram origem a um determinado artefato ou processo [Moreau et al 2011]. Dentre as regras de completude definidas no modelo OPM, apenas a transformação em um passo, que diz que uma aresta “was triggered by” pode ser obtida a partir da existência das arestas “used” e “was generated by”, pode ser inferida por uma ontologia. A regra recíproca a esta, que diz que a introdução de um artefato permite estabelecer que uma aresta “was triggered by” está escondendo a existência de algum artefato usado por P2 e gerada por P1 (Figura 3(a)) não pode ser inferido por uma ontologia, devido ao fato de não ser possível inferir através de uma ontologia qual é o artefato ao qual a regra está se referindo. Isso é consequência da aresta “was triggered by” ser um resumo com perda da composição das arestas “used” e “was generated by”. O SciProvMiner utiliza técnicas de mineração de dados para auxiliar o usuário a inferir qual seria este artefato.

Da mesma forma, a regra de completude que diz que uma aresta “was derived from” esconde a presença de um processo intermediário, conforme Figura 3(b), e

também não pode ser inferida em uma ontologia, por não se saber qual seria o processo a ser incluído. O SciProvMiner também neste caso utiliza de técnicas de mineração de dados para auxiliar o usuário a inferir qual seria o processo a ser incluído na transformação, e, desta forma, viabilizar a transformação. Em relação às regras de inferências de múltiplos passos definidas no modelo OPM em Moreau et al. (2011), todas elas são passíveis de serem capturadas através de uma ontologia OWL. Em Moreau et al. (2011) é definida uma ontologia OWL para capturar os conceitos da versão 1.1 do modelo OPM e as inferências válidas neste conceito, denominada Open Provenance Model Ontology.

A partir de uma análise detalhada desta ontologia, pôde-se constatar que são necessários alguns ajustes desta ontologia para que as inferências passíveis de serem capturadas por ontologia sejam todas elas capturadas na OPMO. Com vistas nisso, propomos as modificações necessárias nesta ontologia com fim de a utilizarmos no SciProvMiner, garantindo assim que o SciProvMiner disponibiliza ao usuário uma cobertura completa das inferências válidas no modelo OPMO. Por exemplo, detectamos que a captura da regra de transformação em um passo que diz que uma aresta “was triggered by” pode ser obtida a partir da existência das arestas “used” e “was generated by”, não está definida nesta ontologia. Como o objetivo de defini-la, inserimos um object property wasTriggeredByOneStep com a classe Processo como Domínio e Range. Para definir a propriedade, utilizamos o conceito de property chains, que define uma propriedade em termos de uma cadeia de propriedades que conectam recursos Sendo assim, a propriedade wasTriggeredByOneStep ficou definida através das propriedades used e wasGeneratedBy escrita da seguinte maneira: “used o wasGeneratedBy subPropertyOf wasTriggeredByOneStep”. As demais regras definidas não serão colocadas neste trabalho por falta de espaço.

5. Considerações Finais

O SciProvMiner busca contribuir para a evolução do SciProv em duas frentes principais, que são a inclusão da captura de proveniência prospectiva e na adição de uma camada de mineração de dados para auxiliar o cientista a inferir as regras de completude válidas no modelo OPM que não podem ser diretamente inferidas pela ontologia OPMO e dessa forma cobrir todas as regras de inferência válidas no modelo OPM, conforme definido em Moreau et al. (2011).

Além dessas contribuições principais o SciProvMiner também evoluiu o SciProv em outros pontos, a saber: O SciProvMiner utiliza a mais recente versão da ontologia OPMO, com várias modificações em comparação a versão utilizada pelo SciProv; O modelo de dados do SciProvMiner se difere substancialmente do modelo de dados do SciProv, pela estrutura do banco de dados, pela inclusão da captura de proveniência prospectiva, pela forma como as entidades do modelo OPM se relacionam com a entidade account, também definida no modelo OPM, e pela inclusão da possibilidade de captura de Anotações, entidade também definida no modelo OPM; A captura dos dados de proveniência no SciProvMiner é realizado utilizando controle de sessão, evitando dessa forma erros na persistência dos dados de proveniência do workflow executado; Considerando que uma das principais características do SciProvMiner é o uso de recursos da Web Semântica para representação e consulta aos dados de proveniência, o SciProvMiner propõe uma extensão da ontologia OPMO para comportar a proveniência

prospectiva, sendo necessário adaptar também algumas ferramentas Java disponíveis para este fim, que atualmente capturam apenas a proveniência retrospectiva.

Com o objetivo de avaliar as funcionalidades do SciProvMiner, será realizado um estudo de caso com o workflow definido para as atividade *do Third Provenance Challanges*, usando as 15 consultas de proveniência definidas por este evento como meio de avaliação da capacidade do SciProvMiner em armazenar e responder às consultas aos dados de proveniência.

6. Referências Bibliográficas

FREIRE, J., KOOP, D., SANTOS, E., SILVA, C. T., Provenance for Computational Tasks: A Survey, Computing in Science and Enginnering, v. 10, n. 3,pp. 11-21, 2008.

LIM, C., LU, S., CHEBOTKOT, A., FOTOUHI, F. Prospective and retrospective provenance collection in scientific workflow environments (2010) *Proceedings - 2010 IEEE 7th International Conference on Services Computing, SCC 2010*, art. no. 5557202, pp. 449-456.

MARINHO, A. S. ProvManager: Uma Abordagem para Gerenciamento de Proveniência de Experimentos Científicos, dissertação de mestrado, Engenharia de Sistemas e Computação, UFRJ/COPPE, Rio de Janeiro, RJ, Brasil, 2011.

MATTOSO, M., WERNER, C., TRAVASSOS, G., BRAGANHOLO, V., MURTA, L., "Gerenciando experimentos científicos em larga escala", In: XXVIII Seminário Integrado de Software e Hardware, pp. 121–135, Sociedade Brasileira de Computação: Porto Alegre, RS, Brazil, 2008.

MOREAU, L., CLIFFORD, B., FREIRE, J., FUTRELLE, J., GIL, Y., GROTH, P., KWASNIKOWSKA, N., MILES, S., MISSIER, P., MYERS, J., PLALE, B., SIMMHAN, Y., STEPHAN, E., DEN BUSSCHE, J. V., The Open Provenance Model core specification (v1.1), Future Generation Computer Systems, v. 27 Issue 6, pp.743 –756, 2011.

VALENTE, W., A., G. SciProv: uma Arquitetura para a Busca Semântica em Metadados de Proveniência no Contexto de e-Science, dissertação de mestrado, UFJF, Juiz de Fora, MG, Brasil, 2011.

ZENG. R., HE. X,.LI. J., LIU. Z., AALST. V.D. A Method to Build and Analyze Scientific Workflows from Provenance through Process Mining. 3rd USENIX Workshop on the Theory and Practice of Provenance. June 2011.

WONG, S. C., MILES, S., FANG, W., GROTH, P. and MOREAU, L. (2005) Provenance-based Validation of E-Science Experiments. In: 4th International Semantic Web Conference (ISWC), 6-10 November 2005, Galway, Ireland. pp. 801-815.