Uma Abordagem Semântica para Linhas de Experimentos Científicos Usando Ontologias*

Daniel de Oliveira[1], Eduardo Ogasawara[1], Fernando Chirigati[1], Vítor Sousa[1], Leonardo Murta[2], Cláudia Werner[1], Marta Mattoso[1]

[1]Programa de Engenharia de Sistemas e Computação – COPPE / UFRJ
Caixa Postal 68.511, Rio de Janeiro, RJ, 21941-972

[2]Instituto de Computação – UFF
Rua Passo da Pátria 156, Niterói, RJ, 24210-240

{danielc,ogasawara,fernando_seabra,silva,werner,marta}@cos.ufrj.br,
leomurta@ic.uff.br

Abstract. *Scientific workflows have been used as an abstraction to compose scientific experiments. However, the composition of these workflows is a complex task. Currently there is no guidance or process to follow to reach a workflow specification. As workflows become more complex, composition needs abstraction and semantic support. Experiment lines are an innovative and promising solution to model and reuse scientific workflows in different levels of abstraction. This paper proposes an ontology coupled to experiment lines to provide semantics and flexibility.*

Resumo. *Workflows científicos vêm sendo usados como uma abstração para compor experimentos científicos. Entretanto, a composição destes workflows é uma tarefa complexa. Atualmente não há um guia ou processo a ser seguido para alcançar uma especificação de um workflow. Conforme os workflows se tornam mais complexos, a composição dos workflows necessita de níveis de abstrações maiores e apoio semântico. As linhas de experimentos são uma solução inovadora e promissora para modelar workflows científicos em diferentes níveis de abstração, apoiando assim a concepção e o reuso de workflows científicos. Este artigo propõe a utilização de ontologias acopladas às linhas de experimentos para oferecer maior semântica e flexibilidade.*

1. Introdução

Workflows científicos são uma abstração poderosa para modelar experimentos científicos (Taylor et al. 2007). Entretanto, os workflows se tornam complexos, principalmente em experimentos de larga escala. Apesar da grande quantidade de Sistemas de Gerência de Workflows Científicos (SGWfC) existentes, eles apresentam limitações no que tange ao ciclo de vida do experimento científico, isto é, à composição de um workflow científico e à seleção de atividades em diferentes níveis de abstração, carecendo, portanto, de modelos que representem o workflow em cada um desses níveis. A grande vantagem de apoiar a composição de um workflow em diferentes níveis de abstração é que o cientista pode se concentrar em compor o experimento, ao

* Os autores agradecem ao CNPq pelo apoio financeiro

invés de se preocupar com problemas de infra-estrutura computacional. Estes sistemas somente permitem que um workflow científico seja modelado diretamente em termos de programas, ou seja, já explicitando qual estrutura computacional será utilizada, caracterizando o que é conhecido por workflow concreto ou executável. Entretanto, é importante que o pesquisador também possa modelar seu workflow em níveis mais altos de abstração, definindo apenas as atividades conceituais do experimento a serem utilizadas, mais próximas do domínio do experimento; a este tipo de workflow, nomeamos de workflow abstrato.

As linhas de experimentos (Ogasawara et al. 2009) representam uma idéia inovadora que pode ser vista como uma abordagem para a representação de experimentos científicos em diferentes níveis de abstração. Elas podem ser caracterizadas como workflows abstratos que são capazes de derivar automaticamente um ou mais workflows concretos. Apesar de trazerem um grande benefício no que tange à representação de workflows com maior nível de abstração, as linhas de experimentos ainda carecem de semântica para apoiar a composição dos workflows.

A utilização de ontologias como o arcabouço para representação de conhecimento é considerada uma das abordagens mais proeminentes na literatura. De acordo com Gruber (1993), uma ontologia pode ser definida como um conjunto de definições de um vocabulário formal, ou ainda como uma descrição de conceitos e relacionamentos que podem existir para um determinado domínio do conhecimento. Assim sendo, ontologias podem ser utilizadas para descrever os conceitos do domínio do workflow e relacioná-los dentre os diversos papéis de abstração de um workflow. Portanto, o uso de ontologias auxilia a encontrar atividades abstratas e concretas necessárias na composição, além de incorporar semântica à composição de workflows e, em particular, às linhas de experimentos. A abordagem de linhas de experimentos se encontra implementada na ferramenta GExpLine (GExp 2009) e a incorporação de ontologias à GExpLine está em andamento.

Este artigo contribui com uma abordagem que combina ontologias e linhas de experimentos, a fim de oferecer apoio semântico durante a composição e análise de workflows científicos. Esta abordagem visa a fornecer meios para: (i) a descoberta de atividades, (ii) a composição dirigida do workflow científico, (iii) a descrição e a checagem de dados envolvidos na execução, e (iv) a análise por meio da busca semântica dos resultados, consultando um repositório representado sob um esquema de proveniência . Além desta introdução, este artigo apresenta na Seção 2 os fundamentos básicos sobre linhas de experimentos. A Seção 3 explica de forma sucinta a ontologia de workflows utilizada. A Seção 4 apresenta como a ontologia foi acoplada às linhas de experimentos. A Seção 5 mostra trabalhos relacionados. A Seção 6 conclui o artigo.

2. Linhas de Experimentos

A linha de experimento (Ogasawara et al. 2009) é uma abordagem para representação de experimentos científicos. O processo pelo qual um workflow concreto é obtido da linha de experimentos é chamado de derivação, e é sucintamente apresentado na Figura 1. Um workflow concreto é derivado da linha de experimento a partir da escolha de uma atividade concreta para cada atividade abstrata (caso seja uma atividade variante, haverá uma lista com todas as possibilidades), e a partir da escolha de inclusão ou não de cada uma das atividades opcionais. Além disso, as atividades escolhidas devem ser

compatíveis entre si. Desta forma, quando uma atividade abstrata pode ser mapeada para mais de uma lista de atividades concretas, é chamada de atividade variante. Quando uma atividade pode ser suprimida no workflow concreto derivado, é chamada de atividade opcional.

Apesar de uma mesma linha poder gerar diversos workflows científicos com seus pontos de opcionalidade e variabilidade, nas linhas de experimento não há nenhuma associação das atividades (sejam elas normais, opcionais ou variantes) com metadados e informações importantes, como com quais algoritmos ou métodos as atividades da linha estão associadas. Por exemplo, uma atividade variante da linha pode nos oferecer três opções, como a atividade *C* da Figura 1. Apesar de sabermos que *C1, C2 e C3* são variações de C, não sabemos quais algoritmos ou métodos estão associados com estas variações, uma vez que esta informação não consta na linha de experimento.

Mesmo que o cientista crie uma linha de experimento representando cada método possível e associe estes métodos a pontos de variação de uma atividade abstrata, isto não seria suficiente. Neste caso, estaríamos compondo uma linha apenas baseados em métodos, perdendo informações de algoritmos, por exemplo, ou necessitando replicar o mesmo encadeamento para representar o mesmo workflow em diferentes níveis de abstração, quando na verdade o desejado é visualizar a mesma atividade em diferentes níveis de abstração. Somente metadados e uma descrição textual poderiam disponibilizar maiores informações sobre a linha. Entretanto, como ressaltado por Gomez-Perez et. al. (2004), a utilização de texto livre limita buscas e gera uma não-uniformização dos termos, já que cada usuário pode descrever uma atividade de maneira diferente. Metadados sem navegação ou inferência dificultam consultas. Em determinados experimentos científicos , este tipo de informação é fundamental para que se extraiam dados de proveniência, como "quais métodos ou algoritmos foram utilizados em workflows executados?".

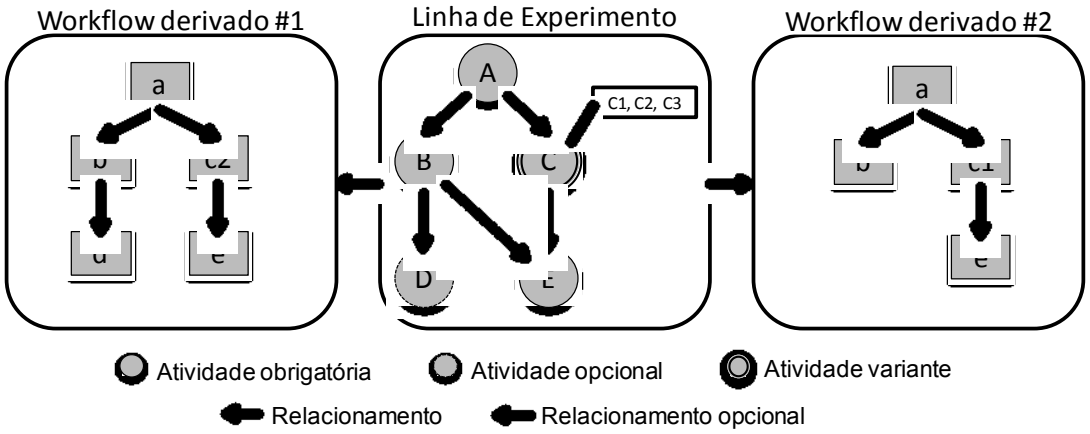

Figura 1 – Exemplo de derivação, adaptado de Ogasawara et al. (2009)

Um problema que pode ser observado é que a descrição das atividades que compõem os workflows gerados e da linha em si é limitada. Estas limitações podem ser contornadas com o acoplamento das linhas com ontologias, que fornecem um arcabouço para modelagem de conhecimento. Na próxima seção, a ontologia utilizada será brevemente descrita.

3. Ontologias para Workflows Científicos

Uma vez que desejamos associar uma ontologia à linha de experimento, devemos descobrir possíveis ontologias candidatas para esta associação. A idéia principal é que acoplemos ontologias previamente avaliadas à GExpLine. Esta seção apresenta sucintamente duas ontologias para workflows científicos que podem ser acopladas à linha de experimentos. Uma das opções é a SciFlow, baseada na MF-Ontology (Cavalcanti et al. 2005, Oliveira et al. 2009), uma ontologia inicialmente modelada para um workflow específico de Mineração de Textos (MT), mas que pôde ser generalizada para workflows científicos, por conter conceitos que são comuns a workflows de diversos domínios do conhecimento, como tarefas, algoritmos e métodos (Cavalcanti et al. 2005). A SciFlow foi baseada nos conceitos descritos por Oliveira et al. (2009) e é baseada em sete conceitos principais: *Passo do Workflow*, *Tarefa*, *Algoritmo*, *Método*, *Medida, Software* e *Dado*. O relacionamento entre esses conceitos pode ser visto na Figura 2. Nesta abordagem, a SciFlow é especializada pelo cientista utilizando ontologias de domínio (já consolidadas e avaliadas) em que se queira compor experimentos científicos.

Uma outra opção é a ontologia myGrid (Wolstencroft et al. 2007). Esta ontologia OWL foi desenvolvida para a descoberta de serviços dentro do SGWfC Taverna (Oinn et al. 2004) via anotação semântica. Ela se encontra subdividida em outras duas: uma ontologia de domínio e outra de serviços. A ontologia de domínio modela o domínio de bioinformática enquanto a de serviços modela a função dos serviços web e parâmetros dentro do Taverna. Na ontologia myGrid, podemos usar inferência para descobrir ancestrais comuns às atividades dos workflows, entretanto, os papéis das atividades e quais métodos ou algoritmos elas compartilham não está explicitamente representado.

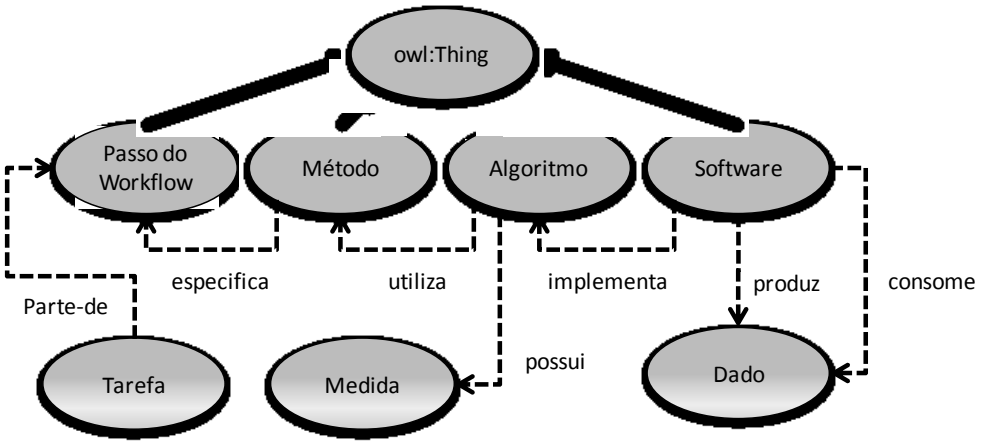

Figura 2 - Parte da ontologia de workflows

Escolhemos acoplar a SciFlow com as linhas de experimento, pois já existe um *middleware* implementado (Oliveira et al. 2009) que nos possibilita trabalhar mais facilmente com esta ontologia. A Figura 2 apresenta parte da SciFlow e mostra os principais conceitos modelados, que podem ser utilizados para adicionar semântica às atividades da linha de experimento, descrevendo-as melhor. É importante ressaltar que esses conceitos são genéricos e que, para cada domínio, é necessário especializar a ontologia SciFlow.

4. Ontologias Apoiando Linhas de Experimentos

A ferramenta GExpLine (GExp 2009) foi desenvolvida para implementar o conceito de linhas de experimentos. Como representação genérica de workflows científicos na linha foi adotada a linguagem padrão do WfMC, denominada XPDL (WfMC 2009), que foi estendida e adaptada para oferecer mais de um nível de abstração aos cientistas e representar conceitos de variabilidade e opcionalidade, essenciais para a linha de experimento.

Atualmente, a GExpLine permite compor workflows de modo descendente. Ou seja, a partir da representação em alto nível da linha de experimento, o usuário pode compor workflows abstratos e, posteriormente, obter a derivação automática do workflow concreto. De forma a possibilitar o processo de derivação, GExpLine prevê a inclusão de cartuchos que permitem a importação e exportação de workflows concretos na linguagem de especificação de um determinado SGWfC. GExpLine conta com cartuchos para os SGWfC Taverna (Oinn et al. 2004) e Kepler (Altintas et al. 2004). A GExpLine também prevê a composição ascendente de workflows, ou seja, a linha de experimento e os workflows abstratos são compostos a partir de workflows concretos previamente importados e comparados.

A utilização de ontologias junto às linhas de experimentos visa a oferecer os seguintes serviços: *(i) Verificação de compatibilidade dos workflows concretos gerados*: na ontologia utilizada, existe o mapeamento da seqüência de programas de um determinado workflow, bem como os tipos de dados que cada programa utiliza; desta maneira podemos saber se dois programas subseqüentes no fluxo são compatíveis, ou seja, se a saída de um possui o mesmo tipo da entrada do próximo; *(ii) Registro de metadados sobre métodos e algoritmos utilizados no experimento*: na ontologia, existe o registro de qual programa implementa qual atividade do workflow, qual algoritmo é utilizado e em qual método é especificado; o cientista poderá então definir a linha de experimentos de maneira descendente, apenas informando qual algoritmo ou método aquela atividade implementa; com a inferência na ontologia, é possível descobrir quais programas são baseados em determinado algoritmo ou método e oferecê-los ao cientista como opção para gerar um workflow concreto; (iii) *Armazenamento junto às definições do workflow*: de forma a apoiar o desenvolvimento colaborativo do workflow, as definições dos workflows (abstrata e concreta) são armazenadas em um esquema relacional do PostgreSQL; ao utilizar uma abordagem semântica, é possível armazenar, no mesmo esquema, os conceitos da ontologia, possibilitando a realização de consultas baseados nos conceitos da ontologia, como, por exemplo, quais workflows gerados utilizam um determinado algoritmo, ou quais linhas de experimento modeladas implementam um determinado método.

Desta maneira, este trabalho visa a completar uma lacuna antes deixada pelos SGWfC e mesmo pelas linhas de experimentos, ou seja, a descrição dos dados, adicionando mais semântica e a validação dos workflows concretos derivados a partir das linhas de experimento. Além disso, a utilização de ontologias facilita a representação de atividades variantes na geração descendente da linha, disponibilizando atividades concretas alternativas para gerar os workflows concretos a partir de um conceito, como, por exemplo, algoritmo ou método, definido na ontologia.

Na Figura 3, podemos ver um exemplo aplicado a um experimento de prospecção de petróleo. A partir da escolha do passo *Análise de Fadiga* para ser

associado à atividade variante C, é possível inferir na ontologia quais são as atividades concretas disponíveis, que, no exemplo da Figura 3, são os softwares γ_1 e γ_2. Desta forma, os softwares γ_1 e γ_2 caracterizam o ponto de variabilidade da linha. Assim, a GExpLine tem meios de verificar se a derivação do workflow abstrato para os workflows concretos correspondentes está sendo feita corretamente. Estas verificações podem ser apoiadas, por exemplo, por serviços de busca (Santanchè e Medeiros 2005).

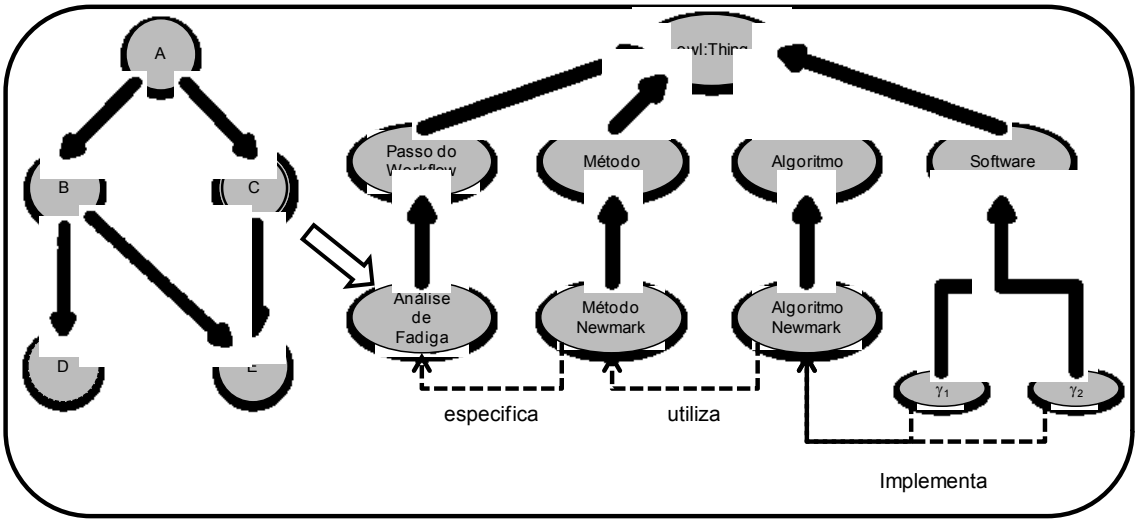

Figura 3 - Linha de experimento alinhada com a ontologia

5. Trabalhos relacionados

Existem diversos SGWfC disponíveis para orquestrar a execução de workflows, alguns deles focados em domínios específicos, outros com o objetivo de serem mais genéricos. O Taverna (Oinn et al. 2004) é um SGWfC inicialmente focado no domínio da bioinformática. Apesar de disponibilizar uma série de serviços agrupados em bibliotecas categorizadas para ajudar na composição do workflow concreto, o Taverna não disponibiliza meios para auxiliar o cientista na composição do experimento, como um guia para a escolha de serviços e programas, por exemplo, e nem fornece diferentes níveis de abstração. O SGWfC Kepler (Altintas et al. 2004) provê um ambiente para gerência de workflows que procura ser independente de domínio. Ele disponibiliza uma interface gráfica e uma biblioteca vasta de componentes. Entretanto, nem sua versão atual, nem seus artigos mencionam apoio semântico para composição de workflows concretos ou abstratos.

O VisTrails (Callahan et al. 2006) foca em capturar e manter uma proveniência detalhada sobre o workflow científico. Ele provê mecanismos para rastreamento de execuções de workflows e suas variações, de forma a facilitar a comparação de resultados e redefinições dos workflows. Embora possua um poderoso mecanismo de proveniência e apoio à composição do workflow concreto (Oliveira et al. 2008), ele ainda não disponibiliza suporte semântico com ontologias ou outro recurso para auxiliar a composição do workflow em diferentes níveis de abstração.

6. Conclusão e Trabalhos Futuros

As linhas de experimentos representam uma contribuição e um avanço na gerência de experimentos científicos, disponibilizando níveis de abstração fundamentais ao

ambiente científico. Entretanto, elas apresentam limitações no que tange à semântica envolvida no processo. As descrições de recursos e atividades das linhas ainda carecem de maior semântica, sendo atualmente representadas por descrições em texto livre. Utilizando conceitos de representação de conhecimento, como ontologias, é possível oferecer um arcabouço que permite a composição facilitada das linhas, verificações de erros e adição de descrição aos dados e atividades baseados em um vocabulário controlado.

Este artigo apresenta formas de aplicação de ontologias no conceito de linhas de experimentos e o seu desenvolvimento em andamento junto à GExpLine. No estágio atual do desenvolvimento, estamos integrando à GExpLine os mecanismos de inferência já existentes na ferramenta MiningFlow (Oliveira et al. 2007). Para avaliar a proposta, utilizaremos dois workflows científicos (prospecção de petróleo e dinâmica de fluidos), que já se encontram especificados em representações concretas, para gerar linhas de experimento usando ontologias. Assim, poderemos comparar os resultados obtidos através de uma especificação na forma concreta com os workflows derivados através da GExpLine.

Desta maneira, o cientista pode modelar seu workflow em níveis mais altos de abstração, sem se preocupar com aspectos de infra-estrutura. Uma representação do workflow com semântica, como é o caso da linha de experimento aliada com ontologias, é particularmente importante para experimentos em larga escala. A GExpLine é parte integrante do projeto de apoio a experimentos em larga escala chamado GExp e se encontra alinhado aos Grandes Desafios da Computação da SBC (SBC 2006).

Referências

Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludascher, B., Mock, S., (2004), "Kepler: an extensible system for design and execution of scientific workflows". In: *Proceedings. 16th International Conference on Scientific and Statistical Database Management*, p. 423-424, Santorini, Greece.

Callahan, S. P., Freire, J., Santos, E., Scheidegger, C. E., Silva, C. T., Vo, H. T., (2006), "VisTrails: visualization meets data management". In: *Proceedings of the 2006 ACM SIGMOD*, p. 745-747, Chicago, IL, USA.

Cavalcanti, M. C., Targino, R., Baião, F., Rössle, S. C., Bisch, P. M., Pires, P. F., Campos, M. L. M., Mattoso, M., (2005), "Managing structural genomic workflows using web services", *Data & Knowledge Engineering*, v. 53, n. 1, p. 45-74.

GExp, (2009), *Brazilian project for supporting large scale management of scientific experiments*, http://gexp.nacad.ufrj.br/.

Gomez-Perez, A., Corcho, O., Fernandez-Lopez, M., (2004), *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. First Edition*. Springer.

Gruber, T. R., (1993), "A translation approach to portable ontology specifications", *Knowl. Acquis.*, v. 5, n. 2, p. 199-220.

Ogasawara, E., Paulino, C., Murta, L., Werner, C., Mattoso, M., (2009), "Experiment Line: Software Reuse in Scientific Workflows". In: *Proceedings of the 21th international conference on Scientific and Statistical Database Management*, p. 264–272, New Orleans, LA.

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M. R., et al., (2004), *Taverna: a tool for the composition and enactment of bioinformatics workflows*. Oxford Univ Press.

Oliveira, D., Baião, F., Mattoso, M., (2007), "MiningFlow: Adding Semantics to Text Mining Workflows". In: *First Poster Session of the Brazilian Simposium on Databases*, p. 15-18, João Pessoa, PB - Brazil.

Oliveira, D., Cunha, L., Tomaz, L., Pereira, V., Mattoso, M., (2009), "Using Ontologies to Support Deep Water Oil Exploration Scientific Workflows". In: *IEEE International Workshop on Scientific Workflows*, Los Angeles, California, United States.

Oliveira, F., Murta, L., Werner, C., Mattoso, M., (2008), "Using Provenance to Improve Workflow Design". In: *2nd International Provenance and Annotation Workshop - IPAW* , p. 136 - 143, Salt Lake City, UT, USA.

Santanchè, A., Medeiros, C. B., (2005), "Self Describing Components: Searching for Digital Artifacts on the Web", *In Proc. of XX Brazilian Symposium on Databases*

SBC, (2006). Grandes Desafios da Computação no Brasil: 2006-2016. Disponível em: http://www.sbc.org.br/index.php?language=1&content=downloads&id=272. Acesso em: 22 Jan 2009.

Taylor, I. J., Deelman, E., Gannon, D. B., Shields, M., (Eds.) , (2007), *Workflows for e-Science: Scientific Workflows for Grids*. 1 ed. Springer.

WfMC, I., (2009), *Binding, WfMC Standards*, WFMC-TC-1023, http://www. wfmc. org, 2000.

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