Desafios no apoio à composição de experimentos científicos em larga escala

Marta Mattoso[1], Cláudia Werner[1], Guilherme Horta Travassos[1], Vanessa Braganholo[2], Leonardo Murta[3], Eduardo Ogasawara[1], Frederico de Oliveira[1], Wallace Martinho[1]

[1] Programa de Engenharia de Sistemas e Computação - COPPE / UFRJ
[2] Departamento de Ciência da Computação - IM / UFRJ
[3] Instituto de Computação – Universidade Federal Fluminense
{marta,werner,ght,ogasawara,ftoliveira,wpereira}@cos.ufrj.br
braganholo@dcc.ufrj.br, leomurta@ic.uff.br

Abstract. *Management of scientific experiments requires a set of specific functionalities. One of such functionalities is the support to experiment composition, which includes scientific workflows conception. However, little support is given to: conception and instantiation for execution in a Workflow Management System; reuse of workflows; control about the evolution of different workflow; and gathering of information for provenance of data. In this paper we present solutions to some of these problems. Such solutions were obtained with Software Engineering and Databases Techniques. Preliminary results with real experiments point to the feasibility of this approach.*

Resumo. *Para que experimentos científicos em larga escala possam ser gerenciados, é necessário que um conjunto de funcionalidades esteja presente. Dentre essas funcionalidades está o apoio à composição dos experimentos, que inclui a concepção de workflows científicos. No entanto, pouco apoio é oferecido à concepção e instanciação para execução num Sistema de Gerência de Workflows, à reutilização, ao controle sobre a evolução dos workflows e à coleta de informações para proveniência de dados. Neste artigo apresentamos soluções para alguns destes problemas a partir de técnicas de Engenharia de Software e Banco de Dados. Resultados preliminares com experimentos reais apontam para a viabilidade dessa abordagem.*

1. Introdução

A busca pelo conhecimento faz com que instituições de pesquisa procurem formas de melhorar a qualidade dos experimentos científicos e reduzir o tempo necessário para a sua execução. A adoção de técnicas que permitam atingir elevados ganhos de produtividade e qualidade na condução de experimentos científicos pode ser vista como um diferencial competitivo para essas instituições e, conseqüentemente, para seus países sede.

Desta forma, a ciência, de um modo geral, tem feito cada vez mais uso de procedimentos computacionais com o intuito de lidar com o aumento constante dos volumes de dados e manipulações necessárias aos experimentos científicos. Além disso, apesar do conhecimento científico continuar a ser gerado por experimentos tradicionais, *in-vivo* e *in-vitro*, novas modalidades de experimentos científicos vêm ganhando importância: *in-virtuo* e *in-silico* (Travassos e Barros 2003). Nesses casos, os objetos de

análise dos experimentos são usualmente processados por simulações computacionais e analisados via técnicas de visualização (Deelman et al. 2008).

Contudo, o cenário atual nos remete aos primórdios da computação, pois centros renomados de pesquisa ainda dependem exclusivamente da capacidade individual dos cientistas no encadeamento dos programas necessários para a execução de experimentos. Esse cenário é sujeito a falhas e improdutivo, especialmente em se tratando de experimentos complexos, que podem envolver muitos programas, grande quantidade de dados e diversos cientistas em localidades geograficamente dispersas.

Com o intuito de atenuar esse cenário, sistemas de gerência de workflows científicos (SGWfC) passaram a ser utilizados. O uso desses sistemas representa um avanço se comparado à abordagem manual inicialmente utilizada, porém, o apoio computacional ao experimento científico em larga escala encontra-se ainda incipiente (Gil et al. 2007) (Mattoso et al. 2008). Um experimento científico se caracteriza por três grandes etapas, a saber, a composição, a execução e a análise dos diversos resultados, entre eles os dados de proveniência do experimento. Entretanto, o estado-da-arte das pesquisas está concentrado no workflow de forma isolada do experimento. Os SGWfC apóiam a execução de workflows de modo controlado e documentado, porém não oferecem recursos para acompanhar as três etapas do experimento científico como um todo. Mais especificamente, a etapa de composição do experimento é a mais incipiente na literatura e quase inexistente nos SGWfC. Assim, há a necessidade de um apoio sistemático à composição de um experimento que envolve a modelagem de workflows e o registro de variações do workflow no contexto do experimento, por exemplo.

Para que experimentos científicos em larga escala possam ser gerenciados, é necessário que um conjunto de funcionalidades esteja presente. Dentre essas funcionalidades, estão o apoio à concepção dos workflows científicos e sua posterior instanciação num SGWfC, a reutilização de workflows previamente concebidos por outros cientistas, o controle sobre a evolução das diferentes versões dos workflows e a coleta de informações que permitam identificar a proveniência dos dados gerados pela execução dos workflows científicos. É fundamental que essas funcionalidades estejam atreladas ao experimento científico que está sendo conduzido por uma equipe de cientistas.

Na ciência da computação, em especial nas áreas de Engenharia de Software e Banco de Dados, existem técnicas que podem ser adaptadas para fornecer o apoio necessário ao gerenciamento de experimentos científicos em larga escala. Especificamente, técnicas de engenharia de requisitos, linhas de produto, mineração de dados, controle de versões e rastreabilidade são de extrema valia para esse fim. Essas técnicas podem ser vistas como formas de implementação das funcionalidades necessárias, descritas anteriormente. Entretanto, não foram encontrados na literatura técnica trabalhos relacionados que façam uso dessas técnicas no apoio à gerência dos experimentos científicos.

O objetivo deste artigo é apresentar direções de pesquisa e algumas soluções para os desafios referentes ao apoio computacional no desenvolvimento de ciência em larga escala. Esse apoio foi obtido por meio da aplicação das técnicas previamente apresentadas ao contexto em questão. A contribuição do trabalho pode ser vista como complementar aos esforços que vêm sendo realizados nas pesquisas de SGWfC. Os resultados preliminares, em sua maioria, estão vinculados à etapa da composição de

experimentos científicos e à infra-estrutura de apoio, e foram aplicados nos domínios de engenharia do petróleo e bioinformática. O foco principal deste trabalho está no segundo Grande Desafio da SBC (2006), onde é dito que: "O objetivo deste desafio é criar, avaliar, modificar, compor, gerenciar e explorar modelos computacionais para todos esses domínios e aplicações". Experimentos científicos representam exemplos concretos dos modelos computacionais mencionados no segundo desafio.

Este artigo está organizado em seis Seções, além desta introdução. A Seção 2 apresenta os desafios detectados para prover apoio computacional ao desenvolvimento de ciência em larga escala e indica os caminhos identificados para responder a esses desafios. As Seções 3 a 5 detalham, respectivamente, os resultados obtidos até o momento no que diz respeito à concepção, reutilização e gerência de configuração em experimentos científicos, além de analisarem os trabalhos relacionados. Finalmente, a Seção 6 apresenta considerações finais e trabalhos futuros.

2. Gerência de Experimentos Científicos em Larga Escala

Um experimento é uma das formas utilizadas pelos cientistas para apoiar a formulação de novas teorias. Como o workflow científico representa a definição da orquestração de seqüências de processos que manipulam dados de modo a construir uma simulação, ele é o principal recurso do experimento científico. Entretanto, neste contexto, um experimento se caracteriza pela composição e execução de diversas variações de um workflow. Essas variações incluem a mudança de dados de entrada, parâmetros, programas ou ainda a combinação delas (Ogasawara et al. 2008, Oliveira et al. 2008). Conforme mencionado anteriormente, os SGWfC se limitam a gerenciar a execução de um workflow científico de forma isolada do experimento ao qual ele faz parte. Assim, variações de um "mesmo" workflow são muitas vezes consideradas workflows diferentes ou então versões do workflow sem o conhecimento da sua razão de existência. Contudo, para representar e apoiar o desenvolvimento do experimento científico, é necessário o registro das variações dos workflows associados a um experimento. O conceito de versões é importante, mas muitas vezes o resultado final do experimento será obtido com resultados de variações dos workflows, não havendo uma única versão final representativa do experimento.

Na visão do projeto myGrid (Oinn et al. 2007), um experimento científico possui um ciclo de vida com cinco etapas. Agrupamos essas cinco etapas em três fases, a saber, composição, execução e análise, para simplificar esse ciclo, conforme apresentamos na Figura 1. Dessa forma, é possível fazer uma analogia entre o experimento e um sistema computacional de forma mais genérica. A seguir, descrevemos o ciclo a partir das três principais fases:

- **Composição**: Essa é a fase responsável pela elaboração dos workflows que devem fazer parte do experimento. A composição trata da conceituação do escopo de uma atividade, da seleção de um programa ou componente adequado para implementar a atividade e também a configuração do fluxo de atividade. Assim, ela é representada na Figura 1 pela busca de atividades adequadas, pelo reúso de workflows já definidos que sejam adequados ao experimento até se chegar a um fluxo de dados e atividades ainda em alto nível de representação. Numa segunda etapa da composição, esse workflow é transformado para ser executado numa máquina de execução de workflows. Após a execução e análise, uma nova composição pode ser feita, caracterizando assim o ciclo;

- **Execução** Nesta fase o foco é na execução dos workflows propriamente ditos, incluindo o monitoramento das execuções e a distribuição de dados e de programas (Couvares et al. 2007). A rastreabilidade também faz parte desta etapa, viabilizando relacionar as atividades do workflow com os resultados obtidos;
- **Análise**: Nesta fase o foco é na avaliação dos resultados experimentais obtidos das execuções dos workflows, que incluem atividades como visualização de dados e consultas (Freire et al. 2008). Além disso, o compartilhamento dos resultados vai ajudar o reúso de workflows definidos ao longo do ciclo de vida do experimento.

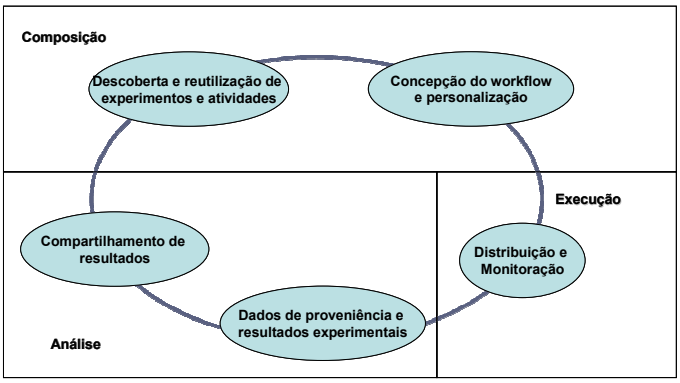


Figura 1. Ciclo de vida de um experimento científico (adaptado de (Oinn et al. 2007))

Em um trabalho prévio (Mattoso et al. 2008), foi feita uma avaliação dos requisitos para se apoiar a experimentação em cada uma das fases discutidas anteriormente, bem como caracterizar as frentes de pesquisas que poderiam contribuir para a solução dos problemas existentes nestas fases. A Figura 2 resume e caracteriza estas frentes de pesquisas e ressalta, dentre os requisitos levantados, quais destes fazem parte do escopo deste trabalho (caixas escuras em alto relevo).

Figura 2. Desafios referentes às etapas do ciclo de experimentação

Reforçando os conceitos apresentados na Figura 1, a fase de composição do experimento possui desafios que podem ser agrupados segundo duas frentes principais, a saber: concepção e reutilização. A concepção engloba as atividades de especificação e modelagem do workflow, bem como a definição do empacotamento do experimento. Na modelagem, é feita a captura do conhecimento referente ao experimento e se projeta o conjunto de workflows para apoiá-lo. No empacotamento, os diversos recursos científicos envolvidos num experimento são combinados em uma estrutura maior. Esses recursos incluem, por exemplo, modelos científicos, algoritmos, programas, versões de programas, workflows, versões de workflows e bases de dados. Os resultados obtidos, referentes à etapa de concepção, são detalhados na Seção 3.

Ainda faz parte da fase de composição do experimento a frente relacionada à reutilização. Esta frente engloba a necessidade de se repetir experimentos e de se tirar vantagem da presença de workflows ou experimentos previamente elaborados para se criar novos workflows ou experimentos. Neste sentido, as experiências obtidas pela área de reutilização de software (Frakes e Kyo Kang 2005) podem ser aplicadas de modo a agilizar e sistematizar a etapa de composição dos experimentos científicos. Os resultados obtidos acerca dos desafios em reutilização de workflows podem ser observados na Seção 4.

É possível também observar, na Figura 2, que a gerência de configuração permeia o ciclo de vida da experimentação. Assim como em software (Estublier 2000), a gerência de configuração no escopo do ciclo de experimentação é uma atividade meio e onipresente. A partir dela se pode inferir qual a versão do workflow que permitiu a obtenção de um resultado em uma simulação e quais foram os passos realizados desde a primeira versão do workflow até a última versão que corroboram para a obtenção do resultado desejado. Neste sentido, ela é utilizada no processo de composição, execução e análise de resultados. Os resultados alcançados em gerência de configuração podem ser observados na Seção 5.

Ao longo das seções dedicadas à concepção, reutilização e gerência de configuração de experimentos científicos, os trabalhos da literatura que mais se aproximam dessas técnicas são analisados.

3. Concepção de Experimentos Científicos

A principal etapa da composição de um experimento é a concepção. Seus objetivos incluem a modelagem e especificação de workflows em seus diversos níveis de abstração e no registro de variações do workflow no contexto do experimento. Frequentemente, workflows são classificados em dois níveis, a saber: concreto e abstrato (Deelman et al. 2008). Um workflow abstrato é modelado sem estar preso a programas e nem à definição de recursos computacionais. Esta abstração permite flexibilidade, visto que não há a necessidade de se entrar no nível de detalhes de desenvolvimento. Neste sentido, diagramas de atividades UML (Pressman 2004) ou especificações em XPDL (Guelfi e Mammar 2006) poderiam ser utilizados para modelar workflows abstratos. Num workflow concreto, entretanto, há a definição de características tecnológicas, como a indicação de programas e de recursos necessários para se executar as atividades. Desta forma, o workflow concreto é uma instância específica de um workflow abstrato para resolver um determinado problema.

De modo a facilitar a concepção do workflow, os SGWfC precisariam apoiar a especificação e modelagem de workflows nesses diferentes níveis de abstração. Entretanto, SGWfC como o Taverna (Hull et al. 2006), Kepler (Altintas et al. 2004), VisTrails (Callahan et al. 2006), Swift (Zhao et al. 2005) e OMII-BPEL (Taylor et al. 2006) apóiam a modelagem de workflows com baixo nível de abstração, muito próximo ao nível concreto de especificação. Este tipo de apoio é insuficiente para a concepção de workflows que deveria começar com níveis mais altos de abstração. Afinal, para o especialista no domínio, inicialmente, é mais importante se preocupar com o conceito do experimento, podendo ser mais fácil trabalhar com um workflow abstrato do que um workflow equivalente no nível concreto, com atividades auxiliares (por exemplo, adaptadores) e detalhes sobre os recursos computacionais (e.g., ferramentas ou programas). Considerando o experimento científico, é importante apoiar tanto o

workflow em vários níveis de abstração quanto o concreto e o relacionamento entre eles. Nos SGWfC esse apoio não é encontrado.

A Figura 3 e a Figura 4 apresentam um workflow de bioinformática no nível abstrato e no nível concreto, respectivamente. O workflow abstrato foi concebido usando um editor gráfico aderente à especificação UML, enquanto o workflow concreto está de acordo com a notação do SGWfC Taverna. No workflow abstrato (Figura 3), é possível identificar, que a partir de uma seqüência de proteínas (ou de ácidos nucléicos) pode-se realizar o alinhamento genético com uma técnica a ser escolhida em outro momento. Essa escolha pode passar pela definição do algoritmo e software até a especificação dos recursos tecnológicos a serem utilizados na execução. As atividades no nível abstrato indicam que algo precisa ser feito, mas não dizem como se deve fazer. Realizando uma analogia com desenvolvimento de software (Pressman 2004), um workflow abstrato está para a etapa de análise da mesma forma que um workflow concreto está para a etapa de projeto.

O workflow concreto (Figura 4), definido no Taverna, tem seus processos representados por retângulos de cor cinza. Os dados de entrada são representados por triângulos voltados para cima. Os dados de saída são representados por triângulos voltados para baixo. Para uniformizar o vocabulário acerca dos diferentes SGWfC, os processos do Taverna serão denominados por atividades. Estas atividades são ligadas diretamente entre si, estabelecendo uma relação de dependência. No nível concreto, as atividades do workflow são pacotes ou programas no domínio da aplicação (atividades em cinza claro, por exemplo, *executaKalign*) ou são atividades que funcionam como adaptadores (atividades em cinza escuro, por exemplo, *desempacotaAlinhamento*), usadas para manipular entrada de dados e transformá-las num formato adequado, de modo a ajustar o fluxo de dados para as demais atividades. Neste sentido, atividades possuem portas de entrada e portas de saída que são utilizadas para se interligarem.

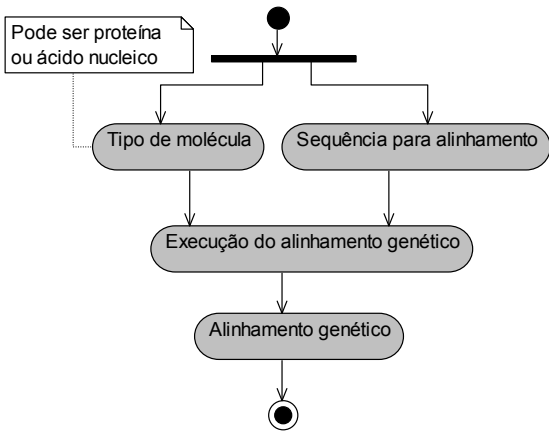


Figura 3. Workflow abstrato para análise de seqüência em UML

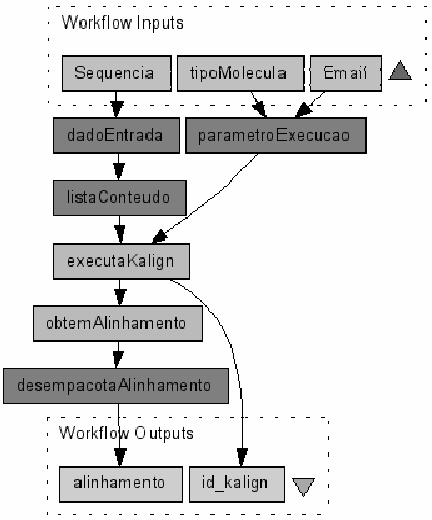


Figura 4. Workflow concreto para análise de seqüência com Kalign

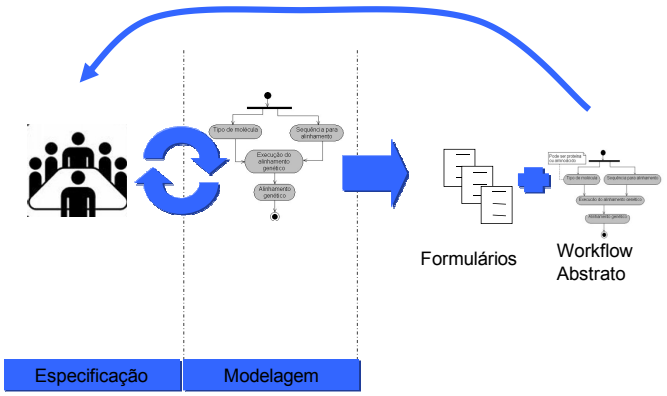
Desta forma, por possuir uma natureza incremental e pela presença de diferentes níveis de abstração, o processo de concepção de workflows científicos, assim como em

software, requer apoio metodológico. Isto é justificado, pois o tamanho dos experimentos e a quantidade de elementos envolvidos vêm aumentando consideravelmente, criando a necessidade de meios para auxiliar o pesquisador na concepção do workflow, partindo dos seus requisitos até chegar à representação executável por um SGWfC. Porém, na literatura, não foram encontrados processos ou metodologias para apoiar este tipo de concepção. Desta maneira, tornou-se necessária a criação de uma abordagem para concepção de workflows.

A abordagem criada tem como princípio o uso de formulários, onde os campos contidos ali representam os requisitos de uma solução possível, isto é, um experimento. Basicamente, existem três tipos de elementos presentes em *workflows* científicos, sendo eles: *atividades*, *resultados* e *ferramentas*. As *atividades* se caracterizam por consumir, modificar ou gerar o conhecimento científico em suas diferentes representações. Estes conhecimentos são definidos como *resultados*, sendo estes os produtos e insumos da execução de uma *atividade*. A Figura 5 mostra parte do formulário de atividades proposto nessa abordagem. Tais *resultados* podem apresentar diversos formatos e padrões, podendo ser digitais ou não, numéricos, textuais, entre muitos outros. Já as *ferramentas* são os apoios computacionais utilizados na execução, representadas, por exemplo, por programas de simuladores de modelo científico. Vale ressaltar que os relacionamentos entre os três elementos estão presentes nos formulários, o que permite uma navegabilidade entre eles. Por exemplo, a partir de uma *atividade* é possível chegar à lista de *ferramentas* que a utilizam, ou de uma determinada *ferramenta* é possível listar os *resultados*, em formato digital, que ela produziria durante a execução.

Além desta especificação textual em formulários, existe também a modelagem do fluxo do experimento, através da notação UML, mais especificamente, do diagrama de atividades (Pressman 2004). Esta representação visual serve para explicitar os pontos de decisão e sincronização (e.g.: divisão, junção) e de encadeamento seqüencial das atividades (e.g.: repetição, associação entre as atividades). Isto acaba por tornar mais legível o experimento, em um primeiro momento, pois assim o pesquisador identifica estes pontos em uma representação gráfica, tendo os detalhes descritos nos formulários. É importante ressaltar que a especificação textual e a modelagem do fluxo representam as mesmas informações, isto é, o encadeamento das atividades, os elementos que compõem o experimento e os pontos de decisão e sincronização.

Durante a etapa de concepção de um workflow, é necessário entender bem o domínio da aplicação. Nesta etapa, são necessárias reuniões entre os membros do grupo de pesquisa, as quais envolvem a tomada de decisão e análises refinadas sobre a especificação e modelagem do workflow. Para executar esta etapa, foi adotado o uso de entrevistas, que na Engenharia de Software é uma das técnicas de elicitação de requisitos mais utilizadas (Davis et al. 2006). A sua estruturação (ver Figura 5) é baseada, primeiramente, na identificação da seqüência das atividades, tendo como resultado a modelagem do workflow no nível abstrato, descrito em um diagrama de atividades. A partir desta modelagem, os formulários são preenchidos de uma maneira incremental e iterativa, com os produtos da especificação e modelagem sendo avaliados pelo pesquisador. Esta ordem se justifica, pois assim os membros do grupo de pesquisa têm uma visão global do experimento, permitindo que, posteriormente, sejam discutidos os detalhes, e.g.: lista de ferramentas, paralelismo de atividades, etc.



Atividade	[NOME DA ATIVIDADE]
Descrição	[Descreva a atividade de forma detalhada.]
Tipo de atividade	[Informe o tipo de atividade, se a atividade é Manual, Semi-Automatizada ou Automatizada]
Ferramentas	[Liste os softwares necessários para a realização da atividade.]
Insumos	[Liste os artefatos consumidos pela atividade. Esses artefatos são as "matérias-primas" da atividade, ou seja, os dados de entrada necessários para executar a atividade.]
Papéis	[Liste os papéis desempenhados por uma ou mais pessoas para a realização da atividade. Responsáveis pela execução da atividade (e.g.: geólogo).]
Pré-condições	[Liste as pré-condições que devem ser atendidas para iniciar a execução da atividade (e.g.: atividade X executada, recurso financeiro suficiente, etc.).]
Pré-atividades	[Liste as atividades que devem ter sido concluídas para que seja possível a execução dessa atividade.]
Pós-condições	[Liste as pós-condições que são geradas ao final da execução da atividade (e.g: dado gravado no repositório, recurso computacional liberado).]

Figura 5. Elicitação de requisitos

Normalmente, é preciso mais de uma reunião para realizar a identificação completa do workflow abstrato, sendo necessário alguns refinamentos no modelo. A cada nova reunião, os produtos gerados na entrevista anterior são validados pelos pesquisadores envolvidos. Assim, o controle sobre o desenrolar dos experimentos científicos é aumentado, registrando etapas e escolhendo parâmetros, de modo a garantir a eficácia e a qualidade dos resultados nos experimentos científicos.

Mesmo em se tratando de concepção de workflows em nível abstrato, as informações sobre *ferramentas* podem ser capturadas nesta fase. Ao realizar isto, o pesquisador pode acelerar a especificação do workflow concreto, pois só precisará escolher o recurso computacional conveniente. Contudo, o pesquisador pode postergar esta tarefa de levantamento ou solicitar que algum responsável por garantir e manter os recursos computacionais os cadastrem, através do preenchimento dos formulários.

Durante as atividades de especificação e modelagem, uma característica muito importante é ter o envolvimento do cliente, que neste caso é representado pelo pesquisador, pois ele tem o papel de fornecer os requisitos do experimento e, também, de avaliar os produtos gerados na elicitação de requisitos. Afinal, ele é possuidor do conhecimento do domínio e pode determinar como o experimento é representado em termos de atividades, resultados e ferramentas. Outra característica importante nestas etapas é aceitar que pode haver mudanças a qualquer tempo. Durante a modelagem é possível perceber, por exemplo, que uma atividade é muito complexa e que a representação via sub-workflow seja mais apropriada. Desta forma, caso haja alguma alteração, é necessário que esta seja refletida tanto na especificação quanto no modelo. Este procedimento foi inspirado em alguns dos princípios apresentados pelos Métodos Ágeis (Beck 1999), que são utilizados na indústria de desenvolvimento de software, como trabalho em conjunto com o cliente e acolhimento de modificações nos requisitos.

Esta abordagem foi utilizada em um projeto de workflow científico real no domínio de Engenharia de Petróleo no Laboratório de Métodos Computacionais e Sistemas Offshore (LAMCSO) na UFRJ em conjunto com a Petrobrás. Foram especificados *atividades* (incluindo as manuais, semi-automatizadas e automatizadas), *resultados* e *ferramentas* (Figura 5). A partir do diagrama de atividades, apresentado parcialmente na Figura 6, geramos uma instanciação concreta no SGWfC Kepler. Para esse workflow, representativo na área, foram realizadas seis reuniões, com uma duração média de uma hora e meia cada. É interessante ressaltar que durante a especificação e

modelagem realizada, o pesquisador ao exercitar o fluxo de atividades pôde perceber que existiam atividades que pertenciam a mais de um workflow. Isto representou um ganho de conhecimento, pois um sub-workflow, oriundo deste experimento, pôde ser construído e sua especificação e modelagem foram reutilizadas. O especialista no domínio conseguiu entender o diagrama de atividades, que representava o workflow científico, depois da adição de uma legenda com a descrição dos elementos UML utilizados (e.g.: decisão, junção), com isso foi capaz de identificar as *atividades* e os *resultados* ainda não mapeados na ocasião. Foi possível também identificar a necessidade de se desenvolver uma ferramenta específica para apoiar esta abordagem. Conforme o andamento da especificação, devido ao aumento no número de elementos presentes no workflow, houve dificuldade de manuseio dos formulários para seu uso e avaliação.

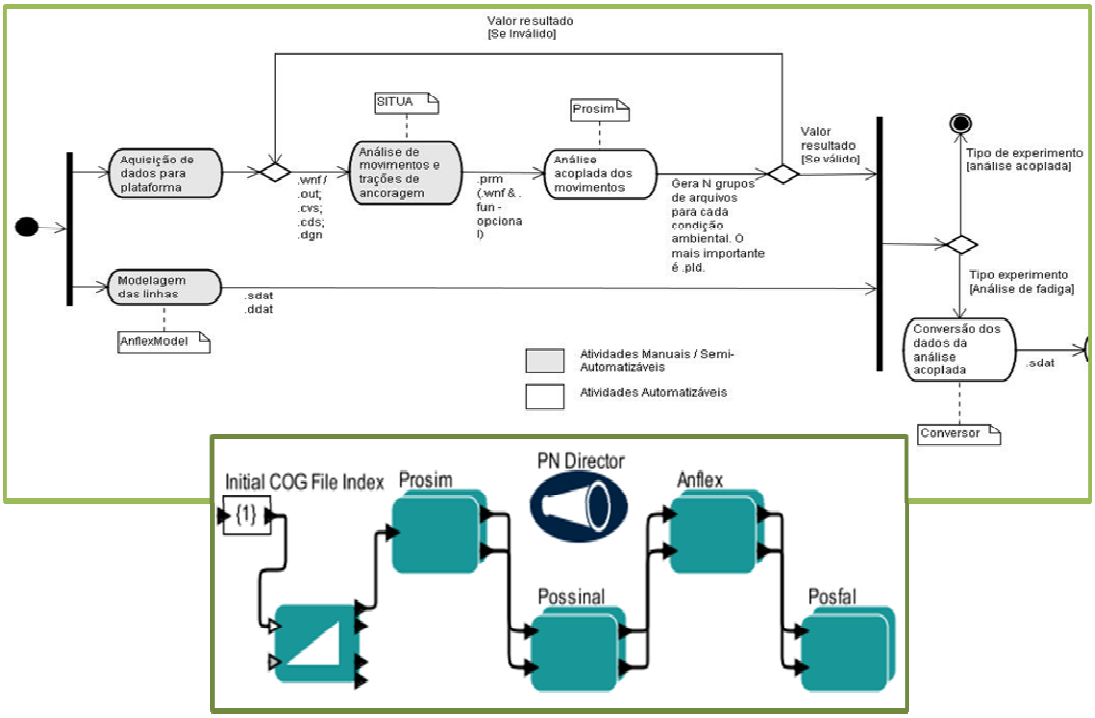


Figura 6. Diagrama de atividades parcial e o workflow concreto no Kepler

4. Reutilização de experimentos científicos

O desafio relacionado à reutilização de experimentos científicos engloba a necessidade de se tirar vantagem da presença de workflows ou experimentos previamente elaborados para se compor novos workflows ou experimentos. Durante o experimento científico, duas situações representativas podem ocorrer. Em uma primeira situação, o pesquisador precisa executar o experimento repetidamente, alterando os dados de entrada e analisando o comportamento do modelo de acordo com as mudanças. Neste caso, o workflow científico precisa ser re-executado apenas variando-se os parâmetros (Goderis et al. 2008). Em uma segunda situação, o pesquisador pode não estar satisfeito com os resultados alcançados e pode decidir explorar o uso de programas alternativos. No primeiro caso, os SGWfC têm primitivas para apoiar a re-execução do workflow, mas no segundo caso eles ainda carecem de recursos para apoiar a experimentação.

Atualmente, esta atividade é feita através de cópias do workflow, seguida de edições específicas. Este tipo de abordagem traz limitações de escalabilidade e manutenibilidade (Taylor et al. 2006), visto que pode ocorrer perda de semântica ou redundância de informações. Os atuais SGWfC consideram cada mudança como um workflow diferente. Mesmo quando existem diferentes programas que desempenham a mesma atividade, não há nenhuma explicitação desse conhecimento via rastreabilidade formal entre eles. Além disto, qualquer mudança no escopo da experimentação demanda alterar todos os workflows concretos relacionados, independentemente de eles pertencerem ao mesmo workflow abstrato. Isso ocorre devido aos SGWfC se concentrarem no workflow concreto e de forma isolada do experimento.

O encadeamento de atividades num workflow não é uma tarefa trivial e, em muitos casos, se transforma em uma barreira para a construção de análises e modelos mais sofisticados. SGWfC, como Taverna, Kepler e VisTrails, oferecem uma interface gráfica rica a partir da qual componentes previamente registrados podem ser arrastados ou colocados na área de edição de workflow. Estas operações geram workflows diretamente no nível concreto. Entretanto, não há apoio às etapas anteriores do processo de concepção de workflows. A busca por atividades é limitada, visto que, por exemplo, não há registro de atividades equivalentes no SGWfC. O conhecimento de quais atividades podem ser interligadas é ainda tácito. É necessário estudar um grande número de exemplos de workflows para se ganhar experiência na configuração do fluxo das atividades.

Uma alternativa usada pelos pesquisadores é tentar compor um workflow a partir de outro previamente montado. O projeto myExperiment (Goderis et al. 2008) provê um sítio com repositório de workflows previamente definidos. A maior parte destes workflows está definida para a linguagem do sistema Taverna e pertence à área de bioinformática. Este repositório de workflows é muito útil quando o workflow do repositório é perfeitamente aderente às necessidades do pesquisador. Entretanto, adaptar o workflow ou tentar remontá-lo em uma outra linguagem não é trivial.

Goderis et al. (2005) e Roure et al. (2007) observaram o crescente interesse acerca do assunto de reutilização de workflow para apoiar a composição de novos workflows. Entretanto, apesar do aumento no uso da expressão "reutilização de workflow", estes trabalhos estão normalmente associados à questão de se variar parâmetros em workflows já existentes. Poucos trabalhos apóiam a concepção de novos workflows, enquanto que nenhum trabalho observado apóia a composição de novos experimentos. Neste sentido, apresentamos a seguir como técnicas de reutilização de software podem ser úteis no apoio a etapa composição de workflows e de experimentos científicos, por exemplo, usando sistemas de recomendação para composição de workflows (Oliveira et al. 2008) e linha de experimento (Ogasawara et al. 2008) para composição de experimentos científicos.

Os sistemas de recomendação são um tipo de filtragem de informação no qual se tenta apresentar itens de informações que provavelmente sejam do interesse de um determinado usuário (Adomavicius e Tuzhilin 2005). Em 2008 surgiram as primeiras tentativas de adaptar técnicas de recomendação para auxiliar a composição de workflows científicos (Ellkvist et al. 2008)(Oliveira et al. 2008).Vale ressaltar que estas iniciativas podem se beneficiar do grande número de atividades disponíveis, por exemplo, em 2007 o Taverna tinha mais de 3500 serviços, enquanto que em 2008 o

VisTrails tinha mais de 1200 módulos. O VisComplete (Koop et al. 2008) baseou-se no trabalho de Ellkvist (2008) e introduziu a técnica de recomendação no VisTrails, permitindo sugerir automaticamente as atividades na composição do workflow. Para isto, ele calcula as correspondências entre o workflow em desenvolvimento e os previamente criados, gerando uma recomendação.

Em nossa abordagem (Oliveira et al. 2008) de recomendação de atividades durante a composição de workflows, optamos por usar a técnica de *sequence mining* (Srikant e Agrawal 1996). Considerando como seqüência cada caminho existente em um workflow previamente criado, foi possível gerar regras respeitando a ordem de encadeamento das atividades. O processamento dessas regras pode gerar novas recomendações de atividades ao identificar similaridades no encadeamento sendo definido. Diferentemente do VisComplete, o uso de *sequence mining* permitiu gerar recomendações para um maior número de casos (ou maior número de workflows em fase de composição, pois são consideradas não apenas as seqüências em que os elementos são consecutivos (A→B→C) como também, as seqüências indiretas (A→ ... →B→C). Atualmente estamos integrando esta técnica baseada em *sequence mining* ao VisTrails, para comparar com a técnica utilizada pelo VisComplete usando *benchmarks*.

No que tange à composição de experimentos, é proposta a adaptação do conceito de linha de produtos em engenharia de software, para se construir o conceito de linha de experimento (Ogasawara et al. 2009a), que vem a ser uma abordagem sistemática para apoiar a reutilização de experimentos científicos. Através desta técnica é possível apoiar a fase de composição de experimentos científicos com a concepção de workflows em diferentes níveis de abstração. Não foram encontrados na literatura outros trabalhos com esse apoio semântico. Em uma linha de experimentos, há uma relação forte entre os experimentos científicos, os workflows abstratos e os múltiplos workflows concretos que podem ser gerados. A partir dela, pode-se separar o papel de se modelar um workflow com alto grau de abstração (no nível do experimento científico) do papel de se reutilizar um workflow previamente montado. O conceito de reutilização de um workflow em uma linha de experimentos está intimamente ligado a uma concepção guiada, a partir de informações que estão no nível de abstração do experimento propriamente dito. Esta separação é um conceito chave para se evitar retrabalho, apoiar melhor a reutilização de workflows e contribuir para que a composição de experimentos possa apoiar experimentos em larga escala. A ferramenta GExpline (GExp 2009) visa a gerenciar linhas de experimentos científicos. Ela auxilia a modelagem de workflows científicos abstratos e gera workflows concretos para serem executados no Kepler e Taverna. O apoio para o Vistrails encontra-se em andamento. Apesar de o conceito da linha de experimento ser independente do SGWfC, a atual implementação da ferramenta apóia um único SGWfC para cada definição de linha de experimento. GExpline foi usada na construção de uma linha de experimento para workflows de bioinformática e esta em andamento a definição de uma linha de experimento no contexto de Engenharia de Petróleo.

5. Gerência de Configuração de Workflows

A gerência de configuração é a disciplina responsável por controlar a evolução de sistemas de software (Conradi e Westfechtel 1998). O foco desta seção está em analisar a gerência de configuração sob a ótica de uma disciplina de desenvolvimento (Conradi e Westfechtel 1998), aplicada no apoio à fase de composição de workflows. Neste

sentido, para os workflows científicos possuírem gerência de configuração, como no caso de software, precisam de ferramentas que ofereçam as seguintes características:

- Um repositório de workflows com controle de acesso e controle de concorrência, de modo que seja possível armazenar workflows e registrar a versão estável (produção) e versões em desenvolvimento;
- Um mecanismo que permita representar e armazenar as versões de atividades que estão sendo utilizadas por uma composição de workflows;
- A presença de conceito de área de trabalho (*workspace*) para apoiar a edição de workflow. A área de trabalho precisa também apoiar a publicação no repositório.

Entretanto, nenhum dos SGWfC apóia gerência de configuração de workflows científicos (Ogasawara et al. 2009b). O VisTrails é um dos sistemas mais ativos nesta área e permite o armazenamento do histórico das versões, mas esta facilidade não é suficiente para apoiar a gerência de configuração, uma vez que não é possível separar a versão de produção (versão disponível para outros usuários reutilizarem (Frakes e Kyo Kang 2005)) das áreas de trabalho para desenvolvimento (versão do workflow sob desenvolvimento, sem garantias de qualidade).

A ferramenta GExpline vem sendo aprimorada no sentido de também oferecer estes recursos. No GExpline, o processo pelo qual um usuário decide usar ou editar um workflow é denominado *check-out* (Conradi e Westfechtel 1998). Tendo terminado o seu trabalho e decidido compartilhar o workflow alterado, o usuário executa uma operação denominada *check-in*. Após o *check-in*, uma nova versão do workflow é criada.

O GExpline oferece dois mecanismos de controle de edição que podem ser utilizados durante o ciclo de *check-out*, edição e *check-in.* O primeiro é uma abordagem otimista, na qual o workflow não é bloqueado para mudanças e dois ou mais usuários podem modificá-los em paralelo. O segundo mecanismo é pessimista, no qual o workflow fica bloqueado para a operação de *check-in* e apenas o usuário que fez o *check-out* com bloqueio pode realizar a operação de *check-in*. Entretanto, outros usuários podem realizar *check-out* simultâneo e esperar até que o usuário que fez o bloqueio deposite o seu trabalho.

Quando o usuário realiza o *check-out* do workflow, a área de trabalho fica associada à versão base utilizada durante o *check-out*. No momento em que o usuário for realizar o depósito do workflow, se o mesmo não estiver bloqueado para *check-in* e se a versão do repositório for a mesma que a versão utilizada para a criação da área de trabalho, então o *check-in* do workflow poderá ser feito sem nenhuma restrição. Entretanto, se a versão do repositório já estiver mais avançada, então, neste caso, será necessário realizar a operação de *diff/merge* (recurso este que atualmente encontra-se em desenvolvimento). Esta restrição garante que o trabalho desenvolvido pelo usuário não será perdido de modo inconsciente.

6. Conclusão

A importância do apoio ao cientista na realização de seus experimentos científicos tem sido evidenciada na academia pela área de e-Science e se fez presente no documento dos "Grandes Desafios" da SBC. Neste artigo, analisamos diversos desafios relacionados à composição de experimentos científicos, como por exemplo, a falta de apoio à concepção dos workflows e sua posterior instanciação para execução num

SGWfC, a reutilização de workflows previamente concebidos por outros cientistas, e o controle sobre a evolução das diferentes versões dos workflows.

Para apoiar a composição de experimentos científicos, destacamos a importância do apoio à gerência de workflows no contexto do ciclo de vida do experimento. Nesse sentido, apresentamos uma abordagem sistemática para concepção de workflows em diversos níveis de abstração. Mostramos como sistemas de recomendação e linhas de experimento podem ser adaptadas para apoiar a reutilização de experimentos científicos. Finalmente, exploramos a gerência de configuração no contexto de workflows científicos. A Tabela 1 apresenta um resumo dos desafios referentes à fase de composição, do ciclo de experimentação e as nossas soluções para o problema.

Tabela 1 - Desafios à composição de experimentos e soluções adotadas

Desafios	**Soluções adotadas**
Concepção	(i) Abordagem para concepção de workflows que apóia diferentes níveis de abstração;
Reutilização	(ii) Sistema de Recomendação usando *Sequence mining* para apoiar a composição de workflows; (iii) Linhas de Experimentos para apoiar a composição de experimentos;
Gerência de configuração	(iv) Ferramentas de gerência de configuração de workflows com apoio a *diff/merge* sintático;

Estas soluções vêm sendo desenvolvidas e aplicadas no contexto do projeto de pesquisa "GExp" voltado à gerência de experimentos científicos em larga escala (GExp 2009). Neste projeto, temos a oportunidade de avaliar nossos resultados preliminares com cientistas em situações reais na área de engenharia de petróleo e bioinformática. Como trabalhos em andamento, estamos abordando problemas relacionados às etapas de execução e análise do ciclo de vida dos experimentos, porém ficaram fora do escopo desse artigo. Destacamos iniciativas na captura da proveniência de modo independente do SGWfC (Marinho et al. 2009) para contribuir com os desafios da fase de análise.

Agradecimentos

Os autores agradecem à equipe do projeto GExp pelo desenvolvimento das técnicas e ferramentas que apóiam esse trabalho e ao CNPq pelo apoio financeiro. G.H.Travassos, como Cientista do Nosso Estado, agradece à FAPERJ pelo apoio.

Referências

Adomavicius, G., Tuzhilin, E., (2005), "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions", *IEEE Transactions on Knowledge and Data Engineering*, v. 17, p. 734-749.

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.

Beck, K., (1999), *Extreme Programming Explained: Embrace Change*. Addison-Wesley Professional.

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.

Conradi, R., Westfechtel, B., (1998), "Version Models for Software Configuration Management", *ACM Computing Surveys*, v. 30, n. 2

Couvares, P., Kosar, T., Roy, A., Weber, J., Wenger, K., (2007), "Workflow Management in Condor", *Workflows for e-Science*, Springer, p. 357-375.

Davis, A., Dieste, O., Hickey, A., Juristo, N., Moreno, A., (2006), "Effectiveness of Requirements Elicitation Techniques: Empirical Results Derived from a Systematic Review". In: *14th IEEE International Conference of Requirements Engineering*, p. 179-188

Deelman, E., Gannon, D., Shields, M., Taylor, I., (2008), "Workflows and e-Science: An overview of workflow system features and capabilities", *Future Generation Computer Systems* (Jul.)

Ellkvist, T., Koop, D., Anderson, E. W., Freire, J., Silva, C., (2008), "Using Provenance to Support Real-Time Collaborative Design of Workflows", *Provenance and Annotation of Data and Processes: 2nd International Provenance and Annotation Workshop, Salt Lake City, UT, USA, LNCS*, Springer-Verlag, p. 266-279.

Estublier, J., (2000), "Software configuration management: a roadmap". In: *Proceedings of the Conference on the Future of Software Engineering*, p. 279-289, Limerick, Ireland.

Frakes, W., Kyo Kang, (2005), "Software reuse research: status and future", *IEEE Transactions on Software Engineering*, v. 31, n. 7, p. 529-536.

Freire, J., Koop, D., Santos, E., Silva, C. T., (2008), "Provenance for Computational Tasks: A Survey", *Computing in Science and Engineering*, v. 10, n. 3, p. 11-21.

GExp, (2009), *Large Scale Managament of Scientific Experiments.*, http://gexp.nacad.ufrj.br/.

Gil, Y., Deelman, E., Ellisman, M., Fahringer, T., Fox, G., Gannon, D., Goble, C., Livny, M., Moreau, L., et al., (2007), "Examining the Challenges of Scientific Workflows", *Computer*, v. 40, n. 12, p. 24-32.

Goderis, A., De Roure, D., Goble, C., Bhagat, J., Cruickshank, D., Fisher, P., Michaelides, D., Tanoh, F., (2008), "Discovering Scientific Workflows: The myExperiment Benchmarks", *IEEE Transactions on Automation Science and Engineering*

Goderis, A., Sattler, U., Lord, P., Goble, C., (2005), "Seven Bottlenecks to Workflow Reuse and Repurposing". In: *The Semantic Web – ISWC 2005*, p. 323-337, Galway, Ireland.

Guelfi, N., Mammar, A., (2006), "A formal framework to generate XPDL specifications from UML activity diagrams". In: *Proceedings of the 2006 ACM SAC*, p. 1224-1231, Dijon, France.

Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M. R., Li, P., Oinn, T., (2006), "Taverna: a tool for building and running workflows of services", *Nucleic Acids Research*, v. 34, n. Web Server issue, p. 729-732.

Koop, D., Scheidegger, C., Callahan, S., Freire, J., Silva, C., (2008), "VisComplete: Automating Suggestions for Visualization Pipelines", *IEEE Transactions on Visualization and Computer Graphics*, v. 14, n. 6, p. 1691-1698.

Marinho, A., Murta, L., Werner, C., Braganholo, V., Cruz, S. M. S. D., Mattoso, M., (2009), "A Strategy for Provenance Gathering in Distributed Scientific Workflows". In: *IEEE International Workshop on Scientific Workflows*, Los Angeles, California, United States.

Mattoso, M., Werner, C., Travassos, G., Braganholo, V., Murta, L., (2008), "Gerenciando Experimentos Científicos em Larga Escala". In: *SEMISH - CSBC*, Belém, Pará - Brasil.

Ogasawara, E., Murta, L., Werner, C., Mattoso, M., (2008), "Linhas de Experimentos: Reutilização e Gerência de Configuração em Workflows Científicos". In: *2 E-Science Workshop co-locado ao SBBD/SBES*, Campinas, Brasil.

Ogasawara, E., Paulino, C., Murta, L., Werner, C., Mattoso, M., (2009a), "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.

Ogasawara, E., Rangel, P., Murta, L., Werner, C., Mattoso, M., (2009b), "Comparison and Versioning of Scientific Workflows". In: *Proceedings of the 2009 international workshop on Comparison and versioning of software models*, Vancouver, Canada.

Oinn, T., Li, P., Kell, D. B., Goble, C., Goderis, A., Greenwood, M., Hull, D., Stevens, R., Turi, D., et al., (2007), "Taverna/myGrid: Aligning a Workflow System with the Life Sciences Community", *Workflows for e-Science*, Springer, p. 300-319.

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.

Pressman, R. S., (2004), *Software Engineering Software Engineering: A Practitioner's Approach*. 6 ed. McGraw-Hill; 6 edition.

Roure, D. D., Goble, C., Stevens, R., (2007), "Designing the myExperiment Virtual Research Environment for the Social Sharing of Workflows". In: *Proceedings of the 3rd IEEE International Conference on e-Science and Grid Computing*, p. 603-610, Bangalore, India.

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.

Srikant, R., Agrawal, R., (1996), "Mining Sequential Patterns: Generalizations and Performance Improvements". In: *Proceedings of the 5th EDBT*, p. 3-17

Taylor, I. J., Deelman, E., Gannon, D. B., Shields, M., (2006), *Workflows for e-Science: Scientific Workflows for Grids*. 1 ed. Springer.

Travassos, G. H., Barros, M. O., (2003), "Contributions of In Virtuo and In Silico Experiments for the Future of Empirical Studies in Software Engineering". In: *Proc. of 2nd Workshop on Empirical Software Engineering the Future of Empirical Studies in Software Engineering, Roma*

Zhao, Y., Dobson, J., Foster, I., Moreau, L., Wilde, M., (2005), "A notation and system for expressing and executing cleanly typed workflows on messy scientific data", *ACM SIGMOD Record*, v. 34, n. 3, p. 37-43.