Gerenciamento de workflows científicos em bioinformática

Luciano Antonio Digiampietri[1], João Carlos Setubal[2], Claudia Bauzer Medeiros[1]

[1]Instituto de Computação, Universidade Estadual de Campinas
Caixa Postal 6176, 13084-971, Campinas, SP, Brazil

[2]Virginia Bioinformatics Institute, Virginia Tech,
Bioinformatics 1, Box 0477, Blacksburg, VA, USA

{luciano,cmbm}@ic.unicamp.br, setubal@vbi.vt.edu

Abstract. *The proliferation of bioinformatics data and tools brings several challenges: how to understand and organize these resources, how to exchange and reuse successful experimental procedures, and how to provide interoperability among data and tools across different sites, for users with distinct profiles. Several approaches are being used to solve these problems, in general for some specific aspects, involving the use of scientific workflows. The solution developed in this thesis is also centered in scientific workflows, but covers all the investigated issues, thus, contributing for giving integrated support to users. The proposed infrastructure supports design, reuse, validate, share and document bioinformatics experiments*[1].

Resumo. *A proliferação de dados e ferramentas de bioinformática originou diversos desafios: como entender e organizar esses recursos, como compartilhar e reusar experimentos bem sucedidos, e como prover interoperabilidade entre dados e ferramentas de diferentes locais e utilizados por usuários com perfis distintos. Várias soluções vem sendo propostas para estes problemas, geralmente para alguns aspectos específicos, envolvendo freqüentemente o uso de workflows científicos. A solução oferecida pela tese também é centrada em workflows científicos, mas cobre todos os pontos levantados, contribuindo assim para dar um apoio integrado aos usuários. A infra-estrutura proposta permite projetar, reusar, validar, compartilhar e documentar experimentos de bioinformática*[1].

1. Introdução e Motivação

Workflows científicos [Wainer et al. 1996] estão sendo cada vez mais adotados como meios para especificar e coordenar a execução de experimentos que envolvem participantes em locais distintos. Eles permitem a representação e execução de tarefas que usam dados e ferramentas heterogêneos [Cavalcanti et al. 2005].

O projeto de workflows científicos é tipicamente manual, sendo uma atividade árdua e suscetível a erros. Além disso, em bioinformática, devido à constante evolução da área e a explosão combinatória de alternativas, há tantas possibilidades para a construção de workflows que é inviável computar e comparar todas elas. Assim, há uma crescente demanda por soluções que ajudem os cientistas a projetar os workflows desejados.

[1]A tese pode ser encontrada em: http://www.ic.unicamp.br/~luciano/digiampietri_tese.pdf

Outro aspecto importante está ligado à proveniência dos dados e ferramentas utilizados em cada experimento. Este tipo de informação é fundamental para que um experimento possa ser reproduzido, além de assegurar sua qualidade [Buttler et al. 2002]. Ambientes laboratoriais de bioinformática são muito dinâmicos. Para cada dado ou ferramenta, um usuário precisa conhecer *quando* o dado foi gerado, *quem* o produziu, *onde* e *como* o dado foi gerado. Para responder a essas perguntas um sistema deve, além de conter anotações sobre a proveniência dos dados, possuir o armazenamento detalhado da execução de cada workflow.

Todos esses requisitos computacionais – para projeto, gerenciamento e rastreabilidade – apresentam desafios em computação. A tese enfrentou esses desafios para facilitar o projeto, execução, reuso, validação e compartilhamento de experimentos científicos.

Os resultados foram publicados em dois congressos nacionais, três revistas nacionais, um congresso internacional e em quatro periódicos internacionais. O trabalho foi parcialmente financiado pela CAPES e por uma bolsa de dois anos da Microsoft Research (uma das duas *Microsoft Research Latin America Fellowships* concedidas em 2006). Durante o doutorado, foram feitos dois estágios internacionais na Microsoft Research, Washington, EUA – um no Grupo de Banco de Dados e outro no Grupo de Ciência Aplicada.

2. Aspectos de Pesquisa Envolvidos

Nossa pesquisa se concentrou em cinco aspectos: especificação de experimentos, anotação, integração de dados, interoperabilidade e rastreabilidade.

Especificação de experimentos. Na tese supomos que os experimentos científicos podem ser representados sob a forma de workflows científicos. Esta hipótese está sendo usada por diversos grupos (por exemplo, [Bausch et al. 2001, The myGrid Consortium]) para documentar ou executar esse tipo de experimento. No Laboratório de Sistemas de Informação (LIS: www.lis.ic.unicamp.br) da UNICAMP foi desenvolvida uma extensão do modelo de representação da *Workflow Management Coalition* (WFMC). Esta extensão permite a representação e o compartilhamento de workflows em vários níveis de abstração [Medeiros et al. 2005]. Partindo deste modelo, esta tese enfrentou os seguintes desafios:

1. Como complementar o modelo existente [Medeiros et al. 2005] de forma a agregar informações semânticas para facilitar o entendimento dos experimentos;
2. Como facilitar a integração de dados e interoperabilidade de ferramentas para a construção dos workflows supondo que os dados são armazenados em estruturas diferentes e as ferramentas não utilizam interfaces padronizadas;
3. Como adicionar dados de proveniência aos experimentos de forma a facilitar as consultas e possibilitar a rastreabilidade;
4. Como identificar as estratégias de composição de atividades que podem ser usadas para ajudar os diversos tipos de usuários a construírem seus workflows.

A especificação de um modelo de representação de workflows utilizou mecanismos da Web Semântica para enriquecer o conhecimento sobre os dados e as ferramentas. Já os mecanismos de composição de atividades foram desenvolvidos estendendo métodos de planejamento em Inteligência Artificial [Long and Fox 2003].

Anotação de dados e ferramentas. Um dos desafios na anotação de dados e ferramentas que são compartilhados entre diversos usuários é o fornecimento de um

vocabulário adequado (e compartilhado). Este desafio é abordado na tese a partir da construção e uso de ontologias. Ontologias estão sendo amplamente utilizadas como mecanismos que fornecem um vocabulário uniforme de conceitos e o relacionamento entre esses conceitos. O trabalho exigiu a criação de ontologias (de domínio e de serviços) que: (i) forneçam um vocabulário adequado para a anotação dos dados e ferramentas; (ii) auxiliem usuários a construir workflows, reusando total ou parcialmente soluções desenvolvidas por outros [Digiampietri et al. 2007c].

Integração de dados. O grande crescimento dos dados sobre genomas (seqüências de DNA, genes, etc) e das ferramentas disponíveis (muitas na forma de serviços Web) traz novos desafios: como aproveitar e integrar os dados disponíveis e prover interoperabilidade entre as ferramentas existentes. A anotação de dados seguindo conceitos de ontologias consensuais permite a redução da ambigüidade e facilita seu entendimento. Porém, laboratórios distintos costumam disponibilizar seus dados de maneira heterogênea, o que dificulta o compartilhamento e o reuso. A tese enfrentou este desafio combinando técnicas de mapeamento de estruturas de dados e o uso de ontologias [Digiampietri et al. 2007a].

Interoperabilidade. Nos últimos anos, diversas ferramentas para o processamento de atividades em bioinformática foram disponibilizadas na forma de Serviços Web. A falta de padronização das interfaces desses serviços e da descrição de suas funcionalidades dificulta seu uso. Para prover interoperabilidade entre os serviços, utilizamos uma estratégia composta de três abordagens: (i) a anotação de cada uma das operações de um serviço conforme nossa ontologia de serviços; (ii) a anotação da interface de cada operação (parâmetros e tipos de resultados produzidos) seguindo nossa ontologia de domínio; e (iii) o uso de algoritmos de casamento de interfaces para verificar a compatibilidade semântica e sintática entre interfaces.

Rastreabilidade é a habilidade de se rastrear o processo (dados e ferramentas) no qual um objeto está envolvido. Mecanismos de rastreabilidade são comumente encontrados em, por exemplo, trabalhos de engenharia de software, comércio eletrônico ou cadeias produtivas. Em bioinformática há dois propósitos para se prover rastreabilidade: assegurar a qualidade de um experimento e permitir consultas mais elaboradas sobre todos os dados e ferramentas envolvidos na produção de um resultado. O desafio quanto a este tema é o desenvolvimento de mecanismos que permitam tais tipos de rastreabilidade. A solução adotada foi adequar o modelo de representação de workflows [Medeiros et al. 2005] e proveniência [Barga and Digiampietri 2007] de forma a possibilitar a rastreabilidade e prover ferramentas que facilitem a navegação dentro de um experimento.

Com os resultados obtidos na pesquisa em cada um dos aspectos citados, especificamos e prototipamos uma infra-estrutura para o gerenciamento de experimentos científicos de bioinformática. Esta infra-estrutura ajuda a solucionar os problemas em aberto citados na motivação, estendendo as funcionalidades encontradas em outros sistemas com objetivos semelhantes.

3. Contribuições

O trabalho apresentado nesta tese resultou na definição da arquitetura apresentada na Figura 1. A arquitetura é composta por quatro camadas principais: repositórios; gerenciador de dados; módulos de processamento; e interface. Cada uma das publicações da

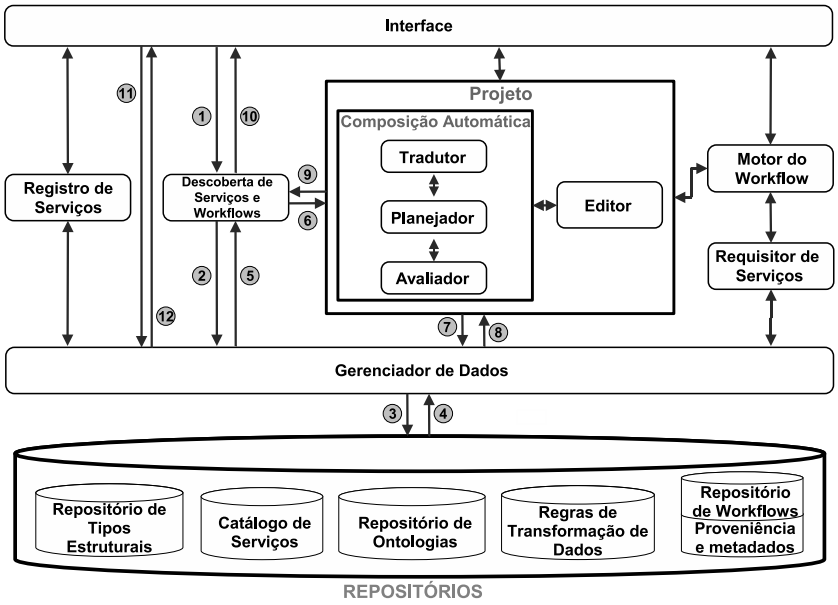

Figura 1. Arquitetura do sistema

tese contribui com partes desta arquitetura. A Figura 1 será utilizada para apresentar um cenário de uso e situar as contribuições e as publicações no contexto geral da tese.

3.1. Cenário de uso

Algumas das principais funcionalidades da solução proposta estão relacionadas à busca por serviços que satisfaçam às requisições do usuário, à composição automática de serviços, à consulta a dados de proveniência e à execução de workflows. Esta seção utiliza a Figura 1 para detalhar uma busca por serviços que aciona a composição automática.

Cenário de uso: cientista consulta o sistema para encontrar um serviço ou workflow desejado, mas não existe um serviço ou workflow que satisfaça os critérios de consulta. O sistema então constrói soluções e as sugere ao cientista. A Interface envia a requisição do usuário para o módulo de Descoberta de Serviços e Workflows (1). Este módulo encaminha a consulta para o Gerenciador de Dados (2) que consulta os repositórios de Ontologias, Workflows e o Catálogo de Serviços (3). Esta consulta não retornará nenhum serviço para o Gerenciador de Dados (4) que encaminha esta informação (5) ao módulo de Descoberta de Serviços e Workflows. Este módulo solicita ao módulo Composição Automática (6) a geração de novos workflows. Este módulo consulta o Gerenciador de Dados (7) para obter os dados necessários ao planejamento. O Gerenciador de Dados encaminha esta consulta para os repositórios: Catálogo de Serviços, Repositório de Ontologias, Regras de Transformação de Dados e Repositório de Workflows (3). O Gerenciador de Dados recebe os resultados da consulta (4) e os envia ao módulo Composição Automática (8). Com estes dados o módulo de Composição Automática projeta novos workflows e os envia para o módulo de Descoberta de Serviços e Workflows (9) que por sua vez os entrega a Interface (10), onde são apresentados ao usuário.

O principal objetivo da tese foi facilitar o trabalho de usuários em laboratórios de bioinformática (biólogos, bioinformatas e cientistas da computação) no que tange a construção, documentação, reuso e gerenciamento de experimentos, ferramentas e da-

dos. Este objetivo foi atingido tanto em ambientes centralizados quanto distribuídos. As soluções adotadas para os diversos desafios se baseiam em quatro eixos: (i) o uso de workflows científicos como a base para a especificação e execução de tarefas em um ambiente laboratorial distribuído; (ii) a adoção de ontologias consensuais, como forma de permitir compartilhamento de recursos, integração e interoperabilidade; (iii) o armazenamento, em bancos de dados, de ontologias e workflows, em diversos níveis de abstração, e (iv) o uso de planejamento em Inteligência Artificial (IA) para facilitar a construção automática de workflows (caixa rotulada Composição Automática, na Figura 1).

3.2. Aspectos de implementação

Implementamos três protótipos no decorrer desta tese para testar diferentes aspectos de nossa solução. Nesta seção descrevemos as características de cada um deles, contextualizando-os dentro de nossa arquitetura.

3.2.1. Extensão do WOODSS - composição automática

Estendemos uma versão do sistema gerenciador de workflows WOODSS (WOrk-flOw-based spatial Decision Support System) [Medeiros et al. 2005, Seffino et al. 1999] para tirar vantagens do uso de ontologias e do planejamento em IA como mecanismo de composição automática. Os módulos em cinza na Figura 2 correspondem aos módulos implementados neste protótipo.

A interface gráfica do protótipo foi desenvolvida em Java. O planejador utilizado foi o SHOP2 (www.cs.umd.edu/projects/shop), sendo que a definição de domínio do planejador é feita em Lisp. Construímos *scripts* em Perl e *bash* para ligar o planejador aos demais módulos do sistema. As ontologias de serviços e de domínio foram especificadas em OWL, usando a ferramenta Protégé e as ligações entre as instâncias dos serviços e as ontologias foram descritas em OWL-S (www.daml.org/services/owl-s).

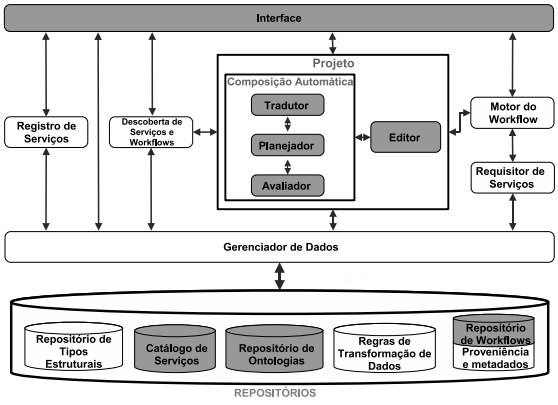

Figura 2. Protótipo 1 - Extensão do WOODSS para planejamento

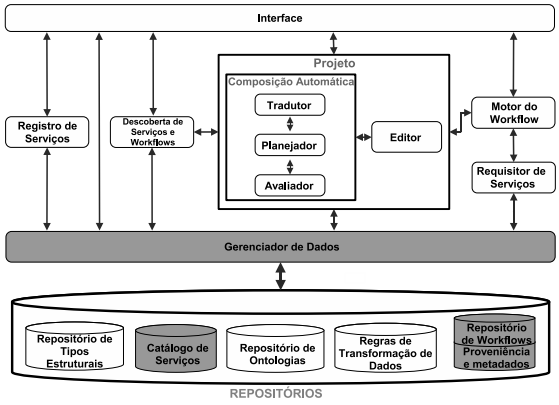

Figura 3. Protótipo 2 - Banco de dados de workflows e rastreabilidade

3.2.2. Banco de dados e rastreabilidade

O segundo protótipo corresponde ao banco de dados e sistema de consultas REDUX [Barga and Digiampietri 2006b] que foi especificado e implementado durante o

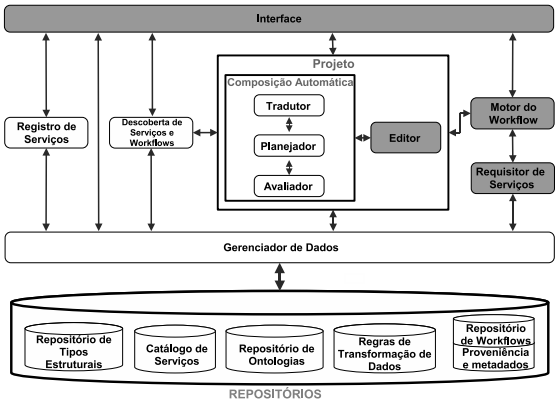

Figura 4. Protótipo 3 - Projeto e execução remotos de workflows

nosso primeiro estágio internacional na Microsoft Research. Os módulos em destaque na Figura 3 foram implementados neste protótipo, cujo objetivo principal foi testar experimentos em rastreabilidade. Além da especificação e implementação do banco de dados, que foi testado com dados biomédicos reais [Barga and Digiampietri 2006b], também implementamos mecanismos de consulta a proveniência e rastreabilidade. Este protótipo satisfez todos os requisitos propostos em um evento internacional sobre proveniência [First Provenance Challenge]. Utilizamos o sistema gerenciador de banco de dados SQLServer 2005 e a linguagem de programação C# para implementar a ferramenta de processamento de consultas.

3.2.3. Projeto e execução remotos de workflows

O terceiro protótipo foi desenvolvido para testar o projeto e a execução de workflows utilizando um navegador de Internet. Ele foi implementado durante o nosso segundo estágio internacional e está sendo testado por oceanógrafos da Universidade de Washington, EUA. Os módulos implementados neste protótipo estão destacados em cinza na Figura 4. Este protótipo utiliza o Windows Workflow Foundation [Andrew et al 2006] como motor de execução. O editor de workflows foi implementado em C# utilizando-se a tecnologia ASPX. O servidor Web utilizado foi o Internet Information Service 7.0.

4. Conclusões

Esta tese apresenta uma infra-estrutura para gerenciamento de experimentos em bioinformática. A pesquisa realizada combinou aspectos de bancos de dados, planejamento em Inteligência Artificial, gerenciamento de ontologias, padrões da Web Semântica e workflows científicos. Como resultado, apresentou soluções para os desafios de facilitar a especificação, reuso, documentação, validação e compartilhamento de experimentos científicos de bioinformática.

A infra-estrutura especificada e prototipada fornece mecanismos de integração de dados, composição de serviços e rastreabilidade para facilitar o gerenciamento de experimentos científicos. As principais contribuições desta tese são, desta forma:

- Proposta de uma solução para o problema da composição de serviços, combinando resultados da IA e de Bancos de Dados, de forma a ajudar o projeto de work-

flows científicos e documentar alternativas de projeto [Digiampietri et al. 2005, Digiampietri et al. 2006, Digiampietri et al. 2007b];

- Definição de um modelo de dados que combine o armazenamento de workflows em camadas com o armazenamento de proveniência de dados e ferramentas [Barga and Digiampietri 2006a, Barga and Digiampietri 2007]. Este modelo participou de um desafio proposto no evento *Provenance and Annotation of Data International Provenance and Annotation Workshop* (IPAW) [First Provenance Challenge] e satisfez a todos os requisitos propostos pelos organizadores, sendo capaz de, por exemplo: (i) representar sub-workflows, (ii) responder perguntas sobre proveniência de dados e ferramentas, e (iii) responpoder consultas sobre a execução de workflows.
- Uso de repositórios de ontologias para enriquecer a semântica na construção automática de workflows e facilitar o rastreamento da proveniência de dados e procedimentos [Digiampietri et al. 2007c];
- Especificação de mecanismos de integração de dados, interoperabilidade de ferramentas e rastreabilidade de experimentos científicos [Digiampietri et al. 2007a];
- Validação de parte da solução proposta pela implementação de três protótipos em e-Science [Barga and Digiampietri 2006a, Barga and Digiampietri 2007, Digiampietri et al. 2007a].

A solução apresentada foi projetada para problemas de montagem e anotação de genomas. Porém, ela pode ser utilizada em outros domínios bastando para isso: (i) a construção de ontologias de domínio e de serviços apropriadas; e (ii) o desenvolvimento de operadores específicos ao domínio adotado. Em particular, aplicamos um dos protótipos à bio-medicina e outro a oceanografia.

5. Possíveis Extensões

Há diversos trabalhos futuros previstos para esta tese. Dentre eles, há trabalhos teóricos de pesquisa, extensão da infra-estrutura adicionando novas funcionalidades e especificação de ontologias mais abrangentes ou que envolvam um domínio de aplicação diferente. Destacamos:

- Avaliação de outros tipos de planejadores além dos hierárquicos;
- Estudo de mecanismos para otimização na geração e execução de planos, tais como re-planejamento e reparo de planos;
- Extensão do escopo das ontologias para que descrevam contextos mais amplos em bioinformática, tais como genômica comparativa e vias metabólicas e especificação de ontologias em outros domínios, por exemplo, geoprocessamento.

Referências

Andrew et al, P. (2006). *Presenting Windows Workflow Foundation*. SAMS Publishing.

Barga, R. and Digiampietri, L. (2006a). Automatic generation of workflow provenance. In *Provenance and Annotation of Data International Provenance and Annotation Workshop (IPAW)*, volume 4145 of *Lecture Notes in Computer Science*. Springer.

Barga, R. and Digiampietri, L. (2006b). Redux - First provenance challenge. http://twiki.ipaw.info/bin/view/Challenge/REDUX (as of 2007-04-04).

Barga, R. and Digiampietri, L. (2007). Automatic capture and efficient storage of escience experiment provenance. *Concurrency and Computation: Practice and Experience*. DOI: 10.1002/cpe.1235.

Bausch, W., Pautasso, C., Schaeppi, R., and Alonso, G. (2001). Bioopera: Cluster-aware computing. In *Proceeding of the the 4th IEEE International Conference on Cluster Computing (Cluster)*, pages 90–94.

Buttler, D., Coleman, M., Critchlow, T., Fileto, R., Han, W., Pu, C., Rocco, D., and Xiong, L. (2002). Querying multiple bioinformatics information sources: can semantic web research help? *ACM SIGMOD Record*, 31(4):56–64.

Cavalcanti, M., Targino, R., Baiao, F., Rössle, S., Bisch, P., Pires, P., Campos, M., and Mattoso, M. (2005). Managing structural genomic workflows using Web services. *Data & Knowledge Engineering*, 53(1):45–74.

Digiampietri, L., Medeiros, C., and Setubal, J. (2005). A framework based in Web services orchestration bioinformatics workflow management. *Genetics and Molecular Research*, 4(3):535–542.

Digiampietri, L., Medeiros, C., Setubal, J., and Barga, R. (2007a). Traceability Mechanisms for Bioinformatics Scientific Workflows. In *Proceedings of the AAAI2007's Workshop on Semantic e-Science (SeS07)*, pages 26–33, Vancouver, Canada.

Digiampietri, L., Pérez-Alcazar, J., and Medeiros, C. (2007b). AI Planning in Web Services Composition: a review of current approaches and a new solution. In *VI ENIA - Proceedings of the XXVII Brazilian Computer Society Conference (CSBC2007)*.

Digiampietri, L., Pérez-Alcázar, J., and Medeiros, C. (2007c). An ontology-based framework for bioinformatics workflows. *International Journal of Bioinformatics Research and Applications*, 3(3):268–285.

Digiampietri, L., Setubal, J. C., and Medeiros, C. B. (2006). Bioinformatics scientific workflows: combining databases, AI and Web services. In *Proceedings of V Workshop de Teses e Dissertaes em Bancos de Dados (WTDBD)*, pages 2–9.

First Provenance Challenge. First Provenance Challenge. http://twiki.ipaw.info/bin/view/Challenge/ (as of 2006-11-13).

Long, D. and Fox, M. (2003). The 3rd International Planning Competition: Results and Analysis. *Journal of Artificial Intelligence Research*, 20:1–59.

Medeiros, C., Perez-Alcazar, J., Digiampietri, L., Pastorello, G., Santanche, A., Torres, R., Madeira, E., and Bacarin, E. (2005). WOODSS and the Web: Annotating and Reusing Scientific Workflows. *ACM SIGMOD Record*, 34(3):18–23.

Seffino, L., Medeiros, C., Rocha, J., and Yi, B. (1999). WOODSS - A Spatial Decision Support System based on Workflows. *Decision Support Systems*, 27(1-2):105–123.

The myGrid Consortium. myGrid: Middleware for in silico experiments in biology. http://www.mygrid.org.uk/ (as of 2008-02-02).

Wainer, J., Weske, M., Vossen, G., and Medeiros, C. (1996). Scientific Workflow Systems. In *The NSF Workshop on Workflow and Process Automation Information Systems*.