Gerência de workflows científicos: oportunidades de pesquisa em bancos de dados

Marta Mattoso, Sérgio Manuel Serra da Cruz
{marta,serra}@cos.ufrj.br

Agenda

- **Parte I**
  - Motivação
  - Objetivo
- **Parte II**
  - Workflows Científicos
  - Sistemas de Gerência de Workflows Científicos
- **Parte III**
  - Proveniência
- **Parte IV**
  - Demonstrações de SGWfC (Kepler, VisTrails, Mashups ...)
- **Pesquisas em andamento**

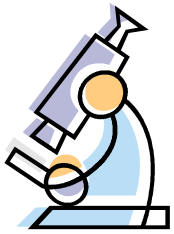
Agenda – Parte I

- **Introdução**
  - Motivação
  - Objetivo
- Vertentes do Trabalho
  - Desempenho
  - Processo de Gerência
  - Apoio Semântico
- Sumário
- Considerações – Parte I

e-Science

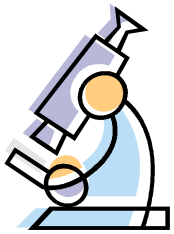
Apoio ao cientista para o desenvolvimento de ciência em larga escala utilizando infra-estrutura computacional correspondente

Experimentos Científicos



- Bancada
- Simulação
- Sensores
- ....

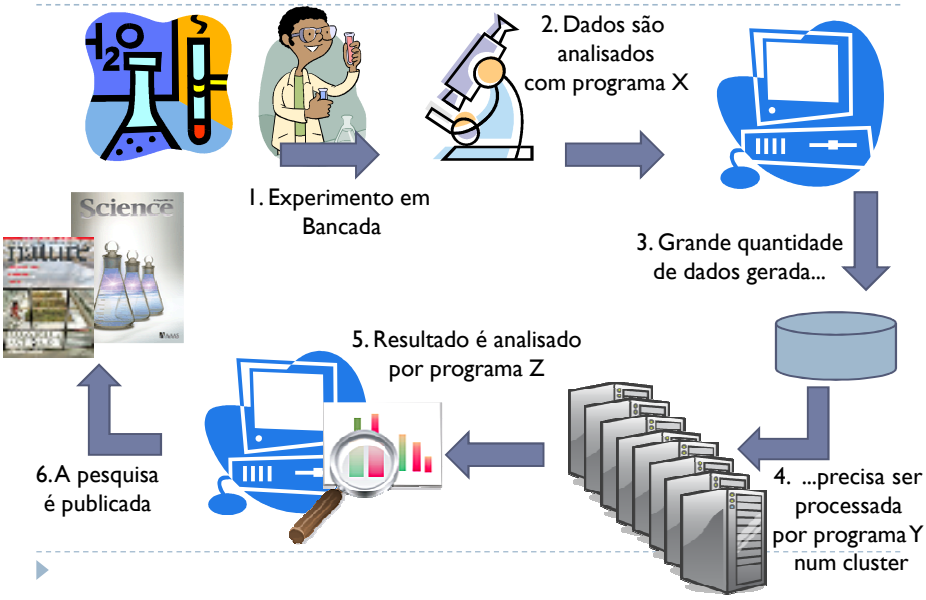
Experimentos Científicos



- Bancada
- Simulação
- Sensores
- ....

- **Enorme quantidade de dados a ser manipulada**

Um Cenário Típico de Experimento



Problemas deste cenário

- A seqüência de ações está na "cabeça" do cientista
  - Ele executa os programas em uma determinada seqüência
  - Ele move os dados de um ambiente para o outro
- Experimentos não são reproduzíveis de forma automática
- Necessidade de acesso a ambientes de grades e clusters
  - Procedimentos bastante complexos para cientistas sem experiência computacional
- Necessidade de compartilhar os resultados com outros pesquisadores

Necessidade...

- Sistema que gerencie a composição de processos e dados num fluxo coerente (workflow científico)
  - que registre as etapas executadas e parâmetros de cada etapa
  - independente do local de execução
  - que associe resultados aos processos e dados que os geraram (proveniência)
  - que trafegue por diferentes ambientes de execução (distribuição, paralelismo, heterogeneidade)
  - que faça uso eficiente dos recursos computacionais disponíveis
  - que permita a comparação de diferentes resultados de um mesmo experimento

[myGrid, Goble 2007]

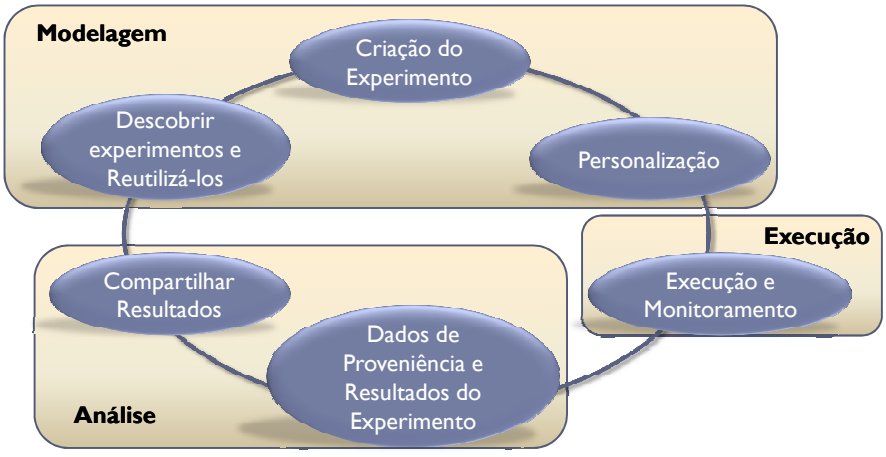
Inúmeros Sistemas de Gerência de Wf

- Apple's Mac OS X Automator
- Askalon
- Bigbross Bossa
- Bea's WLI
- BioPipe
- BizTalk
- BPWS4J
- Breeze
- Carnot
- Con:cern
- **DAGMan**
- DiscoveryNet
- Dralasoft
- Enhydra Shark
- Filenet
- Fujitsu's i-Flow
- **GenFlow**
- GridAnt
- Grid Job Handler
- GRMS (GridLab Resource Management System)
- GWFE
- GWES
- IBM's holosofx tool
- IT Innovation Enactment Engine
- ICENI
- Inforsense
- Intalio
- jBpm
- JIGSA
- JOpera
- **Kepler**
- Karajan
- Lombardi
- **MathWS**
- Microsoft WWF
- NetWeaver
- Oakgrove's reactor
- ObjectWeb Bonita
- OFBiz
- OMII-BPEL
- Open Business Engine
- Oracle's integration platform
- OSWorkflow
- OpenWFE
- Q-Link
- **Pegasus**
- Pipeline Pilot
- Platform Process Manager
- **P-GRADE**
- PowerFolder
- **PtolemyII**
- Savvion
- Seebeyond
- Staffware
- Sonic's orchestration server
- **Swift**
- ScyFLOW
- SDSC Matrix
- SHOP2
- **Taverna**
- Triana
- Twister
- Ultimus
- Versata
- **VisTrails**
- WebMethod's process
- wftk
- XFlow
- YAWL Engine
- Yahoo Pipes
- WebAndFlo
- **WebIOS**
- Wildfire
- Werkflow
- wfmOpen
- WFEE
- **WOODSS**
- Zbuilder
- **E muito mais....**

Ciclo de Vida de um Experimento

Fonte: myGrid (Goble e De Roure, 2007)

Ciclo de Vida de um Experimento



Ênfase na Análise

Fonte: myGrid (Goble e De Roure, 2007)

No contexto deste tutorial

Ciclo de Vida do Experimento Científico

=

Apoio à

Modelagem + Execução + Análise

de processos científicos de forma encadeada e controlada

Tal apoio visa a ...

- Facilitar a concepção do experimento

Modelagem

- Apoiar a execução distribuída com eficiência
- Tornar o experimento reproduzível

Execução

- Permitir rastreamento (proveniência) dos processos e dados que levaram a um determinado resultado
- Apoiar o compartilhamento e compreensão do processo científico por diversos pesquisadores

Análise

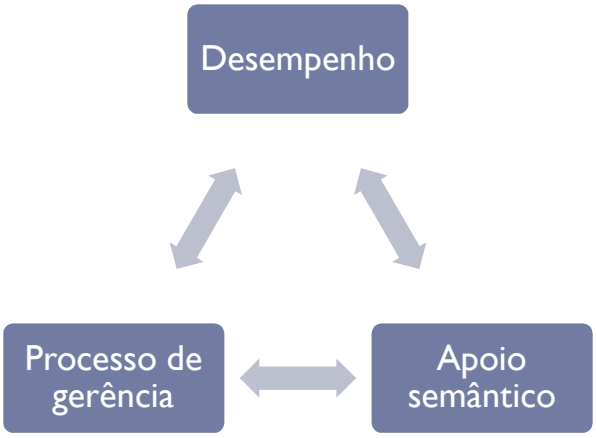
Objetivo deste tutorial

- Identificar e discutir desafios necessários para prover apoio computacional ao desenvolvimento de ciência em larga escala, com foco em
  Gerência de Recursos Científicos
  Gerência de Sistemas de Workflow
  Gerência de Informações de Proveniência

Agenda – Parte I

- Introdução
  - Motivação
  - Objetivo
- **Vertentes de Pesquisa em BD**
  - Desempenho
  - Processo de Gerência
  - Apoio Semântico
- Sumário
- Considerações sobre a Parte I

Vertentes de Pesquisa em BD



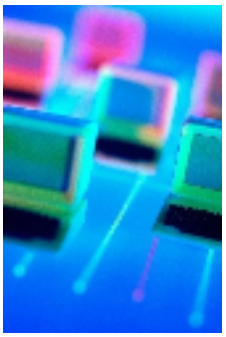
Vertente **Desempenho**

- Foco no 1° Grande Desafio da SBC
  - "Gestão da Informação em grandes volumes de dados multimídia distribuídos"
- Estratégia
  - Dividir para conquistar
- Problema
  - Dados e processamento situados em diferentes locais
  - Heterogeneidade necessária dos SGWfC para viabilizar distribuição

Vertente **Desempenho**

Cientista

Rede de baixa velocidade

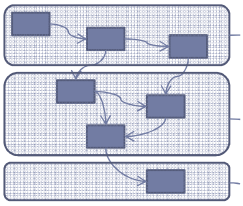


Ambiente de *Grid*
(rede de alta velocidade)

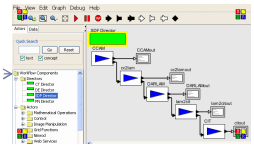
Como trafegar grandes quantidades de dados e tirar proveito do ambiente de Grid de modo simples ?

Vertente **Desempenho**

Cientista



Workflow Científico



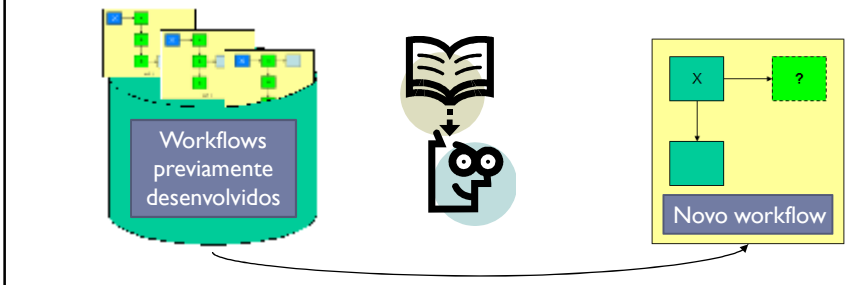
Como viabilizar que um workflow seja executado de forma distribuída por diferentes SGWfCs?

Diferentes SGWfC

Vertente **Processo de Gerência**

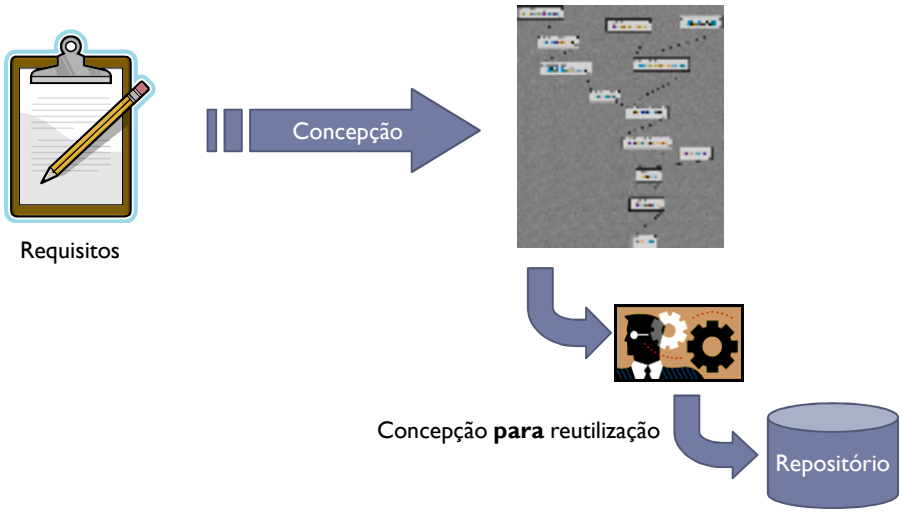
- Foco no 2° Grande Desafio da SBC
  - "Modelagem computacional de sistemas complexos artificiais, naturais e sócioculturais e da interação homem-natureza"
- Estratégia
  - Aplicar Engenharia de Software (Reutilização e Modelagem Ágil) sobre os processos de concepção de workflows científicos
- Problema
  - Como estender técnicas de reutilização e modelagem ágil para diminuir o grau de retrabalho e aumentar a qualidade na concepção de workflows científicos?

Vertente **Processo de Gerência**

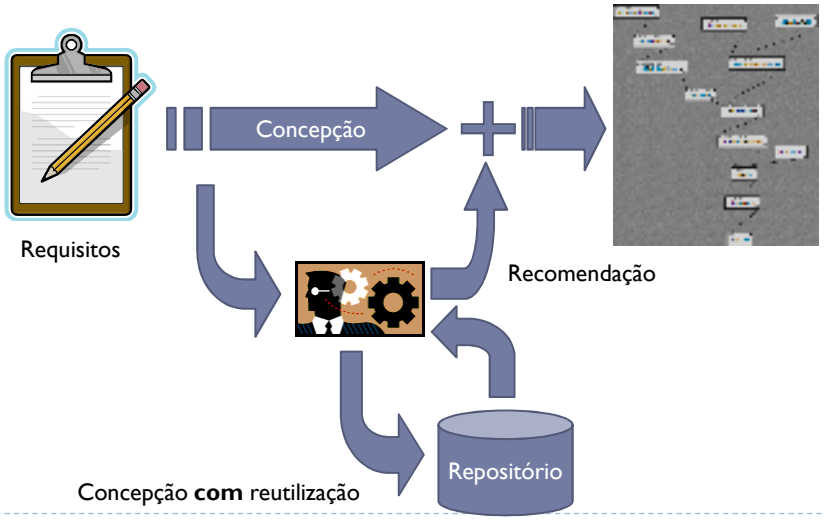


Conhecimento passado pode ser útil na concepção de novos workflows?

Vertente **Processo de Gerência**



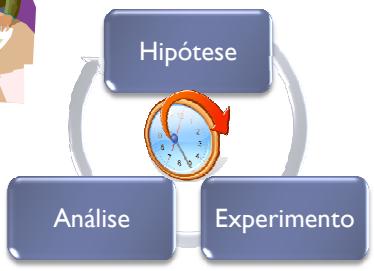
Vertente **Processo de Gerência**



Vertente **Apoio Semântico**

- Foco no 2° Grande Desafio da SBC
  - "Modelagem computacional de sistemas complexos artificiais, naturais e sócio-culturais e da interação homem-natureza"
- Estratégia
  - Aplicar Engenharia de Software (Gerência de Configuração e Rastreabilidade) sobre os workflows científicos
- Problema
  - Como estender técnicas de gerência de configuração e rastreabilidade para produzir dados de proveniência dos experimentos científicos?

Vertente **Apoio Semântico**



Ad hoc e com diversas iterações!

É possível repetir o experimento?

Como explicar as conclusões?

Como comparar resultados?

Vertente **Apoio Semântico**

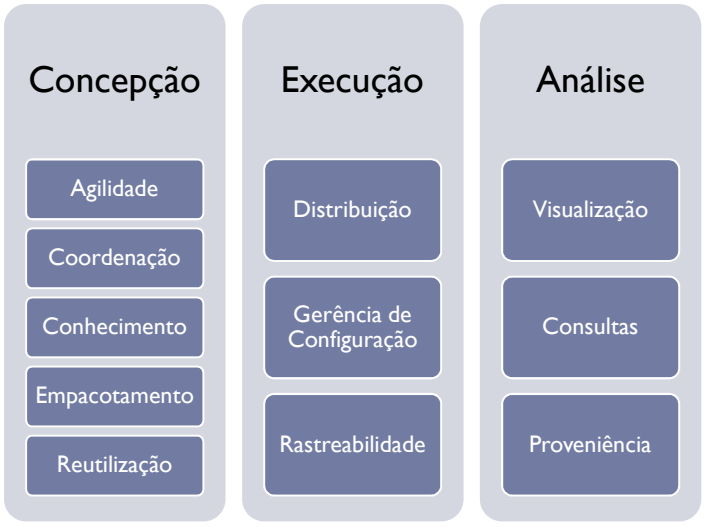
Agenda – Parte I

- Introdução
  - Motivação
  - Objetivo
- Vertentes do Trabalho
  - Desempenho
  - Processo de Gerência
  - Apoio Semântico
- **Sumário**
- Considerações – Parte I

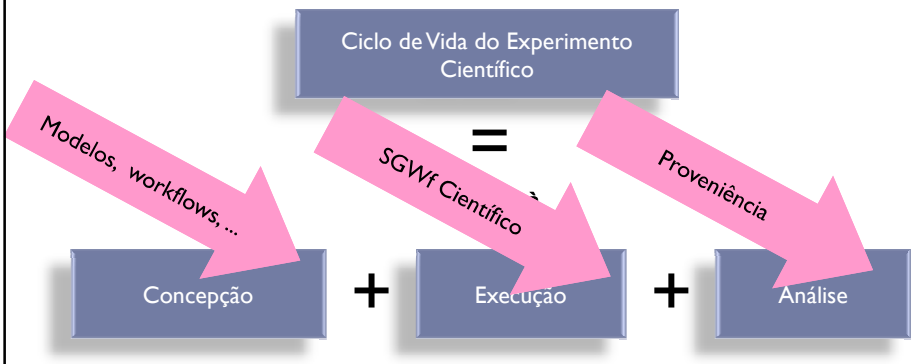
Sumário

- Resumo dos pontos nos quais acreditamos que as pesquisas atuais podem trazer ganhos
- Categorizados de acordo com o ciclo de vida de um experimento científico
  1. Concepção de workflows
  2. Execução de workflows
  3. Análise de workflows

Sumário- apoio ao ciclo de vida



No contexto deste tutorial



de processos científicos de forma encadeada e controlada

Agenda – Parte I

- Introdução
  - Motivação
  - Objetivo
- Vertentes do Trabalho
  - Desempenho
  - Processo de Gerência
  - Apoio Semântico
- Sumário
- **Considerações – Parte I**

Considerações – Parte I

- Vertentes complementares a outras iniciativas de apoio a E-Science
- No BR
  - IC-UNICAMP- *WOODSS ,WeBIOS*
  - IC-UFJF- *MathWS*
  - IME-USP- *GenFlow*
  - *COPPE-UFRJ – GExp*
  - *FIOCRUZ / UFRJ / IME-RJ/ UFF – BioWebDB*

GExp- Gerência de Experimentos científicos em larga escala

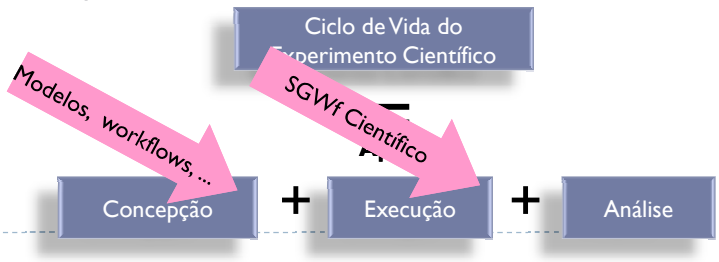
Considerações – Parte I

- Vertentes com lementare ativas de apoio a E-Sc
- No BR
  - UNICA
  - IME/US
  - UFJ
  - *UFRJ – C*
  - *FIOCRUZ/ UF bDB*
- **E-Science Workshop , SBBD/SBES 2008**
  5ª. feira – 16/10/2008

10:30 – 12:00 Palestra –
Is e-Science more than
"e + Science" ?
Cláudia Bauzer Medeiros
- IC/UNICAMP

Agenda – Parte II

- **Concepção**
  - Recursos científicos
  - Modelos, Algoritmos, Programas, Dados, Workflows, Experimentos
- **Execução**
  - Sistemas de Gerência de Workflows
  - Gerência de Experimentos
- Sumário



Alguns Conceitos

- **Modelos**
  - Conjunto de idéias que descrevem um processo natural
  - Algoritmos
- **Programas**
  - implementação computacional de um modelo
- **Dados**
  - Obtidos em redes de sensores
  - Gerados por mecanismos de "coleta"
  - Gerados por programas científicos
- **Metadados**
- **Workflows**
  - Combinação de modelos ou programas e dados
- **Experimentos**
  - Conjunto de especificações/execução de workflows

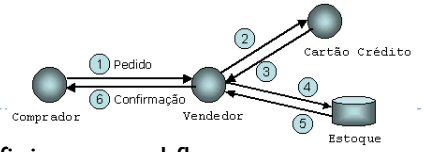
Workflow (WfMC, 1995)

"A automação de um processo de negócio, completo ou apenas parte dele, através do qual, documentos, informações ou tarefas são transmitidas de um participante a outro por ações, de acordo com regras procedimentais."

- Provê a abstração necessária para descrever uma série de processos estruturados e suas atividades, oferecem um contexto robusto de resolução de problemas e promovem o uso efetivo e otimizado dos recursos computacionais
- Cada programa participante opera segundo um conjunto de regras definidas *a priori*
  - Recebem conjuntos de dados, processam e enviam para o próximo programa.

39

Workflow



- Mais formalmente, podemos definir um workflow como uma coleção de atividades organizadas para acompanhar algum experimento (processo de negócio).
- As **atividades** ou **tarefas** são os componentes de software independentes que implementam alguma funcionalidade e são executadas por um ou mais sistemas de softwares. Exemplos de atividades incluem executar um programa, transformar um arquivo ou atualizar um banco de dados.
- Um workflow define a **ordem de execução** dessas atividades ou as **condições** em que essas atividades serão executadas e a sua eventual sincronização.
- Os **dados** de entrada e saída das atividades (variáveis) são definidos como o fluxo de dados do workflow.

40

Definição de workflow

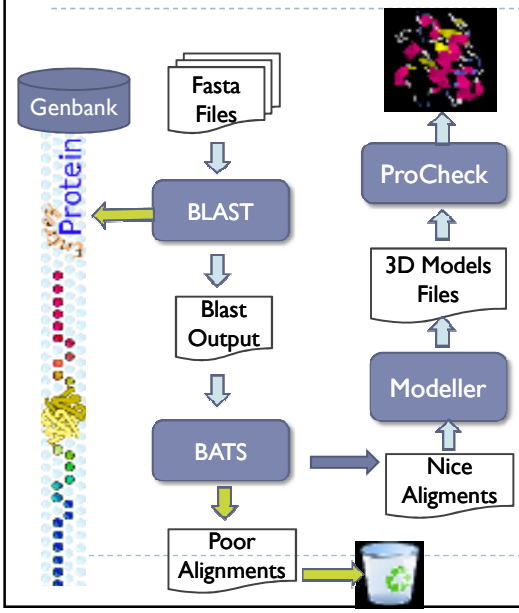
- Resumindo, um workflow *W* é representado pela quádrupla *(T,V, Sf, Cf)* onde:

- *T* é um conjunto {*t1, t2,. . . ,tn*} de tarefas de W,
- *V* é um conjunto de variáveis {*v1, v2,. . . ,vn*} de *W* definindo um fluxo de dados,
- *Sf* é uma função sucessora associada a cada tarefa $t \in T$, e
- *Cf* é uma função de condição associada a cada tarefa $t \in T$.

SGWf e Máquina de Wf

- Um sistema de gerência de workflows – SGWf (WfMS - *Workflow Management Systems*)
  - Software que fornece toda a infra-estrutura para definir, executar e monitorar workflows.
- Máquina de Workflow
  - Na utilização deste tipo de software é importante para separar a definição do workflow, da **máquina**(*engine*) encarregada de executar o mesmo, permitindo dessa forma que modificações em uma das partes não afete o funcionamento da outra.

Workflow Científico - bioinformática



- Conjunto de dados de entrada a serem processados
- Programas legados e de terceiros
- Tempo de execução dos programas desbalanceado
- Tamanhos de conjuntos de dados distintos
- Dependências entre os programas
- Ambientes computacionais distintos

Workflows: Científicos vs. Comerciais

Científicos	Comerciais
Centrados em dados	Centrados em controles (ex."BPEL")
Manipula vastas coleções de dados e recursos computacionais distribuídos	Manipula "Poucos dados", orientados a tarefas
Dinâmicos, recursos não são conhecidos a priori	Estáticos, poucas alterações
Stateless	Statefull
Dados: heterogêneos, distribuídos, não estruturados	Dados: estruturados
Não Lineares e de natureza exploratória	Lineares, foco nos "Biz procs"
Monitoramento, controles de execução, mudanças *ad-hoc*	
Natureza colaborativa, sub-workflows	
Execução parametrizada, aplicações legadas	

Rastreabilidade, segurança, desempenho, confiabilidade, reprodutibilidade,

"Focados em resolver diferentes problemas, mas…existem oportunidades de "fertilização cruzada" "(Davidson e Freire, 2008)

Workflows: Científicos vs. Comerciais

Científicos	Comerciais
Concepção dinâmica- o processo de especificação é interativo, a busca da melhor seqüência de execução de tarefas nem sempre é clara. Necessita da execução do wf ao longo da especificação.	Concepção estática- o processo de especificação deve ser finalizado antes da execução.
Na execução, um mesmo cientista executa várias vezes num mesmo dia o mesmo workflow mudando apenas alguns parâmetros.	São executados várias vezes, mas normalmente invocados por usuários diferentes, como por exemplo, os vários clientes que fazem um pedido em uma loja.
A definição de um workflow envolve a tomada de diversas decisões, análises e trabalho de equipe	A definição de um workflow é modelada para atender a um processo de negócio relativamente fixo, como processar um pedido, verificar o estoque, enviar a mercadoria e emitir a fatura.

Características desejáveis dos SGWf científicos (1)

Característica	Descrição
Interface	Desenho intuitivo, voltado para o usuário final. Detalhes de implementação (baixo nível) devem ser escondidos, foco no nível conceitual
Reuso	Apresentar componentes reutilizáveis e intercambiáveis, idealmente devem ter capacidade de adicionar/remover novos processos dinamicamente
Transformação de dados	Permitir consecutivas transformações de dados entre as atividades
Interação e batch	Prover acompanhamento "process steering" (play, pause e stop) durante a execução do workflow.
Monitoração	Monitorar processos em tempo de execução mesmo em segundo plano, máquinas e ambientes distintos
Distribuição	Permitir processamento local e/ou distribuído
Dados	Apoiar transferências de dados (moderada a intensa)

(Addis (2003)Altintas et al. (2003) e Sangeeta (2005)

Características desejáveis dos SGWf científicos (2)

Característica	Descrição
Flexibilidade	Facilitar alterações na descrição do workflow (coleções de dados e programas)
Complexidade	Manipular fluxos de dados complexos, controles e eventos.
Desempenho e planejamento	Informar o desempenho e os custos de execução. Capaz de coletar dados de diferentes processos e usar métricas para predizer os tempos de execução
Tolerância a falhas	Alta disponibilidade e tolerante a falhas. Execução parcial.
Verificação/ Validação	Verificar e validar a construção ou importação de workflows
Proveniência	Rastreio dos dados/processos utilizados e gerados em cada etapa.
Acesso a "como foi feito"	Consultar dados de proveniência: de onde vieram os dados, onde e quais as transformações que foram feitas neste dado e como cada resultado foi obtido é fundamental para a análise final.
Segurança, etc...	

(Addis (2003)Altintas et al. (2003) e Sangeeta (2005)

O SGWf deve gerenciar

pelo menos 2 níveis de abstração!

- ***Workflow Abstrato*** - descreve um dado workflow sem especificar quais recursos serão utilizados na execução, oferece grande flexibilidade libera o projetista das preocupações relacionadas com os detalhes de implementação, criando uma camada capaz de representar o "comportamento" do workflow.
- ***Workflow Concreto*** – intimamente ligado à uma dada tecnologia, associa quais recursos computacionais são requeridos para a execução das tarefas do workflow.

Especificação do workflow em SGWf

2 tipos de representação!

- ***DAG*** - Contém estruturas (controle, comunicação e dados) do tipo seqüenciais, paralelas ou livres. Workflow é um grafo, onde cada tarefa é um nó e cada nó pode ter um número arbitrário de filhos. As seqüências definem a ordem de execução das tarefas.
- ***Non-DAG*** – Incluem estruturas de iteração (também conhecido como *loops* ou *ciclos*). Elas permitem que os workflow executem tarefas ou sub-workflows de forma repetida.

Especificação do workflow

Alguns formalismos

- ***Padrões de controle de fluxo*** – van der Aalst
- ***Redes de Petri*** -
- ***Álgebra de processos*** – *J. E. Ferreira*

- Permitem validações, equivalências e diversas propriedades, porém não há um padrão, abrangência/simplicidade

[myGrid, Goble 2007]

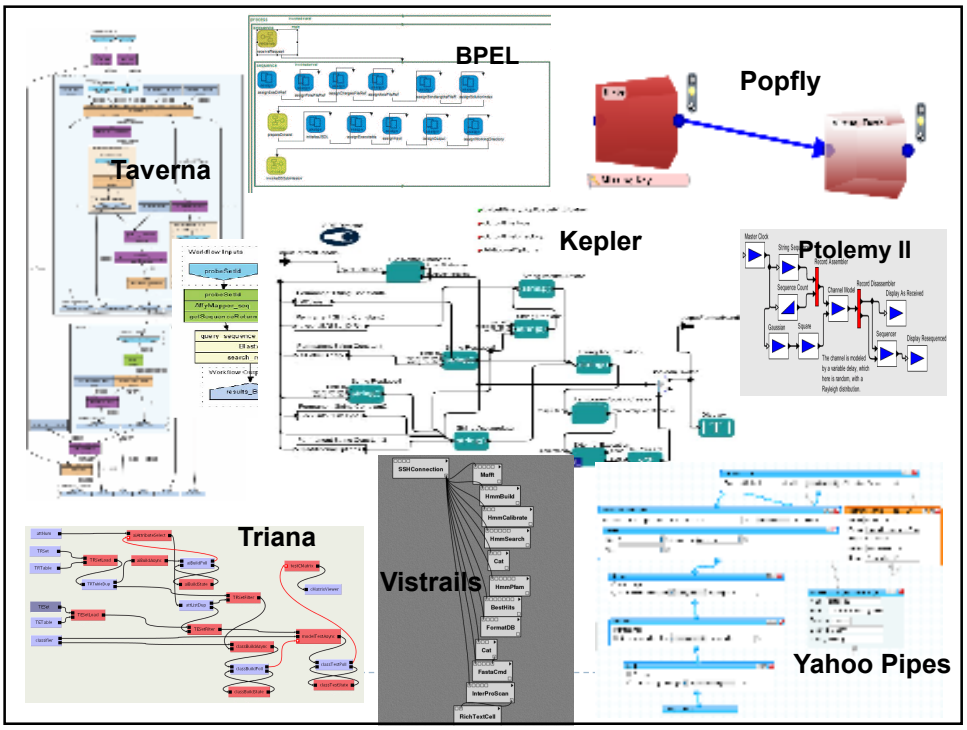
Inúmeros Sistemas de Gerência de Wf

- Apple's Mac OS X Automator
- Askalon
- Bigbross Bossa
- Bea's WLI
- BioPipe
- BizTalk
- BPWS4J
- Breeze
- Carnot
- Con:cern
- **DAGMan**
- DiscoveryNet
- Dralasoft
- Enhydra Shark
- Filenet
- Fujitsu's i-Flow
- **GenFlow**
- GridAnt
- Grid Job Handler
- GRMS (GridLab Resource Management System)
- GWFE
- GWES
- IBM's holosofx tool
- IT Innovation Enactment Engine
- ICENI
- Inforsense
- Intalio
- jBpm
- JIGSA
- JOpera
- **Kepler**
- Karajan
- Lombardi
- **MathWS**
- Microsoft WWF
- NetWeaver
- Oakgrove's reactor
- ObjectWeb Bonita
- OFBiz
- OMII-BPEL
- Open Business Engine
- Oracle's integration platform
- OSWorkflow
- OpenWFE
- Q-Link
- **Pegasus**
- Pipeline Pilot
- Platform Process Manager
- **P-GRADE**
- PowerFolder
- **PtolemyII**
- Savvion
- Seebeyond
- Staffware
- Sonic's orchestration server
- **Swift**
- ScyFLOW
- SDSC Matrix
- SHOP2
- **Taverna**
- Triana
- Twister
- Ultimus
- Versata
- **VisTrails**
- WebMethod's process
- wftk
- XFlow
- YAWL Engine
- Yahoo Pipes
- WebAndFlo
- **WebIOS**
- Wildfire
- Werkflow
- wfmOpen
- WFEE
- **WOODSS**
- Zbuilder
- **E muito mais....**

Linguagens de Definição de Workflows

Dezenas!

- Diagramas, grafos, documentos XML e outros formatos .
- de acordo com van der Aalst *et al*, independente da quantidade de linguagens existentes, diversas permitem especificar os padrões principais de modelagem de workflows, o que poderia dar origem a modelos canônicos para o qual essas linguagens poderiam ser convertidas e mapeadas.



Concepção- linguagem não é suficiente

Protocolo científico

1. "...para cada alinhamento gerar modelos HMM (com o hmmbuild).
2. Calibrar com o hmmcalibrate e posteriormente utilizar para uma busca por similaridade contra bancos de proteínas dos cinco protozoários retirados do GenBank.
3. Buscar com o hmmsearch, e utilizar e-value de corte de 0.1.

Taxonomias de Workflows

- Classificação do ***workflows comerciais*** (Georgakopoulos *et al.* (1995) e Plesums (2002), Leyman & Roller, 1999)
  - Ad-hoc
  - Administrativos
  - Produção
  - Colabotativos

  Baseados em processos

- Weske, Vossen, Medeiros (1996) e Singh e Vouk (1998) adicionam um quarto tipo ***workflows científicos***
  - Auxiliam os cientistas desde a concepção do experimento até análise/visualização dos dados experimentais
- Cao et al (2003), Slominski, Gannon, Fox (2004) Yu e Buyya (2005) caracterizam os ***grid workflows***

55

Tipos de Sistemas de Gerência de Wf

- Web- controle centralizado, ênfase em semântica:
  - VisTrails, Taverna, Kepler, WOODSS, ...
- Grades- controle distribuído, ênfase em desempenho:
  - Swift, Triana, Pegasus, Askalon, P-Grade, ...
- Resumindo
  - falta distribuição aos centralizados (execução remota de subworkflow, captura remota de proveniência, etc.
  - falta semântica aos distribuídos (proveniência, compartilhamento, etc.
  - **o SGWf não apoia o ciclo de vida do experimento**

[GridAsia07, Goble]

SGWf... um caldeirão de tecnologias

Muitos Modelos…

- Computação – control flow, data flow, pipelines baseados em scripts
- Recursos– alocação de tarefas, delegação, Bag-of-tasks
- Interação– interativos, batch
- Representação de Wf – DAG, non-DAG
- Execução- DAG, funcional
- Data types –tipagem forte, fraca
- Adaptabilidade – dinâmicos, estáticos
- Tolerância a falhas– quando as coisas vão mal…

Existem muitos SGWf, cada um tem seus prós e contras

Muitos Ambientes…

- Componentes – recursos de grid, aplications, serviços
- Tipo de Usuários – cientistas, especialistas, leigos
- Natureza- open-source, comerciais
- Escala– long running, large data sets, data staging, data streaming, etc…

Muitas tecnologias….

centralizado, distribuído, local, web…

Workflows Científicos – na literatura

- Muitos workshops e conferências
- Edições especiais em revistas:
  - SIGMOD 2005, JOGC 05, SciProg 06, CCPE 07, DPDJ 08, etc.
- Keynote de Bill Gates no SC 2005
- NSF Workshop - Challenges of Scientific Workflows (2006)
- Provenance Challenges (wiki)
- WWWF, CCGRID, IPAW SBBD-E-Science 08, ICWS08 etc.

Em suma – recursos científicos

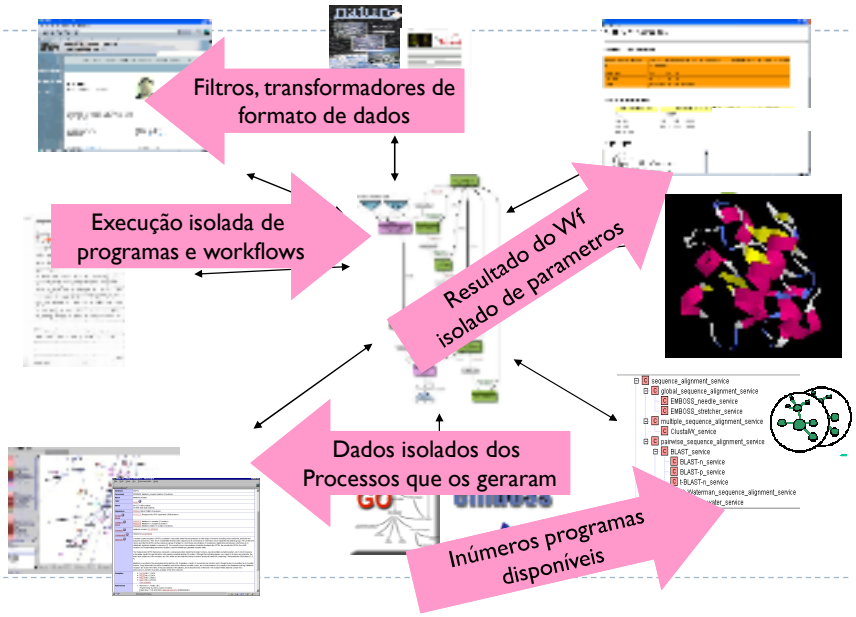
- **Modelos**
  - Conjunto de idéias que descrevem um processo natural
  - Algoritmos
- **Programas**
  - implementação computacional de um modelo
- **Dados**
  - Obtidos em redes de sensores
  - Gerados por mecanismos de "coleta"
  - Gerados por programas científicos
- **Metadados**
- **Experimentos**
  - Conjunto de especificações/execução de workflows

Toda a ênfase nos workflows "isolados" !

Com os recursos isolados …

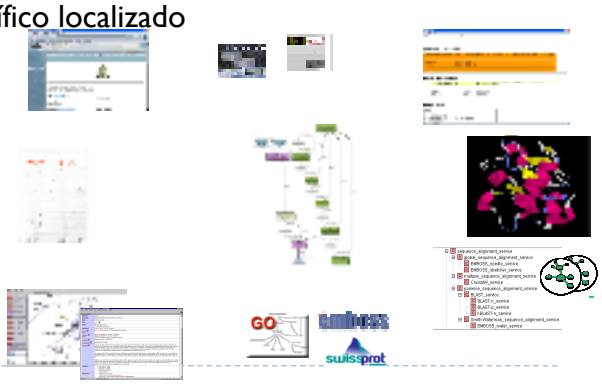
- Um wf científico encadeia programas e dados, mas...
  - não associa programas com algoritmos ou modelos
  - não associa programas genéricos com executáveis específicos
- Um SGWf científico gerencia a execução de um wf, mas...
  - não associa diversas execuções a uma única definição abstrata de um wf
  - não associa variações de um wf a um mesmo experimento científico
- Um repositório de experimentos
  - não associa o experimento à definição do wf

E o relacionamento entre conceitos ?



Recursos Científicos isolados não “escalam”

- Re-trabalho, Inconsistências
- Relacionamento implícito
- Sem compartilhamento de experiências
- Conhecimento científico localizado
- Perda de informação



Como Apoiar o Ciclo de Vida de um Experimento ?

Fonte: myGrid (Goble e De Roure, 2007)

BD pode ajudar

- Modelando relacionamento entre recursos
  - ...
- Modelando proveniência
  - ...
- Usando ontologias para associações
  - Enriquecimento Semântico - Achar e Usar
  - ...

ES pode ajudar

- Testes de software e planejamento de experimentos
  - ...
- DBC- reuso de componentes
  - ...
- Gerência de configuração
  - ...

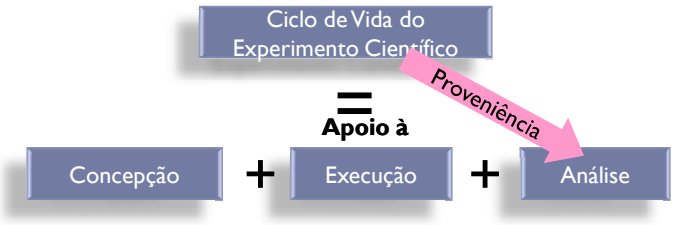
Sumário – Parte II

Levantamento dos principais conceitos ligados à

- **Concepção:**
  Modelos, algoritmos, programas, dados, metadados, workflows, experimentos
- **Execução:**
  SGWf científico
- **Panorama atual –**
  - **Inúmeros SGWf científicos, mas**
  - **Ausência de ferramental para Gerência de Experimentos Científicos em Larga Escala**

Agenda – Parte III

- **Análise**
  - Curadoria
  - Proveniência
  - Sistemas de Proveniência
  - OPM – *Open Provenance Model*
- Sumário



Curadoria de dados

[Day, 2008]

" Representa a manutenção e adição de valor a um conjunto de informações digitais confiáveis para uso imediato ou futuro; relaciona-se especialmente com a manutenção ativa e a revisão dos dados ao longo do ciclo de vida da produção acadêmica ou de materiais científicos"" (http://www.dcc.ac.uk/)

- Curadoria de dados é um processo contínuo que mantém os dados íntegros ao longo do tempo
- Agrega valor através de **anotações** em:
  - Dados
  - Serviços
  - Workflows
  - Experimentos

Curadoria Digital é o planejamento, manutenção e preservação de dados digitais ao longo do tempo!

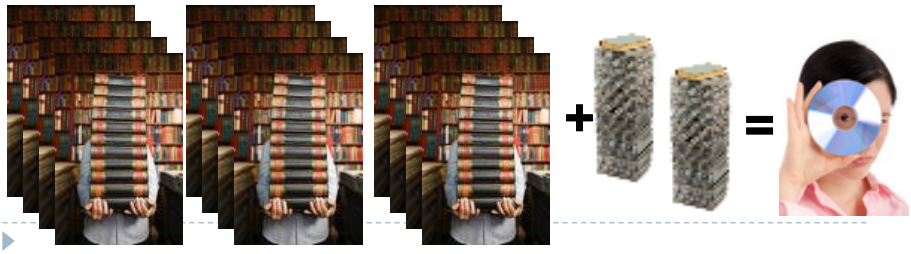
Benefícios da curadoria digital em e-Science

- **Curto prazo**
  - Acesso imediato a dados confiáveis
  - Dados de melhor qualidade
  - Uso de padrões comuns em diferentes *datasets*, ampliando as oportunidades de colaboração
  - Facilita verificações de autenticidade, ampliando a confiabilidade nos dados
  - Aumenta as possibilidades de compartilhamento de dados e análises

- **Longo prazo**
  - Fornece informações sobre o ***contexto do experimento*** e ***dados de proveniência***
  - Encoraja ***reuso e compartilhamento*** dos dados
  - Preserva e protege contra perdas e obsolescência (crucial quando os dados do experimento não podem ser reproduzidos ou são extremamente valiosos)
  - Uso de ferramentas e serviços para migrar dados, metadados e outras representações da informação para novos formatos, assegurando sua perpetuação.

Por que a curadoria de dados científicos é importante em e-Science?

- Faz parte do processo de pesquisa
- Agrega valores *Intrínsecos* e *Extrínsecos* à pesquisa
- Aumenta o potencial de criar "novos conhecimentos" a partir de dados pré-existentes
- Crescentes exigências dos órgãos de fomento
- Preserva o contexto da descoberta científica



Proveniência – no dicionário

- Muitas definições:
  1. (i) ato ou efeito de proceder. (ii) ***lugar de onde alguém ou algo procede.*** (iii) origem. (iv) proveniência (Dic. Aurélio)
  2. (i) lugar de onde alguma coisa provém. (ii) ***fonte, origem***, procedência (Dic. Michaelis)
  3. (i) the fact of coming from some particular source or quarter; origin, derivation. (ii) the history or pedigree of a work of art, manuscript, rare book etc.; *a* ***record of the ultimate derivation and passage*** of an item through its various owners. (Dic. Oxford)
  4. (i) the origin, source; (ii) ***the history of ownership*** of a valued object or work of art or literature (Dic. Merrian-Webster)

Proveniência de dados

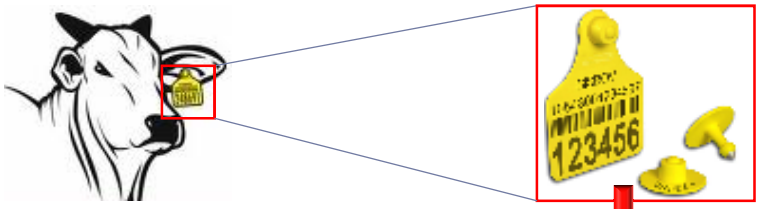
- O termo proveniência de dados (*data provenance*) se refer
fontes de consultas ou a serviços baseados no processame

"dado que descreve dado" é metadado
Proveniência descreve dado
Logo...
Proveniência é metadado.

a literatura que poderiam ser interpretados como
proveniência de dados: *afiliation, data genealogy, data set dependence, data archeology, audit trail* e *derivation history.*

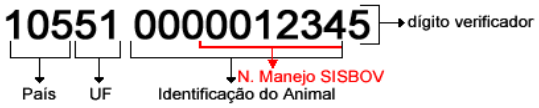
"Proveniência" nas artes

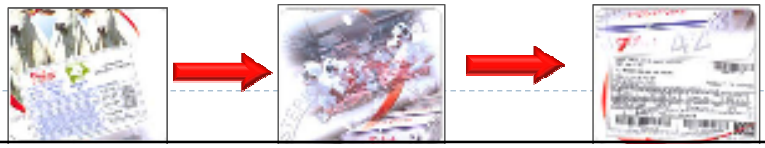
Proveniência no dia-a-dia



Fonte: www.abczcertificadora.com.br/identificacao.php

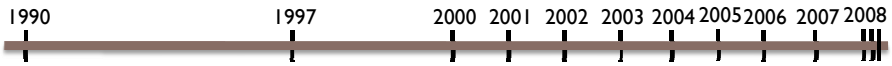
Número SISBOV





Proveniência em BD

- Rastrear mudanças na base de dados, no esquema, etc.



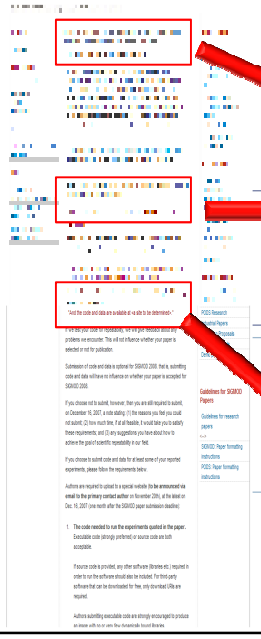
- A Polygen Model for Heterogeneous Database Systems: The Source Tagging Perspective. Y. R. Wang and S. E. Madnick. VLDB 1990.
- Supporting Fine-grained Data Lineage in a Database Visualization Environment. A. Woodruff and M. Stonebraker. ICDE 1997.
- Tracing the Lineage of View Data in a Warehousing Environment. Y. Cui, J. Widom and J. L. Wiener. TODS 2000.
- Why and Where: A Characterization of Data Provenance. P. Buneman, S. Khanna, T. ICDT 2001.
- …
- On the Expressiveness of Implicit Provenance in Query and Update Languages. P. Buneman, J. Cheney and S. Vansummeren. ICDT 2007.
- Intensional Associations Between Data and Metadata. D. Srivastava and Y. Velegrakis. SIGMOD 2007.
- Provenance Semirings. T. J. Green, G. Karvounarakis and V. Tannen. PODS 2007.

A lista **NÃO** é completa !!!

75

ACM SIGMOD VANCOUVER, CANADA 2008

Falta proveniência na Computação?



ACM SIGMOD/PODS Conference: Vancouver, 2008
SIGMOD: Guidelines for Research Papers

EXPERIMENTAL REPEATABILITY REQUIREMENTS

Papers that are accepted and are verified this way will be eligible to include the following text in the proceedings:

"The results in this paper were verified by the SIGMOD repeatability committee"

If verified code and data is also made available for archiving, the following phrase may be added:

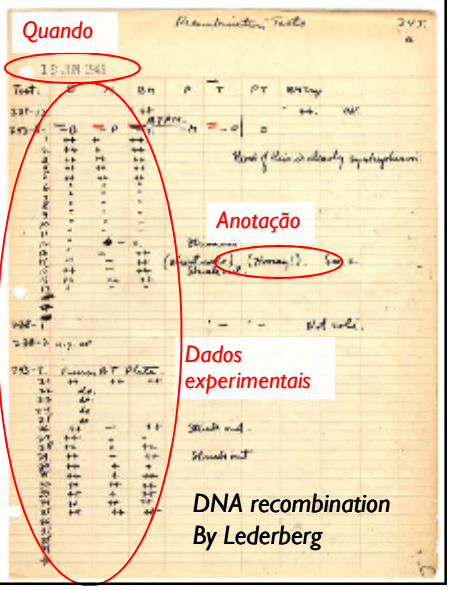
"And the code and data are available at <a site to be determined>."

http://www.sigmod08.org/sigmod_research.shtml

[Davidson, Freire, 2008]

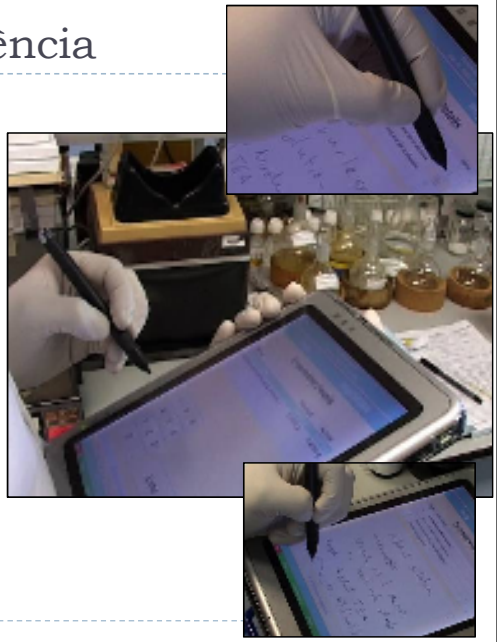
Importância da proveniência na Ciência

- Seu uso ***não é*** novo!
- "Lab notebooks" são usados a muito, muito tempo
  - Auxiliam na reprodução dos resultados
  - São evidências em disputas de patentes
- O que há de novo?
  - Grande volumes de dados
  - Análises muito + complexas
- Anotação manual não é mais uma boa opção
  - Cientistas precisam de métodos sistemáticos para capturar a proveniência de dados



Proveniência na Ciência

- Seu uso não é novo!
- "Lab notebooks" são usados a muito tempo
  - Auxiliam na reproduçaõ dos resultados
  - São evidências em disputas de patentes
- O que há de novo?
  - Grande volumes de dados
  - Análises muito + complexas
- *Anotação manual* não é mais uma boa opção
  - Cientistas precisam de métodos sistemáticos para *capturar* e *compartilhar* a proveniência de dados



[SmartTear, 2004 e myTea,2005]

Proveniência na Ciência

- Seu uso não é novo!
- "Lab notebooks" são usados a muito tempo
  - Auxiliam na reproduçaõ dos resultados
  - São evidências em disputas de patentes
- O que há de novo?
  - Grande volumes de dados
  - Análises muito + complexas
- ***Anotação manual*** não é mais uma boa opção!
  - Cientistas precisam de métodos sistemáticos para ***capturar*** e ***compartilhar*** a proveniência de dados

[SmartTear, 2004, myTea,2005, CombeChem, 2008]

Ainda falta proveniência na Ciência do séc. XXI?

YES

NO T ENOUGH PROVENANCE

nature

www.nature.com/nature Vol 442 | Issue no. 7098 | 6 July 2006

Illuminating the black box

Note to biologists: submissions to *Nature* should contain complete descriptions of materials and reagents used.

This journal aims to publish papers that are not only interesting and thought-provoking, but reproducible and useful. In order to do this, novel materials and reagents need to be carefully described and readily available to interested scientists.

That might seem obvious. But despite the efforts of our editors and referees, papers in the biological sciences are still being submitted — and occasionally published — that do not adequately describe the reagents used. Unless efforts are redoubled to eliminate this practice, we could see an era of 'black box' biology, in which outside researchers cannot work out what was done in an experiment.

Some of these 'rogue reagents' are the offspring of well-established technologies that have fallen under commercial control. A frequent example involves antibodies, reagents used throughout biology. The standard for publication should be that the authors can show an antibody's specificity and utility. But this information is sometimes missing from published papers, and readers must rely on a description from a company with a stake in the antibody in question.

But the problem is most acute in the case of new technologies, which sometimes experience a period of rapid development during which the rules just don't seem to apply. The desire to exploit the latest innovation can outstrip a field's ability to impose a common standard for description and documentation.

Strictly business

An example is RNA interference (RNAi), which allows researchers to manipulate gene expression. The critical reagents in RNAi studies are short interfering RNAs, the sequences of which determine which genes are affected. The approach has unleashed a frenzy of papers describing RNAi strategies that illuminate questions in biological systems and could have commercial application.

Manuscripts submitted from the RNAi field show a pattern typical of such new technologies. As the technology becomes commercialized, it may array tips. *Nature* journals have reviewed too many submissions with reagents incompletely described, or where material-transfer agreements limited availability.

Authors face conflicting pressures. Journals set publication standards for openness and availability, but many researchers have personal business interests, universities want to turn research into patent revenue, and biotechnology companies want to keep as much as possible under wraps.

In one case, an author published a paper without sequences of the RNAi reagents used. Editors and referees missed this. A year or so after publication, the author could not meet requests for her reagents, or provide the sequences for his short interfering RNAs. When the journal editors learned of this they asked for the sequences to be added to the paper. But the author said that patent issues and a business deal prevented him from doing so. The author's lab could not produce the short interfering RNAs, and a company that he had established didn't want the author to reveal the sequences, as this would jeopardize its *raison d'être*. This kind of stalemate matters, because it prevents the replication of experiments and inhibits the selection of appropriate controls in subsequent work.

Some authors claim replication is possible without full sequence information or the details of novel compounds. They say that the materials in question are for sale, enabling anyone to duplicate the paper. This misses the point. Scientific progress revolves around producing data that can drive the next stage of investigation. Even if consistent results can be achieved with a black-box reagent, knowing the composition and sequence can be the difference between making an observation and drawing insightful conclusions about it.

False premise

Another specious claim is that putting the structures of chemical compounds or sequences for short interfering RNAs in a manuscript is unnecessary, because if appropriate controls are described other investigators will know how to control their experiments. This is a false premise: the controls for an experiment designed to test one hypothesis are likely to be different from those needed to tackle another problem. To design appropriate controls one must know what to control for — information that comes only from structures or sequences.

"The desire to exploit the latest innovation can outstrip a field's ability to impose a common standard for description and documentation."

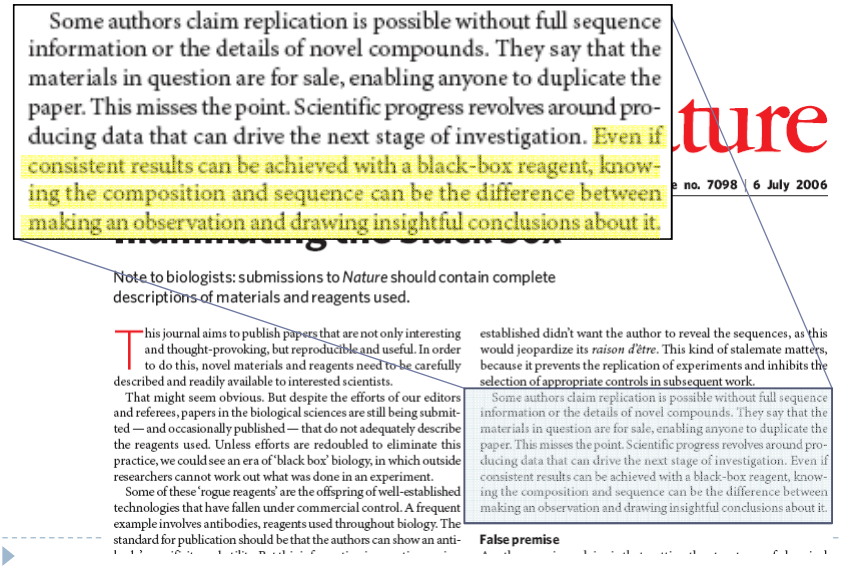
A troubling situation can also arise when a paper is published and the reagent provided, but a project ending or a person moving and a this provision. Journals have to decide whether to push for continued provision, and, if so, how. Is there a sensible statute of limitations on an author's commitment to provide? Input from the community would help publishers make a sensible policy; comments can be sent to author@nature.com.

In the meantime, new authors must, when they are arranging the commercialization of their work, bear in mind the implications that these deals may have on their freedom to publish to the standards that the community is entitled to expect.

Nature will continue to require that, at acceptance, materials should be fully described in the paper or in the online Supplementary Information. If limited disclosure is noted between acceptance and publication, refusal to provide information may forfeit any offer to publish. Non-disclosure or inaccessibility by ought to the editor's attention after publication will be investigated and actions taken, including, if necessary, public rebukes posted online and appended to the archived paper. Authors are expected to make reasonable arrangements, described accurately in the published work, by which readers can obtain the reagents described in their papers. ■

Nature, v. 442, no. 7098, pp. 1, 2006

This journal aims to publish papers that are not only interesting and thought-provoking, but reproducible and useful. In order to do this, novel materials and reagents need to be carefully described and readily available to interested scientists.



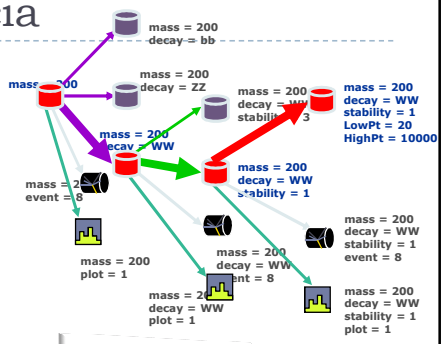
Importância da proveniência na Ciência

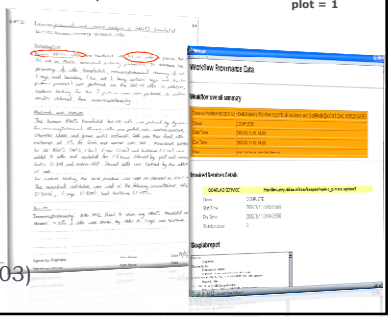
- Útil na ***interpretação e reprodução*** dos resultados
- Útil para ***a compreensão do experimento e da cadeia de eventos*** usados na produção de um resultado
- Permite v***erificar se um experimento foi realizado de acordo com procedimentos científicos*** aceitáveis
- ***Identifica os dados*** de entrada de um experimento e sua origem
- Contribui para assegurar a ***qualidade dos dados***
- ***Rastreio*** de *quem* executou o experimento e quem foi o responsável pelos resultados

Proveniência é tão (ou mais) importante quanto os resultados (Davidson, Freire, 2008)

Formas da proveniência

- Derivação
  - Representação de um caminho, workflow, script ou consulta.
  - Associação entre itens (grafos)
  - Geralmente é uma explicação (2W1H =when, who, how)
  - Centrada no processo
- Anotação
  - Anexado aos items ou coleções de dados (estruturado ou semi ou texto livre)
  - Geralmente é uma explicação (5W1H = why, when, where, who, what, how)
  - Centrada no dado





Fonte: Knowledge and Provenance (Goble, 2003)

As 2 granularidades da proveniência

- Coarse-grained (workflow provenance):
  - Registra ***a história da execução de um workflow*** e/ou da derivação de conjuntos de dados
  - Casos típicos em SGWf
    - Registro das sequências de atividades executados pelos SGWf
      - Pode incluir os uso de dispositivos externos, por ex, sensores, cameras, satélites ou outros dispositivos ou métodos de coleta de dados
    - Algumas etapas sao tratadas como caixa-preta
- Fine-grained (data provenance):
  - Registra a ***história do item de dado*** em um dataset
    - Descreve como o item de dado é movimentado através de uma "rede de BDs"
    - **Maiores informações no tutorial Provenance in Databases (Bunemann & Tan, SIGMOD, 2007)**

85

[Davidson, Freire, 2008]

Captura de dados de proveniência

- 3 Níveis de Captura, mas mecanismos distintos

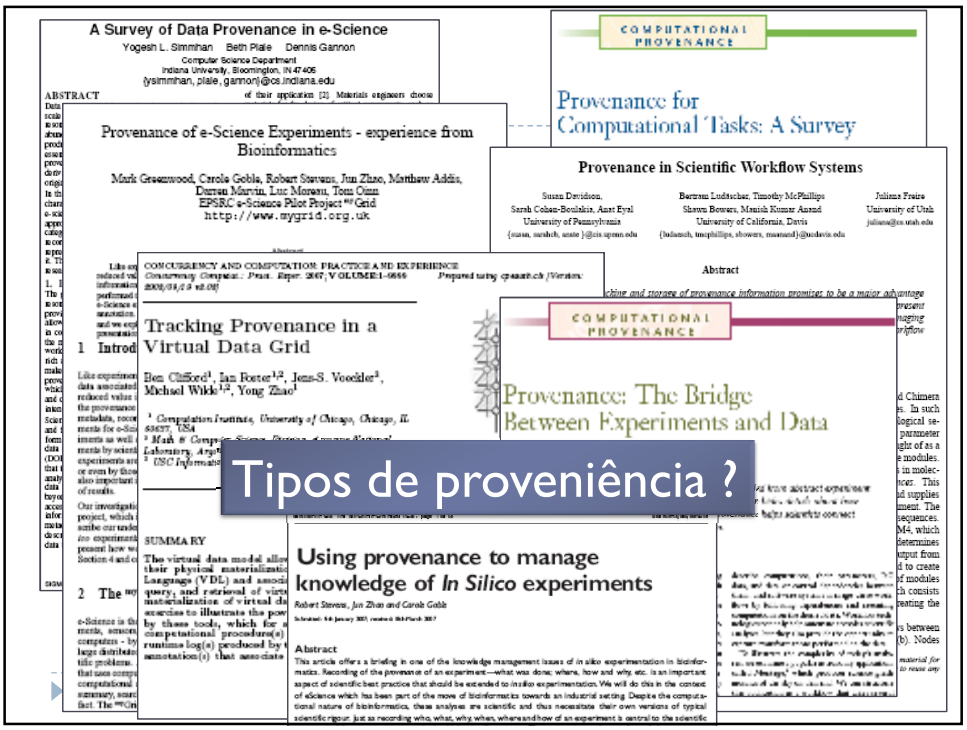
	Workflows	Baseado em Processo	Sist. Operacional
Causa	Na especificação do workflow	Na especificação do processo	Na especificação do processo
Requer	Captura "manual" em cada novo workflow, requer wrapper	Captura "manual" em cada novo workflow, requer wrapper	Captura "automática", não requer wrapper
Tipo Captura	Centralizado	Distribuído	Centralizado

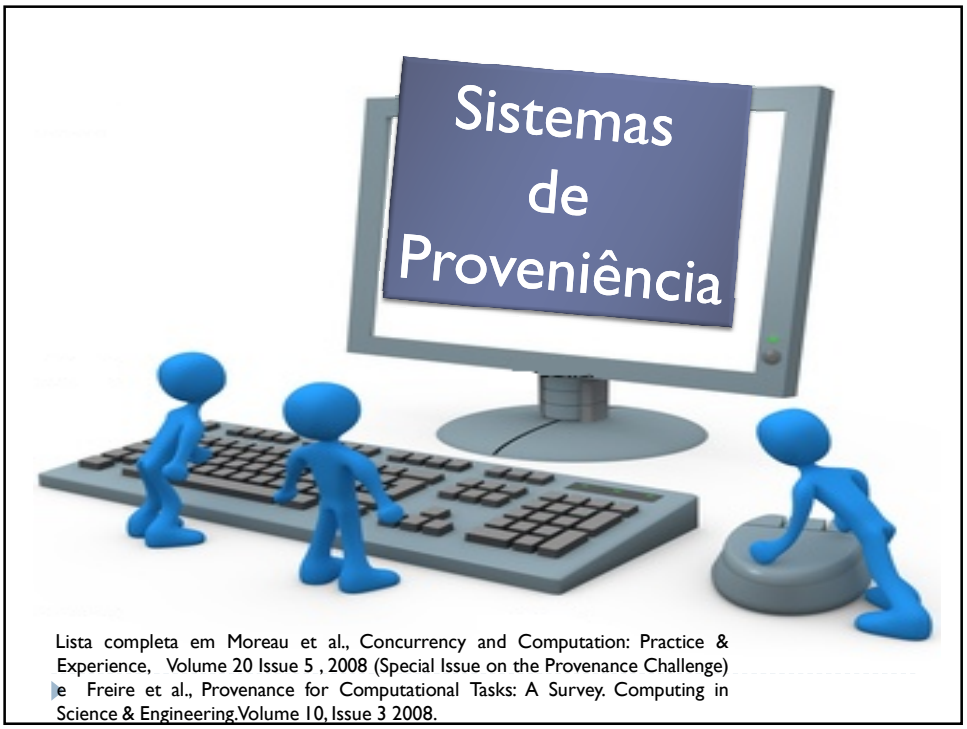
Comparação completa em Moreau et al., CCPE 2008 (Special Issue on the Provenance Challenge) e Freire et al., CISE 2008.

Tipos de proveniência?

- Cliford, Foster,Voeckler, Zhao (2007)
  - Program structure
  - Runtime logs
  - Annotation
- Freire, Koop, Santos, Silva (2008)
  - Prospectiva
  - Restrospectiva
- Stevens, Zhao, Goble (2007)
  - Process Level
  - Data Level
  - Organizational Level
  - Knowledge Level

- Wild (2005)
  - Virtual data
- Simmhan, Plale, Gannon (2005)
  - Taxonomia para classificar os sistemas de proveniência





Lista completa em Moreau et al., Concurrency and Computation: Practice & Experience, Volume 20 Issue 5 , 2008 (Special Issue on the Provenance Challenge) e Freire et al., Provenance for Computational Tasks: A Survey. Computing in Science & Engineering.Volume 10, Issue 3 2008.

PASOA - www.pasoa.org

- Mecanismo de Captura
  - Cada serviço registra proveniencia em um repositório
- Modelo de Dados
  - Prov. Retrospectiva → Grafos e p-structure/XML
  - Prov. Prospectiva → Não há
- Execução do Wf não é identificada,
- Registra Relacionamentos entre itens de dados
- Descoberta da Proveniência é realizadas através dos relacionamentos

Karma - http://www.extreme.indiana.edu/karma

- Mecanismo de Captura
  - Cada serviços publica a proveniência para um serviço central
  - Baixo acoplamento e arquitetura publish/subscriber
- Modelo de Dados
  - Prov. Retrospectiva → Graph represented in XML stored in an RDBMS;
  - Prov. Prospectiva → BPEL
- Coleta a proveniência de processos e dados
- Workflows dinâmicos (meteorologia)
  - Usa a proveniência para verificar e validar os resultados das simulações
- Sobrecarga na captura é minimizada

Pegasus - http://pegasus.isi.edu

- Mecanismo de Captura
  - workflow engine
- Modelo de Dados
  - Prov. Prospectiva → OWL (Wings);
  - Prov. Retrospectiva → Relational
- Focos
  - Criação e Execução de wf sobre recursos distrib. Ex. Grid
- Prov. Retrospectiva capturada peloVDS
  - Armazenada em SGBD
  - *Traduz a proveniência de wf abstract para concreto*
- *2 mecanismos de consulta*
  - SPARQL → sobre os workflows
  - SQL → sobre os dados de execução

Swift - http://www.ci.uchicago.edu/swift/

- Mecanismo de Captura
  - workflow engine
- Modelo de Dados
  - Prov. Prospectiva → XML;
  - Prov. Retrospectiva → Relational
- Focus
  - Criação e Execução de wf sobre recursos distrib. Ex. Grid
  - Usa e expande o VDS
  - Escalonamento e otimização em Grid
    - workflow engine é o **Karajan**
    - Faz seleção de sítios e mvimentação de dados
    - Faz Caching, pipelining, clustering, load balancing
    - Fault tolerance, exception handling

Taverna - http://www.mygrid.org.uk

- Mecanismo de Captura
  - workflow engine
- Modelo de Dados
  - Workflows are serialized in Scufl--an XML dialect
  - Retrospective provenance represented in RDF and stored in MySQL
- Foco
  - Bioinformática e Web services
- Usa as tecnologia de semantic web e ontologies disponíveis na bioinformática

"Workflow has to reflect the experiment not the services invocation interface"

Kepler - http://www.kepler-project.org

- Mecanismo de Captura
  - workflow engine
- Modelo de Dados
  - XML/MOML
- Proveniência
  - COMAD
  - RWS
  - Alterar o código das Classes de Provenance

+ detalhes a seguir no live Demo

Vistrails - http://www.vistrails.org

- Mecanismo de Captura
  - workflow engine
- Modelo de Dados
  - XML
- Proveniência
  - Prospectiva
  - Retrospectiva

+ detalhes a seguir no live Demo

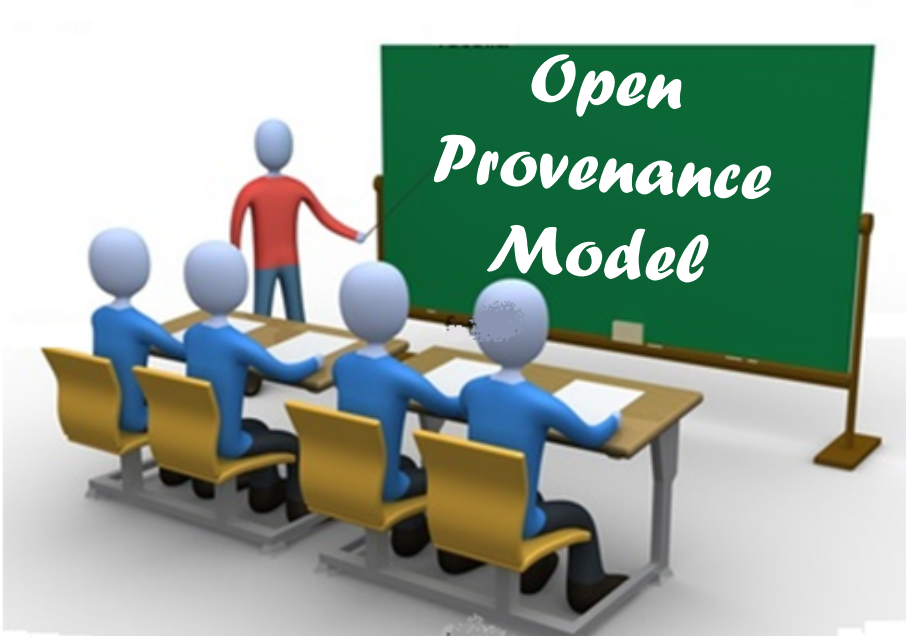
Sistemas de Proveniência: - Sumário

		Data Model					
System	**Capture Mechanism**	**Prospective Provenance**	**Retrospective Provenance**	**Workflow Evolution**	**Storage**	**Query support**	**Available as open-source**
REDUX	workflow-based	Relational	Relational	No	RDBMS	SQL	No
Swift	workflow-based	SwiftScript	Relational	No	RDBMS	SQL	Yes
VisTrails	workflow-based	XML and Relational	Relational	Yes	RDBMS and files	Visual QBE, specialized language	Yes
Karma	workflow- and process-based	BPEL	XML	No	RDBMS	Proprietary API	Yes
Kepler	workflow-based	MOML	MOML variation	Under development	Files, RDBMS planned	Under development	Yes
Taverna	workflow-based	Scufl	RDF	Under development	RDBMS	SPARQL	Yes
Pegasus	workflow-based	OWL	Relational	No	RDBMS	SPARQL for metadata and workflow; SQL for execution log	Yes
PASS	OS-based	n/a	Relational	No	Berkeley DB	nq-proprietary query tool	No
ES3	OS-based	n/a	XML	No	XML database	XQuery	No
PASOA/PReServ	process-based	n/a	XML	No	File system, Berkeley DB	XQuery, Java query API	Yes

[Freire et al., CISE 2008]

Resumo (1/2)

- Os sistemas de captura de proveniência possuem modelos de casualidade, mas são incompletos
  - O quê, Quem, Quando, Onde
- Diferentes modelos de dados e armazenamento
  - OPM ?
- Proveniência não é o foco dos sistemas
  - Ela é conseqüência dos sistemas e não a sua causa
- Consultas de Proveniência são limitadas e a as interfaces são complexas



OPM

- O modelo ainda está em discussão, é um draft.
- Requisitos fuincionais:
  - Habilitar a troca de informações de proveniência entre sistemas
  - Permitir que desenvolvedores construam/compartilhem ferramentas que operem sobre o modelo
  - Definir um modelo independente de tecnologia
  - Suportar a representação digital de proveniência de "qualquer coisa" produzido ou não por computador
  - Definir um conjunto de regras que podem ser usadas para identificar inferências realizadoas sobre os grafos de proveniência

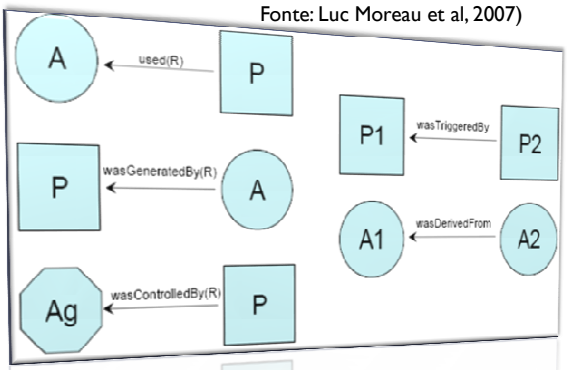
OPM

- **NÃO** é objetivo do OPM:
  - Especificar a representação de proveniência que os sistemas devem utilizar, cada sistema é livre utilizar representações individualizadas
  - Definir sintaxes ou parsers XML ou RDF
  - Definr protocolos para armazenar a proveniência em repositórios
  - Definir protocolos para consultar repositórios de proveniência

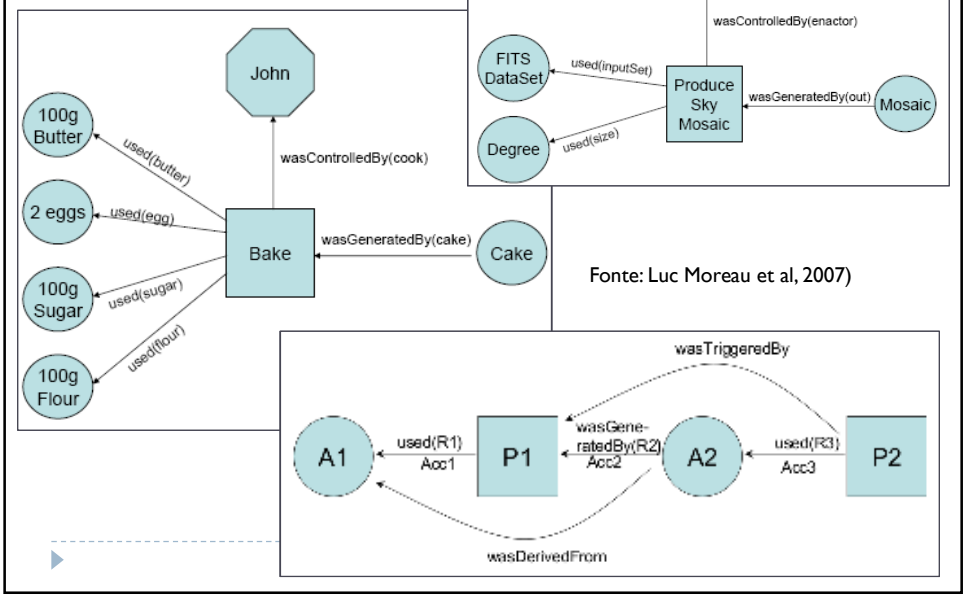
OPM

- Entidades primárias
  - Artefato -
  - Processo -
  - Agente -
- Dependências
  - Used by
  - wasGenetatedBy
  - wasTriggeredBy
  - wasDerivedFrom
  - wasControledBy
- Regras - role designates an artifact's or agent's function in a process.

Fonte: Luc Moreau et al, 2007)



OPM -exemplos



Fonte: Luc Moreau et al, 2007)

1st e 2nd Provenance Challenges

- **Objetivo:**
  - Avaliar os diferentes sistemas de proveniência, testar suas habilidades de consulta e a expessividade das suas representações de proveniência
  - *17 times, 1 workflow, 9 queries*
- **Resultados:**
  - Questões e respostas com baixa precisão
  - Validação difícil
  - Poucos sistemas responderam todas as queries
- Ainda há muito trabalho pela frente…
  - Falta de formalismo para a proveniência e queries [Hidders et al., DILS2007]

http://twiki.ipaw.info/bin/view/Challenge/WebHome

Agenda – Parte IV

- **Demonstrações de SGWfC**
  - Kepler
  - VisTrails
  - Workflows Científicos com Mashups

Kepler – Visão geral

Característica	
Responsáveis	U. San Diego/U. C. Davis -www.kepler-project.org
Sistemas operacionais	Win, MacOX, Linux
Linguagem do Sistema	Java (Código aberto)
GUI	SIM (Vergil)
Representação do Wf	Non-DAG
Controle de Concorrência	SIM (só no MoC PN)
Colaboração	NÃO
Tolerância a falhas	NÃO
QoS constrains	NÃO
Grid-based services	Previsto (mas não funcionam!)
Escalonamento de Jobs	NÃO (Manual - Aplicações em Cluster/Grids)
Linguagem de Especificação	MoML (XML)
Captura automática de proveniência	NÃO
Subworkflows	SIM

107

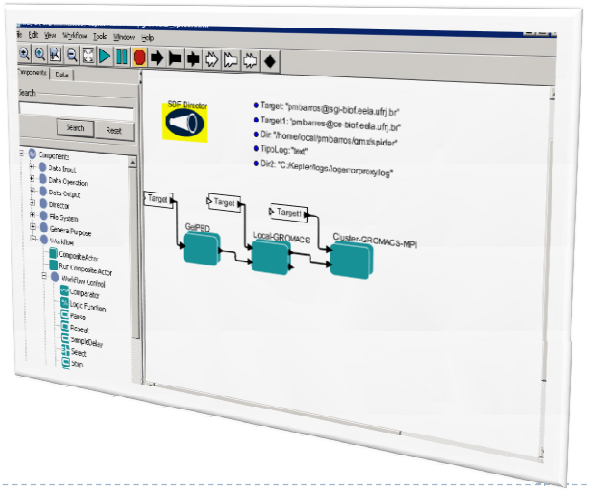
Kepler (detalhes)

- Construído sobre o Ptolemy II (+- 20 anos)
- Vários Modelos de Computação - (semântica e escalonamento próprio)
- Acesso a dados heterogêneos
  - Data access wizard
  - Acesso/Consultas BD Relacional
- Execução de Wf científicos
  - Processamento local ou remoto
- Possui interface para outras linguagens de Java para…
  - C++, Python, C
- Wfs baseados em XML (MoML)
- Separa 2 modelos:
  - Comunicação entre componentes(dataflow)
  - Coordenação do Wf (orquestração).

www.kepler-project.org

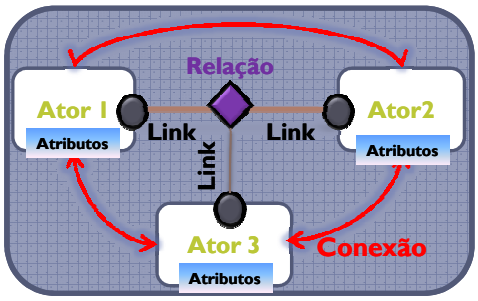
Componentes no Kepler

- Diretores
- Atores
- Parâmetros
- Portas
- Relações



Sintaxe Abstrata - Ptolemy II

Entidades Hierárquicas, Portas, Conexões e Atributos



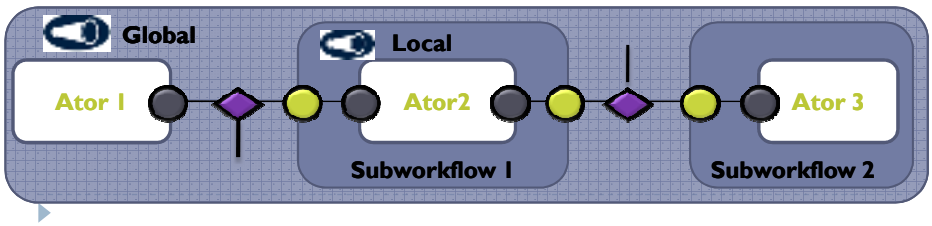
A **sintaxe abstrata** define apenas a estrutura do modelo!

Sintaxe abstrata:

- ✓ Hirarquias representam estruturas
- ✓ As relações são mediadores de conexões
- ✓ Portas podem linkar multiplas relações e relaçoes podem linkar várias portas
- ✓ Portas podem mediar conexões entre diferentes níveis hierárquicos

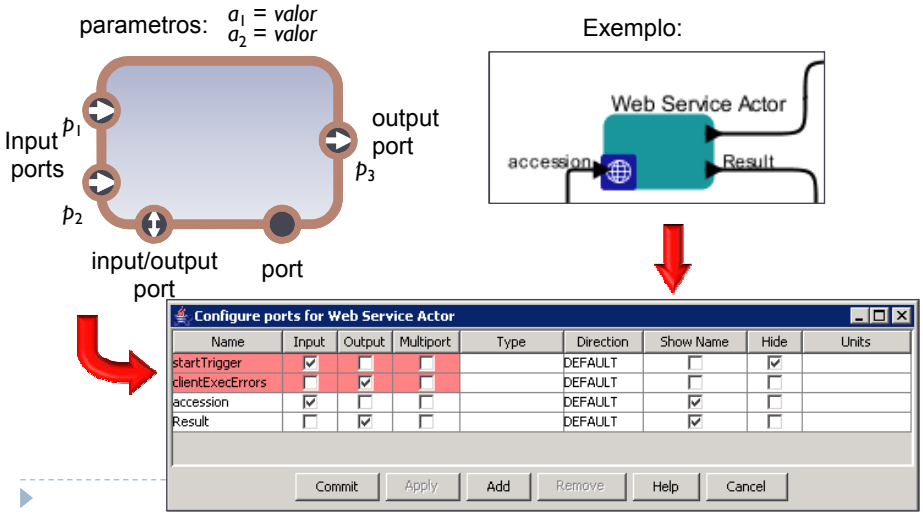
Diretor

- Governa a execução de um workflow ou subworkflow
  - Escalonamento, Disparo, Threads, geração de código, etc.
- Uma entidade composta (subworkflow) é dita ***Opaca*** se não possuir um diretor local
  - Uma entidade opaca HERDA o diretor do seu container (diretor executivo).
- Diretores facilitam: **abstração**, **modelagem**, **reuso de componentes...**



Actor Interfaces: *Portas & Parâmetros*

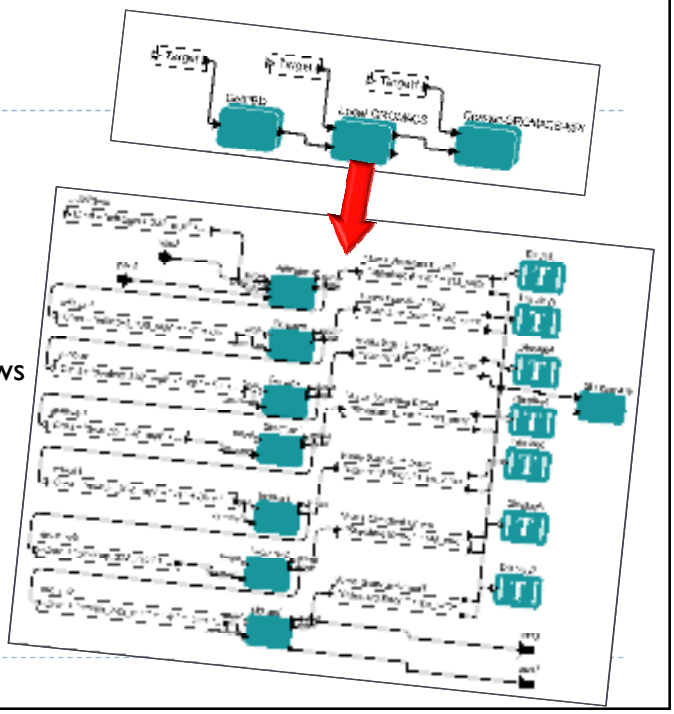
3 tipos de porta (input, output e input/output)



Atores

- Classes Java
- Simples
- Compostos
  - Subworkflows

Workflow Concreto!!!



Acesso a SGBD

- Database connection.
  - Abre e compartilha uma conexão com BD com todos os atores
  - Conecta vários SGBDs
- Database Query
  - Ator genérico de consulta.
- Close Database
  - Encerra uma conexão

Esquemas muitos grandes causam erro (ex.GUS)

Kepler – Características desejáveis

Característica	Atende
Interface	SIM
Reuso	+-
Transformação de dados	SIM
Iteração e batch	NÃO
Monitoração	NÃO
Distribuição	NÃO
Flexibilidade	NÃO
Desempenho e planejamento	NÃO
Tolerância a falhas	NÃO
Verificação e Validação	NÃO
Proveniência	+- (Artigos)
Segurança	NÃO

115

Kepler…. *Demo*

Vistrails – visão geral

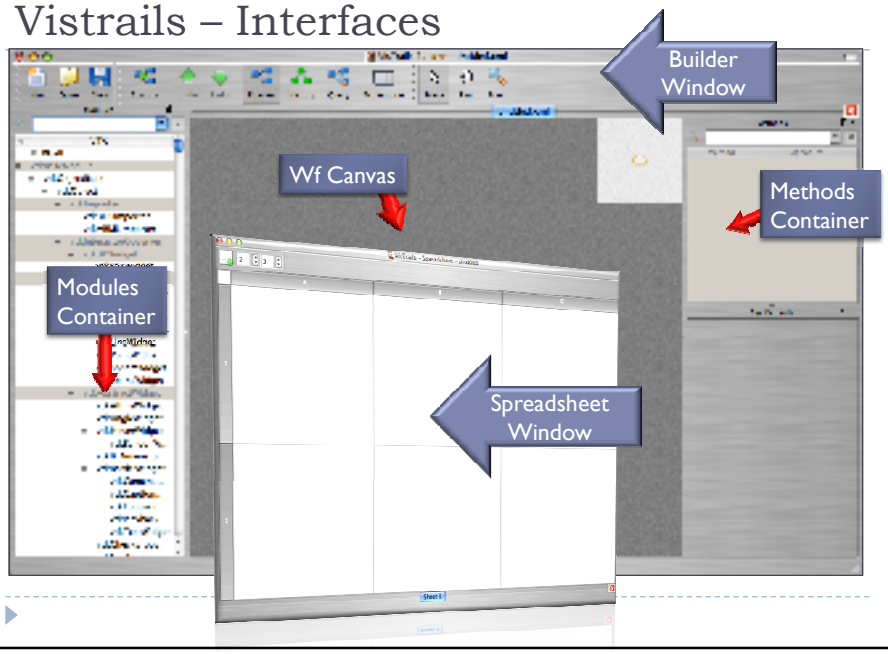
Característica	
Responsáveis	Univ. UTAH e startup - www.vistrails.org
Sistemas operacionais	Win, MacOX, Linux
Linguagem do Sistema	Phyton (Código aberto)
Representação do Wf	DAG
Controle de Concorrência	NÃO
Colaboração	SIM
Serviços Web	SIM
Tolerância a falhas	NÃO
QoS constrains	NÃO
Grid-based services	NÃO
Escalonamento de Jobs	MANUAL
Linguagem de Especificação	.VT (XML)
Captura automática de proveniência	SIM (prospectiva e restrospectiva)
Subworkflows	SIM

Vistrails (detalhes)

- Principais Características:
  - Registra alterações do wf, opcionalmente também nas informações de execução (quem, quando, onde, quanto tempo)
    - Possui área de anotações onde o cientista pode registrar/consultar suas próprias informações de proveniência.
  - Querying and Re-using History
    - Proveniência armazenada em XML ou Relacional (MySQL, DB2)
    - Consultas baseadas em palavra chave
  - Suporte para "collaborative exploration"
    - Repositório de workflows e controle de versões
    - Concepção e colaboração assíncrona de workflows
  - Permite a adição dinâmica de pacotes Python

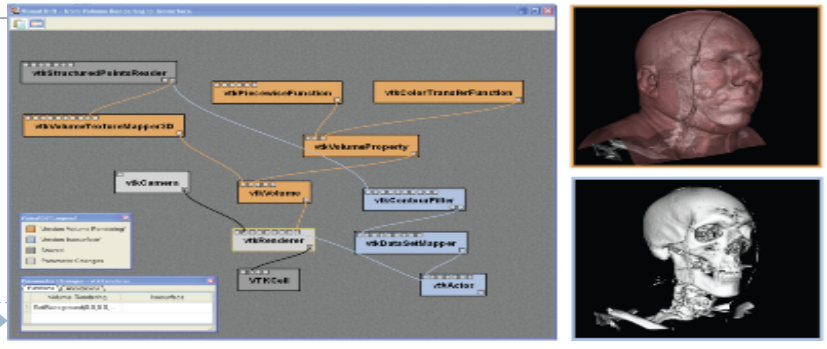
www.vistrails.org

Vistrails – Interfaces

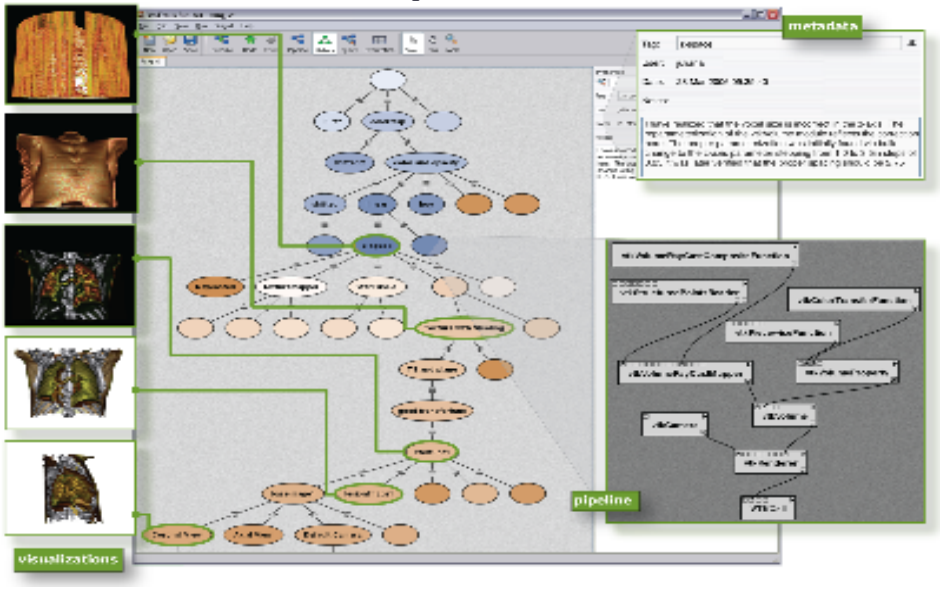


Componentes no VisTrails

- Construtor (Builder) : Pipeline, History, Query e Parameter Exploration
  - History
    - Armazena o histórico do workflow (as versões anteriores, assim como seus parâmetros, são armazenadas);
    - A **janela Visual Diff** permite a comparação entre duas versões;
    - O usuário pode fazer anotações das características das versões e também pode buscar versões específicas do workflow (Query).

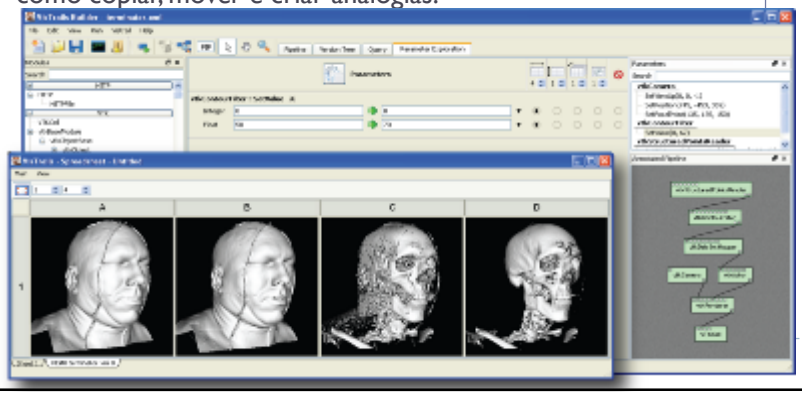


Histórico de ações sobre o wf



Componentes no VisTrails

- Planilha (Spreadsheet)
  - Componente que permite a comparação visual dos resultados;
  - Os resultados são apresentados nas chamadas células de visualização;
  - No modo de edição, algumas opções são mostradas nas células, como copiar, mover e criar analogias.



Vistrails – Características desejáveis

Característica	Atende
Interface	SIM
Reuso	SIM
Transformação de dados	NÃO
Iteração e batch	NÃO
Monitoração	NÃO
Distribuição	NÃO
Flexibilidade	SIM
Desempenho e planejamento	SIM-NÃO
Tolerância a falhas	NÃO
Verificação e Validação	NÃO
Proveniência	SIM
Segurança	NÃO

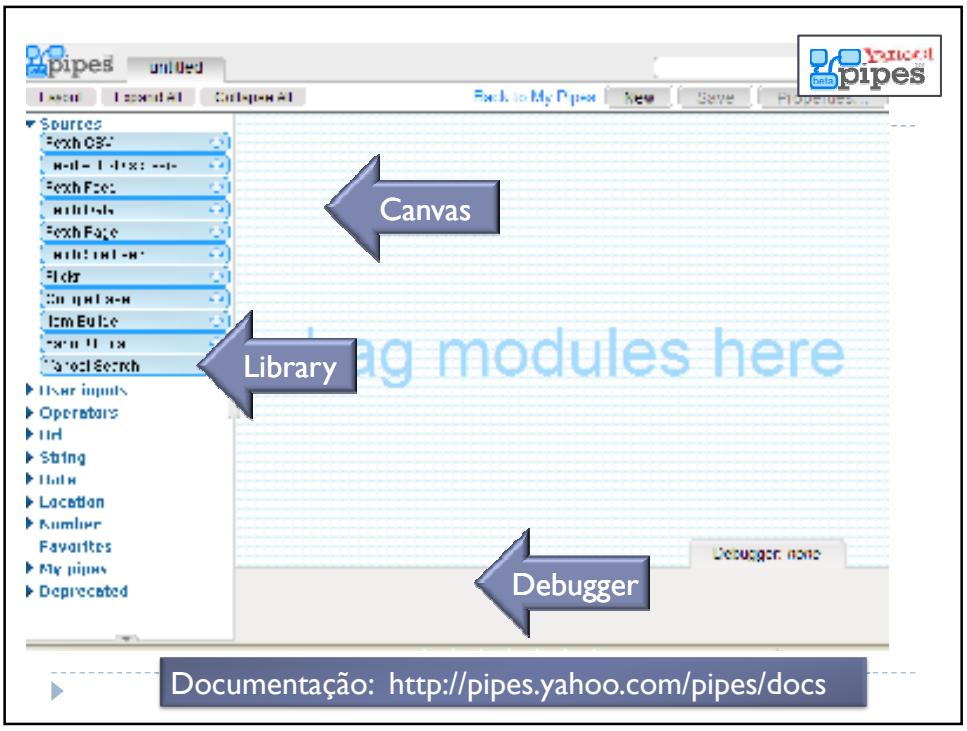
124

Vistrails.... *Demo*

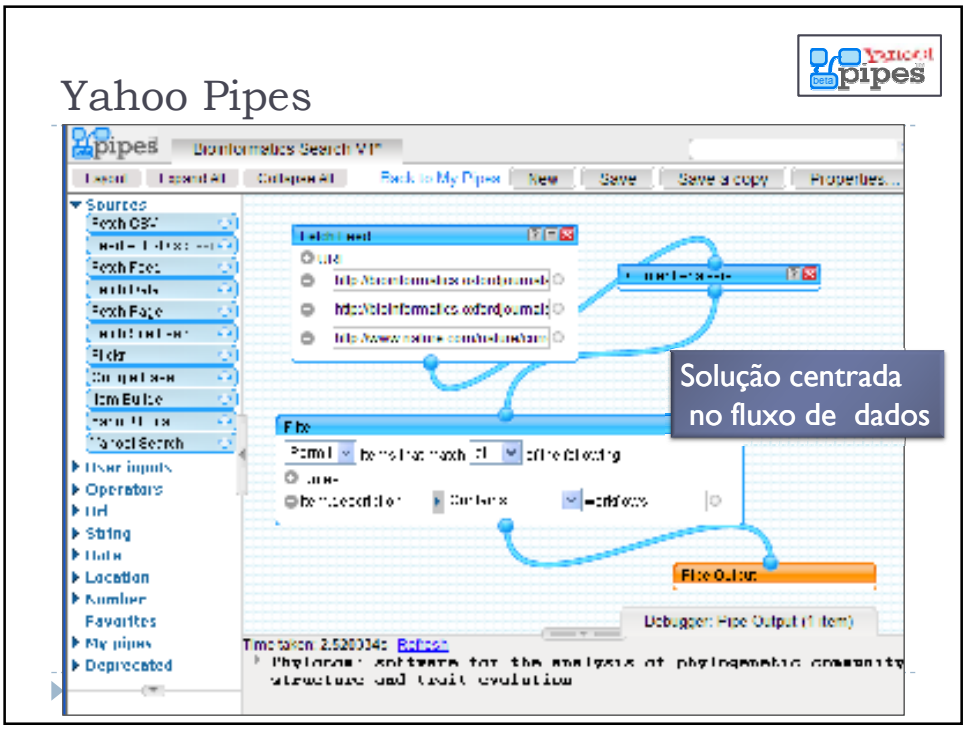
Workflows Científicos com Mashups

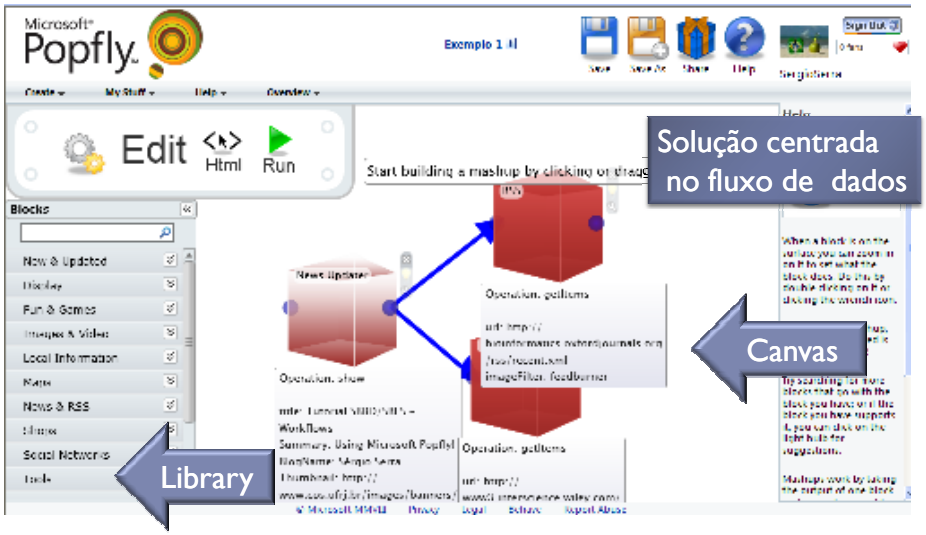
- **Mashups** – Combinação de Dados and Softtware/Serviços de várias fontes em uma nova aplicação Web
- Formam um ecosistema vibrante
- Vantagens
  - Criatividade
  - Inovação
  - Exploração
  - Compartilha idéias
  - "Programação Zero"
- Desvantagens
  - *Threshold* alto (requer conhecimento)
  - APIs são "cripticas e mutantantes"!
  - Fluxos centrados em dados

- Yahoo Pipes
- MS Popfly
  - Silverlight
- Google Mashups
- Intel MashMaker
- Lotus Mashups
- Serena Mashup Composer
- E muito mais…


Canvas
Library
Debugger
Documentação: http://pipes.yahoo.com/pipes/docs

Yahoo Pipes


Solução centrada no fluxo de dados



Popfly

- Genbank 2 Pubmed
- http://www.popfly.com/users/jsum/GeneId%20Article%20Viewer
  - Dado um Genbank ID, o mashup recupera um artigo da base de dados do Pubmed.
  - É uma tarefa rotineira em Bioinformática

Fonte: BioMashups: the new world of exploratory bioinformatics?
Jiro Sumitomo, James M. Hogan and Paul Roe

Yahoo Pipes & PopFly–*resumo*

Característica	Yahoo Pipes	Popfly
Responsável	Yahoo	Microsoft
Sistemas operacionais	Win, MacOX, Linux	Win, MacOX, Linux
Linguagem do Sistema	??? (proprietário)	??? (proprietário)
GUI	SIM	SIM
Representação do Wf	DAG	DAG
Controle de Concorrência	NÃO	NÃO
Serviços Web	SIM	SIM
Tolerância a falhas	NÃO	NÃO
QoS constrains	---	---
Grid-based services	---	---
Escalonamento de Jobs	---	---
Linguagem de Especificação	???	???
Captura automática de proveniência	NÃO	NÃO
Subworkflows	NÃO	NÃO

132

Oportunidades de pesquisa em BD

Retomando: Características desejáveis dos SGWf científicos (1)

Característica	Descrição
Interface	Design intuitivo, voltado para o usuário final. Detalhes de implementação (baixo nível) devem ser escondidos, foco no nível conceitual
Reuso	**Apresentar componentes reutilizáveis e intercambiáveis, idealmente devem ter capacidade de adicionar/remover novos processos dinamicamente**
Transformação de dados	**Permitir consecutivas transformações de dados entre as atividades**
Iteração e batch	Suportar "process steering" (play, pause e stop) durante a execução do workflow.
Monitoração	**Monitorar processos em tempo de execução mesmo em segundo plano, máquinas e ambientes distintos**
Distribuição	**Suportar processamento local e/ou distribuído**
Streamming	Ajuda a transferências de dados (moderada a intensa)

Características desejáveis dos SGWf científicos (2)

Característica	Descrição
Flexibilidade	**Suportar alterações na descrição do workflow (coleções de dados e programas)**
Complexidade	**Manipular complexos fluxos de dados, controles e eventos.**
Desempenho e planejamento	**Informar o desempenho e os custos de execução. Capaz de coletar dados de diferentes processos e usar métricas para predizer os tempos de execução**
Tolerância a falhas	Alta disponibilidade e tolerante a falhas
Verificação e Validação	Verificar e validar a construção ou importação de workflows
Proveniência	**Rastreio dos dados/processos utilizados e gerados em cada etapa**
Segurança	Armazenamento e tráfego de dados de proveniência de forma segura

Alguns trabalhos em andamento, COPPE/UFRJ, coordenação Marta Mattoso

Ciclo de Vida do Experimento Científico

Apoio à

Modelagem + Execução + Análise

- **Linha de Experimento**
- **Recomendação**
- **MiningFlow**
- **ControlFlow**

- **MidMon**
- **Segurança**
- **Matrioska**
- **OrthoSearch**

- **Matrioska**
- **MiningFlow**

MATTOSO, M. L. Q. ; C. M. L. Werner ; TRAVASSOS, G. H. ; BRAGANHOLO, V. P. ; MURTA, L. . Gerenciando Experimentos Científicos em Larga Escala. In: SEMISH, Congresso da SBC, 2008.

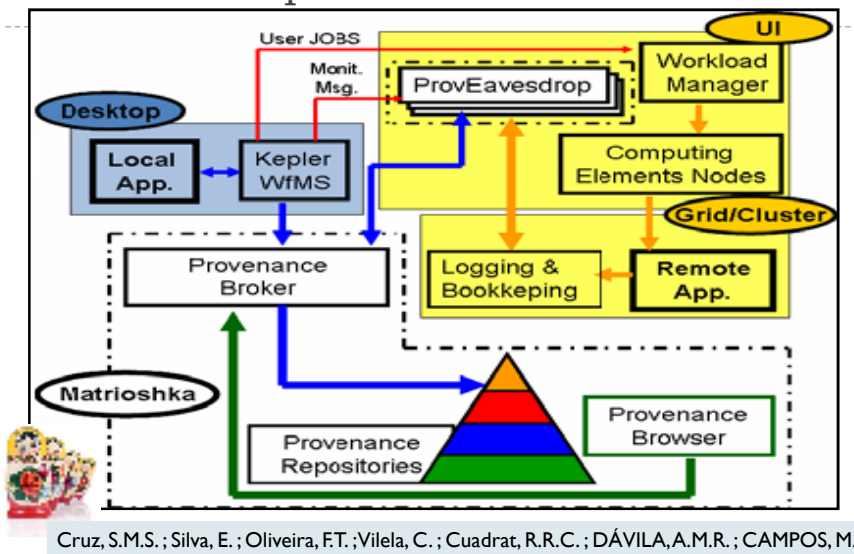
Proveniência

Matrioshka (Cruz. Barros et al.)

- **Motivação**: Coleta de proveniência em ambientes distribuídos (clusters e grids) com baixo acoplamento aos SGWf
- **Objetivos**:
  - Mecanismo independente de SGWf centralizados e storages, capaz de manipular diferentes representações de proveniência
  - Suporte captura e consulta a diferentes granularidades
  - Manter integridade e confidencialidade da proveniência em domínios distintos
  - Modelo Software as a Service

Cruz, S.M.S. ; Barros, P. ; BISCH, P. ; CAMPOS, M. L. M. ; MATTOSO, M. L. Q. . Provenance services for distributed workflows. In: 8th IEEE CCGrid´08, 2008. p. 526-533.

Matrioshka provenance services



Cruz, S.M.S. ; Silva, E. ; Oliveira, F.T. ; Vilela, C. ; Cuadrat, R.R.C. ; DÁVILA, A.M.R. ; CAMPOS, M. L. M. ; MATTOSO, M. L. Q. . OrthoSearch: A Scientific Workflow Approach to Detect Distant Homologies on Protozoans. In: ACM SAC 2008, v. II. p. 1281-1285.

138

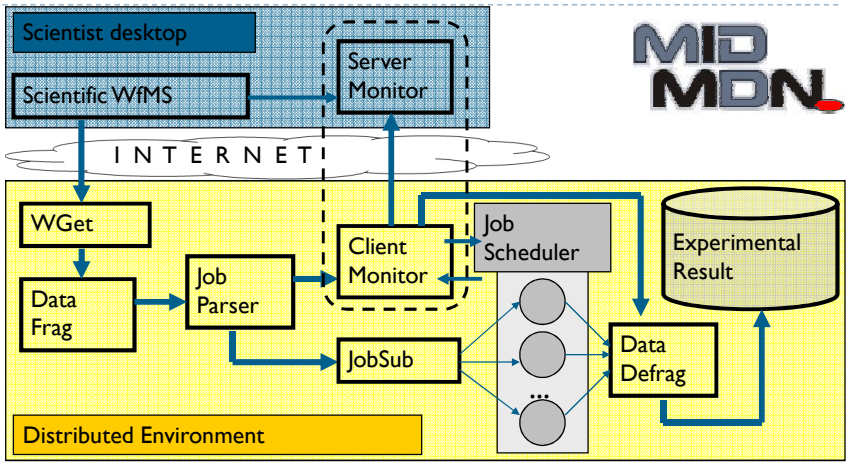
Monitoração

MidMon (Cruz et al)

- **Motivação**: Monitorar workflows científicos executando em ambientes distribuídos
- **Características**:
  - Deploy e manutenção simplificada de submissão de execução paralela de processos
  - Auxilia aplicações legadas e recursos distribuídos
  - Infraestrutura modular, baseado na troca de mensagens
  - Desacoplado do SGWf

Cruz, S.M.S. ; SILVA, F. N. ; Gadelha Jr., L.M.R. ; CAVALCANTI, M. C. ; CAMPOS, M. L. M. ; MATTOSO, M. L. Q. . A Lightweight Middleware Monitor for Distributed Scientific Workflows. In: Int. Workshop on Workflow Systems in e-Science, 8th IEEE CCGrid´08, 2008. p. 693-698.

MidMon Middleware Architecture



Complexidade

Controles de Fluxo (Chirigati, Dahis, et al

Motivação: Construção de wf científicos mais flexíveis, expondo diferentes semânticas para os usuários, facilitando a captura de dados remotos de proveniência

Objetivo:

Construir controles independentes de linguagem de especificação

Diminuir a dependendência das linguagens de definições dos SGWf e desacoplar-se dos mecanismos de captura de proveniência

Implementado no VisTrails

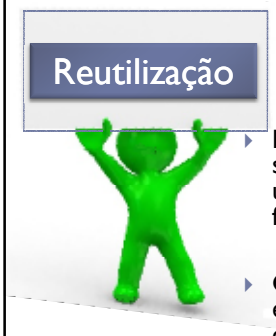
Cruz, S.M.S. ; Chirigati, F S ; DAHIS, R. ; CAMPOS, M. L. M. ; MATTOSO, M. L. Q. . Using explicit control processes in distributed workflows to gather provenance. In: International Provenance and Annotation Workshop, 2008, Salt Lake City. IPAW 2008.

Controle de fluxo (cont)

- Controles de fluxo genéricos
- Baseado nos padrões de van der Aalst et al.

Workflow Pattern	Módulo
Structured Discriminator	Mux
Exclusive Choice	Demux
Deferred Choice	String Control
Multiple Instances without synchronization	Number Control
Synchronization	Number Compare
Exclusive Choice	If

Cruz, S.M.S. ; Chirigati, F S ; DAHIS, R. ; CAMPOS, M. L. M. ; MATTOSO, M. L. Q. . Controles de Fluxo Explícitos em Workflows Científicos. In: E-Science Workshop, SBBD 2008, Campinas.



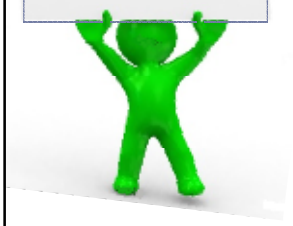
Reuso (E. Ogasawara, F. Oliveira)

- **Motivação**: Com o aumento da complexidade das simulações científicas, as atividades de concepção e utilização de workflows científicos não podem mais ser feitas de modo ad-hoc.
- **Objetivo**: Aplicar técnicas de engenharia de software existentes, adaptadas do contexto de software, para o contexto de workflows científicos.
- **Como**: Através do conceito de ***Linhas de Experimentos***, que aplica duas técnicas tradicionais de engenharia de software na concepção e utilização de workflows científicos: **recomendação** e gerência de configuração.

OGASAWARA, E. ; MURTA, L. ; C. M. L. Werner ; MATTOSO, M. L. Q. . Linhas de Experimentos: Reutilização e Gerência de Configuração em Workflows Científicos. In: E-Science Workshop, SBBD 2008, Campinas.

Oliveira, F.T. ; MURTA, L. ; C. M. L. Werner ; MATTOSO, M. L. Q. . Using Provenance to Improve Workflow Design. In: International Provenance and Annotation Workshop, 2008, Salt Lake City. IPAW 2008.

Flexibilidade

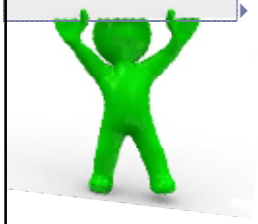


MiningFlow (D. Oliveira)

- **Motivação**: Apoio ao ciclo de vida de mineração de texto via workflows
- **Objetivo**: Aplicar ontologias na concepção de workflows científicos
- **Como**: Associação de conceitos de ontologias a dados e processos, oferecendo semântica durante a concepção e execução do experimento. Gera "código" para : Kepler, Taverna e VisTrails

Oliveira, D. ; BAIÃO, F. ; MATTOSO, M. L. Q. . MF-Ontology, uma ontologia para o processo de mineração de textos. In: Seminário de Pesquisa em Ontologia no Brasil, 2008..

Segurança



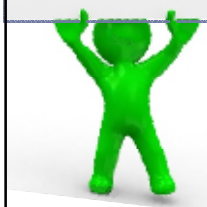
Segurança (L. Gadelha)

Motivação: A maioria dos SGWFs com gerência de proveniência não implementa controles de segurança aos dados de proveniência. A proteção dos registros de proveniência é um requisito freqüente em algumas comunidades científicas específicas, como a de bioinformática.

Objetivo: Inclusão de mecanismos adicionais de segurança para proteção de autoria em experimentos científicos: assinatura digital e datação criptográfica (time-stamping) de registros de proveniência.

Como: Uso de técnicas adequadas de controle de acesso aos registros de proveniência.

Planejamento



Formalismos em Wf (E. Silva)

Motivação:

Construção de wf científicos, tendo como base formalismos matemáticos, facilitando a definição correta e execução livre de erros.

Objetivo:

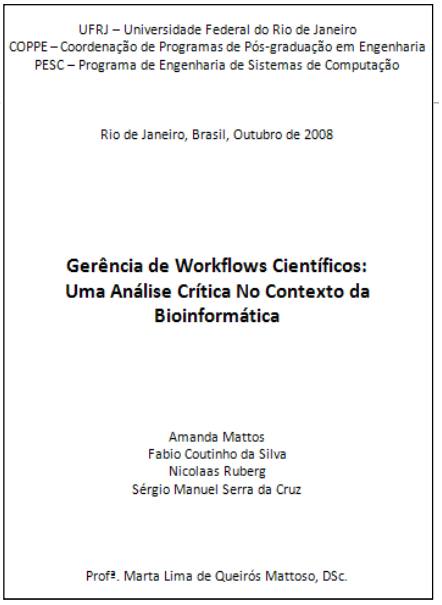
Representar precisamente padrões de comportamento de workflows por meio de álgebras de processo.

Representação dos workflows independente de linguagem de especificação dos SGWf,

Como: Uso de técnicas de verificação de modelos, garantindo a correção da especificação

Referências - principais

- S. B. Davidson and J. Freire. Provenance and scientific workflows: challenges and opportunities. In Proceedings of ACM SIGMOD, pages 1345–1350, 2008.
- J. Freire, D. Koop, E. Santos, and C. T. Silva. Provenance for computational tasks: A survey. Computing in Science and Engineering, 10(3):11–21, 2008.
- I.J. Taylor, E. Deelman, D. B. Gannon, M. Shields (eds) "Workflows for E-Science", Springer 2007.
- Y. L. Simmhan, B. Plale, and D. Gannon. A survey of data provenance in e-science. SIGMOD Record, 34(3):31–36, 2005.
- L. Moreau, editor. Concurrency and Computation: Practice and Experience– Special Issue on the First Provenance Challenge, 2008.
- L. Moreau and I. Foster, editors. Provenance and Annotation of Data - International Provenance and Annotation Workshop, volume 4145. Springer-Verlag, 2006.
- Yu, J., Buyya, R. "A Taxonomy of Scientific Workflow Systems for Grid Computing", SIGMOD Record, Vol. 34, No. 3, Sept, 2005.



UFRJ – Universidade Federal do Rio de Janeiro
COPPE – Coordenação de Programas de Pós-graduação em Engenharia
PESC – Programa de Engenharia de Sistemas de Computação

Rio de Janeiro, Brasil, Outubro de 2008

Gerência de Workflows Científicos: Uma Análise Crítica No Contexto da Bioinformática

Amanda Mattos
Fabio Coutinho da Silva
Nicolaas Ruberg
Sérgio Manuel Serra da Cruz

Profª. Marta Lima de Queirós Mattoso, DSc.

www.cos.ufrj.br/uploadfiles/1204740464.pdf

Referências

- Aalst. W. van der, Hee, K. M. Van "Workflow Management: Models, Methods, and Systems" MIT Press. 2002.
- Aalst, W. M. P., Hofstede, A. H. M., Kiepuszewski, B., Barros, A. P. "Workflow Patterns". 2003.
- Aalst. W. van der; Hofstede, A. H. M. ter, YAWL: Yet another workflow language. Information Systems, 30(4):245–275, 2005.
- Addis, M., Ferris, J., Greenwood, M., Li, P., Marvin, D., Oinn, T. and Wipat, A., "Experiences with e-Science workflow specification and enactment in bioinformatics", Proceedings of e-Science All Hands Meeting 2003, p. 459-466, East Midlands Conference Centre, Nottingham, 2003.
- Altintas, I., Barney, O., & Jaeger-Frank, E. (2006). Provenance collection support in the Kepler Scientific Workflow System In International Provenance and Annotation Workshop (IPAW), LNCS, Provenance and Annotation of Data, 4145: 118-132, 2006.
- I. Altintas, O. Barney, and E. Jaeger-Frank. Provenance collection support in the kepler scientific workflow system. In Proceedings of the International Provenance and Annotation Workshop (IPAW), pages 118–132, 2006.

Referências

- E. Andersen, S. P. Callahan, D. A. Koop, E. Santos, C. E. Scheidegger, H. T. Vo, J. Freire, and C. T. Silva. Vistrails: Using provenance to streamline data exploration. In Poster Proceedings of the International Workshop on Data Integration in the Life Sciences (DILS), page 8, 2007.
- R. Barga and L. Digiampietri. Automatic generation of workflow provenance. In IPAW, pages 1–9, 2006. Invited paper.
- R. S. Barga and L. A. Digiampietri. Automatic capture and efficient storage of escience experiment provenance. Concurrency and Computation: Practice and Experience, 20(5):419–429, 2008.
- O. Biton, S. C. Boulakia, and S. B. Davidson. Zoom*userviews: Querying relevant provenance in workflow systems. In Proceedings of VLDB, pages 1366–1369, 2007.
- L. Bavoil, S. Callahan, P. Crossno, J. Freire, C. Scheidegger, C. Silva, and H. Vo. Vistrails: Enabling interactive multiple-view visualizations. In Proceedings of IEEE Visualization, pages 135–142, 2005.
- BioWebDB Portal, in: http://www.biowebdb.org/index.html/
- Braghetto, K. R., Broinizi, M. E. B., Ferreira, J. E., Pu, C., "Simplifying the representation and execution of Workflow Patterns through Navigation Plan Definition Language", 2005.

Referências

- P. Buneman and W. Tan. Provenance in databases. In Proceedings of ACM SIGMOD, pages 1171–1173, 2007.
- S. Callahan, J. Freire, E. Santos, C. Scheidegger, C. Silva, and H. Vo. Managing the evolution of dataflows with vistrails. In IEEE Workshop on Workflow and Data Flow for Scientific Applications (SciFlow), 2006.
- A. P. Chapman, H. V. Jagadish, and P. Ramanan. Efficient provenance storage. In Proceedings of ACM SIGMOD, pages 993–1006, 2008.
- B. Clifford, I. Foster, M. Hategan, T. Stef-Praun, M. Wilde, and Y. Zhao. Tracking provenance in a virtual data grid. Concurrency and Computation: Practice and Experience, 20(5):565–575, 2008.
- S. B. Davidson, S. C. Boulakia, A. Eyal, B. Ludäscher, T. M. McPhillips, S. Bowers, M. K. Anand, and J. Freire. Provenance in scientific workflow systems. IEEE Data Eng. Bull., 30(4):44–50, 2007.
- S. B. Davidson and J. Freire. Provenance and scientific workflows: challenges and opportunities. In Proceedings of ACM SIGMOD, pages 1345–1350, 2008.

Referências

- Digiampietri, Luciano Antonio ; Perez-Alcazar, Jose ; Medeiros, C. M. B. . An Ontology-based Framework for Bioinformatics Workflows. International Journal of Bioinformatics Research and Applications, v. 3, p. 268-285, 2007
- Gil, Y., E. Deelman, et al. Examining the Challenges of Scientific Workflows. IEEE Computer, 2007.
- Goble, C., Wroe, C., Stevens, R., and the myGrid consortium, "The myGrid Project: Ser-vices, Architecture and Demonstrator", Proceedings UK e-Science All Hands Meeting 2003 Editors - Simon J Cox, p. 595-603, 2003.
- Goderis, A., Sattler, U., Lord, P., Goble, C., "Seven Bottlenecks to Workflow Reuse and Repurposing". ISWC, LNCS 3729, pp. 323-337, 2005.
- Hollingsworth, D. - WFMC, "Workflow Management Coalition Terminalogy & Glossary" WFMC-TC 1011, http://www.wfmc.org/standards/docs/
- J. Kim, E. Deelman, Y. Gil, G. Mehta, and V. Ratnakar. Provenance trails in the wings/pegasus system. Concurrency and Computation: Practice and Experience, 20(5):587–597, 2008.

Referências

- Ludäscher, B. Altintas, I., Berkley, et al., "Scientific Workflow Management and the Kepler System" Concurrency and Computation: Practice & Experience, 18(10), pp. 1039-1065, 2006.
- L. Moreau, J. Freire, J. Futrelle, R. McGrath, J. Myers, and P. Paulson. The open provenance model, December 2007. http://eprints.ecs.soton.ac.uk/14979.
- Medeiros, C. M. B. ; Perez-alcazar, Jose ; Digiampietri, Luciano Antonio ; Pastorello Jr, Gilberto Zonta ; Santanchè, André ; Torres, Ricardo da Silva ; Madeira, Edmundo Roberto Mauro . WOODSS and the Web: Annotating and Reusing Scientific Workflows. SIGMOD Record, New York, v. 34, n. 3, p. 18-23, 2005.
- myGrid. Disponível em http://www.mygrid.org.uk.
- http://www.myexperiment.org/workflows.
- myTea Project, http://mytea.org.uk/

Referências

- T. Oinn, M. Greenwood, M. Addis, et al. Taverna: lessons in creating a workflow environment for the life sciences: Research articles. Concurrency and Computation: Practice & Experience, 18(10):1067–1100, 2006.
- C. Scheidegger, D. Koop, E. Santos, H. Vo, S. Callahan, J. Freire, and C. Silva. Tackling the provenance challenge one layer at a time. Concurrency and Computation: Practice and Experience, 20(5):473–483, 2008.
- Smart tea projecthttp://www.smarttea.org/
- Wroe, C., Goble, C., Goderis, A., Lord, P., Miles, S., Papay, J., Alper, P., Moreau, L. "Recycling workflows and services through discovery and reuse". 2005.

Considerações Finais

- Levantamento faz parte do projeto **GExp – Gerência de Experimentos Científicos em Larga Escala**, financiado pelo Edital dos Grandes Desafios-MCT/CNPq/CT-INFO n. 07/2007

http://gexp.nacad.ufrj.br/

SBBD SBES 2008

XXIII SIMPÓSIO BRASILEIRO DE BANCO DE DADOS
XXII SIMPÓSIO BRASILEIRO DE ENGENHARIA DE SOFTWARE
Campinas, Brasil - 13 a 17 de Outubro

Muito Obrigado!

Gerência de workflows científicos: oportunidades de pesquisa em bancos de dados

Marta Mattoso, Sérgio Manuel Serra da Cruz

{marta,serra}@cos.ufrj.br

CNPq

PESC Programa de Engenharia de Sistemas e Computação

UFRJ COPPE