Pontifícia Universidade Católica
do Rio de Janeiro

Luciana da Silva Almendra Gomes

Proveniência para Workflows de Bioinformática

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós-Graduação em Informática da PUC-Rio.

Orientador: Edward Hermann Haeusler

Rio de Janeiro
Abril de 2011

Luciana da Silva Almendra Gomes

Proveniência para Workflows de Bioinformática

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre pelo Programa de Pós-graduação em Informática do Departamento de Informática do Centro Técnico e Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Edward Hermann Haeusler
Orientador
Departamento de Informática – PUC-Rio

Prof. Sérgio Lifschitz
Departamento de Informática – PUC-Rio

Profª Marta Lima de Queirós Mattoso
UFRJ

Prof. Laurent Emmanuel Dardenne
LNCC

Prof. José Eugenio Leal
Coordenador Setorial do Centro
Técnico Científico – PUC-Rio

Rio de janeiro, 27 de abril de 2011

Luciana da Silva Almendra Gomes

Graduou-se em Informática na Pontifícia Universidade Católica do Rio de Janeiro (2008).

Ficha Catalográfica

Gomes, Luciana da Silva Almendra

Proveniência para Workflows de Bioinformática / Luciana da Silva Almendra Gomes ; orientador: Edward Hermann Haeusler. – 2011.
104 f. : il. ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2011.
Inclui bibliografia

1. Informática – Teses. 2. Bioinformática. 3. Workflow. 4. Proveniência. I. Haeusler, Edward Hermann. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

À minha avó Judith.
Aos meus pais, Lucia e Geraldo, e meu irmão, Vinicius.
Ao meu esposo, Adriano.
Às minhas filhas Beatriz e Alice.

Agradecimentos

A Deus, por tudo.

Aos meus pais, meus primeiros mestres, por todo o amor, dedicação e incentivo.

Ao meu irmão, um grande amigo, que sempre me ajudou em tudo.

Ao meu esposo, por todo o seu carinho, paciência, compreensão e colaboração nesses anos de estudo intenso.

Às nossas filhas, Beatriz e Alice, a primeira nascida no meio do mestrado, e a segunda um pouco depois da defesa. Beatriz e Alice chegaram para encher nossas vidas de alegria!

À minha família e meus amigos, que sempre me apoiaram em minhas decisões.

Ao meu orientador, Professor Sérgio Lifschitz, que desde a graduação direcionou os rumos da minha pesquisa, ajudando-me a distinguir entre um trabalho meramente técnico e uma pesquisa científica.

Aos professores que participaram da banca examinadora, por todos os excelentes comentários.

A Philippe Picouet e Sébastien Bigaret, pela cooperação com o nosso trabalho.

Aos amigos do LabBio, pelas inúmeras ajudas. À Márcia, por ter contribuído com tanta prontidão com a descrição de um dos estudos de caso. Aos alunos Waldecir e Diego, pela ajuda com a implementação do projeto.

Aos amigos Elaine e Hugo, pelas idéias e ajudas com o texto.

Aos professores e à secretaria de pós-graduação, por todo o serviço prestado com excelência e solicitude.

Ao CNPq, pelos auxílios concedidos, sem os quais não teria sido possível realizar esse trabalho.

Resumo

Gomes, Luciana da Silva Almendra; Haeusler, Edward Hermann. **Proveniência para Workflows de Bioinformática**. Rio de Janeiro, 2011. 104p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Muitos experimentos científicos são elaborados como fluxos de tarefas computacionais, que podem ser implementados através do uso de linguagens de programação. Na área de bioinformática é muito comum o uso de *scripts ad-hoc* para construir fluxos de tarefas. Os Sistemas de Gerência de Workflow Científico (SGWC) surgiram como uma alternativa a estes *scripts*. Uma das funcionalidades desses sistemas que têm recebido bastante atenção pela comunidade científica é a captura automática de dados de proveniência. Estes permitem averiguar quais foram os recursos e parâmetros utilizados na geração dos resultados, dentre muitas outras informações indispensáveis para a validação e publicação de um experimento. Neste trabalho foram levantados alguns desafios na área de proveniência de dados em SGWCs, como por exemplo (i) a heterogeneidade de formas de representação dos dados nos diferentes sistemas, dificultando a compreensão e a interoperabilidade; (ii) o armazenamento de dados consumidos e produzidos e (iii) a reprodutibilidade de uma execução específica. Estes desafios motivaram a elaboração de um esquema conceitual de proveniência de dados para a representação de *workflows*. Foi implementada também uma extensão em um SGWC específico (BioSide) para incluir dados de proveniência e armazená-los utilizando o esquema conceitual proposto. Foram priorizados neste trabalho alguns requisitos comumente encontrados em *workflows* de Bioinformática.

Palavras-chave

bioinformática; workflow; proveniência.

Abstract

Gomes, Luciana da Silva Almendra; Haeusler, Edward Hermann (Advisor). **Provenance for Bioinformatics Workflows**. Rio de Janeiro, 2011. 104p. MSc. Dissertation – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Many scientific experiments are designed as computational workflows, which can be implemented using traditional programming languages. In the Bioinformatics domain *ad-hoc* scripts are often used to build workflows. Scientific Workflow Management Systems (SWMS) have emerged as an alternative to those scripts. One particular SWMS feature that has received much attention by the scientific community is the automatic capture of provenance data. These allow users to track which resources and parameters were used to obtain the results, among many other required information to validate and publish an experiment. In the present work we have elicited some data provenance challenges in the SWMS context, such as (i) the heterogeneity of data representation schemes that hinders the understanding and interoperability; (ii) the storage of consumed and produced data and (iii) the reproducibility of a specific execution. These challenges have motivated the proposal of a data provenance conceptual scheme for workflow representation. We have implemented an extension of a particular SWMS system (Bioside) to include provenance data and store them using the proposed conceptual scheme. We have focused on some requirements commonly found in bioinformatics' workflows.

Keywords

bioinformatics; workflow; provenance.

Sumário

Lista de figuras

Lista de tabelas