Referências

[Altintas et al, 2006] ALTINTAS, I. O.; BARNEY, O.; JAEGER-FRANK, E. **Provenance collection support in the kepler scientific workflow system**. In: Proc. International Provenance and Annotation Workshop (IPAW'2006), Chicago, Illinois, USA, 2006. LNCS 4145, pp. 118-132.

[Barga e Digiampietri, 2008] BARGA, R.; DIGIAMPIETRI, L. **Automatic capture and efficient storage of e-Science experiment provenance**. Concurrency and Computation: Practice & Experience 20 (Abril 2008): 419–429.

[Benson, 2010] BENSON, G. Nucleic Acids Res. 2010 July 1; 38 (Web Server issue): W1–W2. **Editorial**. Disponível em: < http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2896091/>

[Berman et al., 2000] BERMAN, H. et al. P.E. Bourne (2000) **The Protein Data Bank**. Nucleic Acids Research, 28: 235-242. URL: <www.pdb.org>.

[BioSide, 2011] **BioSide: a tool for biologists performing, learning, replicating and storing complex bio-analyses**. Disponível em: <http://www.bioside.org/>.

[Buneman et al., 2001] BUNEMAN, P. et al. **Why and Where: A Characterization of Data Provenance**. Lecture Notes in Computer Science, Volume 1973, International Conference on Database Theory (ICDT 2001), pages 316-330. 2001.

[Callahan et al., 2006] CALLAHAN, S. et al. **Managing the Evolution of Dataflows with VisTrails**. In Proceedings of the 22nd International Conference on Data Engineering Workshops, 71–. ICDEW '06. Washington, DC, USA: IEEE Computer Society, 2006.

[Callahan et al, 2006] CALLAHAN, S. et al. **VisTrails: visualization meets data management**. In Proceedings of the 2006 ACM SIGMOD international conference on Management of data, 745–747. SIGMOD '06. New York, NY, USA: ACM, 2006.

[Capriles et al., 2010] CAPRILES, P. et al. **Structural Modelling and Comparative Analysis of Homologous, Analogous and Specific Proteins from*Trypanosoma cruzi* versus *Homo sapiens*: Putative Drug Targets for Chagas' Disease Treatment**. BMC Genomics 2010, 11:610.*

doi:10.1186/1471-2164-11-610<http://www.biomedcentral.com/1471-2164/11/610>. URL: http://www.mholline.lncc.br.

[Clifford et al., 2008] CLIFFORD, B. et al. **Tracking provenance in a virtual data grid**. Concurrency and Computation: Practice & Experience 20 (Abril 2008): 565–575.

[Clustalw2, 2011] **Clustalw2**. Disponível em: <http://www.ebi.ac.uk/Tools/msa/clustalw2/help/index.html>. Acesso em: 13 abr. 2011.

[Curcin e Ghanem, 2008] CURCIN, V.; M. GHANEM. **Scientific workflow systems - can one size fit all?**. In 2008 Cairo International Biomedical Engineering Conference, 1-9. Cairo, Egypt, 2008.

[Cruz et al., 2009] CRUZ, S.; CAMPOS, M.; MATTOSO, M. **Towards a Taxonomy of Provenance in Scientific Workflow Management Systems**. IEEE International Workshop on Scientific Workflows, Los Angeles, California, USA, 2009.

[Davidson e Freire, 2008] DAVIDSON, S.; FREIRE, J. **Provenance and scientific workflows: challenges and opportunities**. In Proceedings of the 2008 ACM SIGMOD international conference on Management of Data, Vancouver, Canada, 2008. pp 1345-1350.

[Deelman et al., 2007] DEELMAN, E., et al. **Pegasus: mapping large-scale workflows to distributed resources**. Workflows for e-Science, Springer, pp.376–394, 2007.

[Deelman et al., 2009] DEELMAN, E. et al. **Workflows and e-Science: An overview of workflow system features and capabilities**. Future Generation Computer Systems 25, n. 5 (Maio 2009): 528-540.

[Ellkvist et al., 2008] ELLKVIST, T. et al. Using **Provenance to Support Real-Time Collaborative Design of Workflows**. In: Proc. International Provenance and Annotation Workshop (IPAW'2008), LNCS 5272, pp. 266-279, 2008.

[Freire et al., 2008] FREIRE, J. et al. **Provenance for Computational Tasks: A Survey**. Computing in Science & Engineering 10, n. 3 (2008): 11-21.

[Gil et al., 2007] GIL, Y. et al. **Examining the Challenges of Scientific Workflows**. Computer 40 (Dezembro 2007): 24–32.

[git, 2011] **git, the fast version control system**. Disponível em: <http://git-scm.com/>. Acesso em: 13 abr. 2011.

[Goble et al., 2010] GOBLE, A. et al. **myExperiment: a repository and social network for the sharing of bioinformatics workflows**, Nucl. Acids Res., 2010. doi:10.1093/nar/gkq429

[Greenwood et al., 2003] GREENWOOD, M. et al. **Provenance of e-Science Experiments - experience from Bioinformatics**. The UK OST e-Science second All Hands Meeting 2003 (AHM'03), September, Nottingham, UK. pp. 223-226, 2003.

[Guimaraes, 2009] GUIMARAES, M. **Uma Abordagem para capturar a proveniência de dados na área da Bioinformática**. Dissertação (Mestrado). Instituto Militar de Engenharia, 2009.

[Holland et al., 2008] HOLLAND, D. et al. **Choosing a Data Model and Query Language for Provenance**. In Proceedings of the 2008 ACM SIGMOD international conference on Management of Data, Vancouver, Canada, 2008.

[HSQLDB, 2011] **HSQLDB, 100% Java Database**. Disponível em: <http://hsqldb.org/>.

[Hull et al., 2006] HULL, D. et al. **Taverna: a tool for building and running workflows of services**. Nucleic Acids Research, vol. 34, iss. Web Server issue, pp. 729-732, 2006.

[Jones, 2011a] JONES, M. kepler-users Mailing List. Tópico: **wrapping command-line tools**. Mensagem enviada em 28 fev. 2011. Disponível em: <http://lists.nceas.ucsb.edu/kepler/pipermail/kepler-users/2011-February/002463.html>. Acesso em: 13 abr. 2011.

[KeplerLSID, 2011] **Kepler Life Science Identifiers (KLSID)**. Disponível em: <https://kepler-project.org/developers/teams/framework/kepler-life-science-identifiers-keplerlsid/?searchterm=lsid>. Acesso em: 13 abr.2011.

[Kepler Prov, 2010] Kepler Provenance, 2010. **Getting Started with the Kepler Provenance Module**. Versão 2.1.0, 08 ago. 2010. Disponível em: <https://code.kepler-project.org/code/kepler/trunk/modules/provenance/docs/provenance.pdf>. Acesso em 13 abr. 2011.

[Koop et al., 2010a] KOOP, D. et al. **The Provenance of Workflow Upgrades**. In: Proc. International Provenance and Annotation Workshop (IPAW'2010), LNCS 6378, pp. 2-16, 2010.

[Koop et al., 2010b] KOOP, D. et al. **Bridging workflow and data provenance using strong links**. In Proceedings of the 22nd international conference on Scientific and statistical database management, 397–415. SSDBM'10. Berlin, Heidelberg: Springer-Verlag, 2010.

[Koop, 2011a] KOOP, D. vistrails-users Mailing List. Tópico: **persistence package**. Mensagem enviada em 11 jan. 2011. Disponível em: <https://lists.sci.utah.edu/sympa/arc/vistrails-users/2011-01/msg00009.html>. Acesso em: 13 abr. 2011.

[Koop, 2011b] KOOP, D. vistrails-users Mailing List. Tópico: **database model**. Mensagem enviada em 6 jan. 2011. Disponível em: <https://lists.sci.utah.edu/sympa/arc/vistrails-users/2011-01/msg00004.html>. Acesso em: 13 abr. 2011.

[Koop, 2011c] KOOP, D. vistrails-users Mailing List. Tópico: **query**. Mensagem enviada em 16 fev. 2011. Disponível em: <https://lists.sci.utah.edu/sympa/arc/vistrails-users/2011-02/msg00010.html>. Acesso em: 13 abr. 2011.

[Lee e Neuendorffer, 2000] LEE, E.; NEUENDORFFER, S. **MoML – A Modeling Markup Language in XML** – Version 0.4. Technical Memorandum ERL/UCB M 00/12. Disponível em: <http://ptolemy.eecs.berkeley.edu/publications/papers/00/moml/moml_erl_memo.pdf>. Acesso em: 13 abr. 2011.

[Ludascher et al, 2006] LUDASCHER, B. et al. **Scientific workflow management and the Kepler system**. Concurrency and Computation: Practice and Experience, Vol. 18, No. 10. pp 1039-1065, 2006.

[Marinho et al., 2010] MARINHO, A. et al. Integrating **Provenance Data from Distributed Workflow Systems with ProvManager**. Lecture Notes in Computer Science, vol 6378, pp 286-288, 2010.

[Mattoso e Cruz, 2008] MATTOSO, M.; CRUZ, S. **Gerência de workflows científicos: oportunidades de pesquisa em bancos de dados**. In Proceedings of the 23rd Brazilian symposium on Databases, 313–314. SBBD '08. Porto Alegre, Brazil, 2008.

[Mattoso et al., 2010] MATTOSO, M. et al. **Towards supporting the life cycle of large scale scientific experiments.** International Journal of Business Process Integration and Management 5, n. 1 (2010): 79 - 92.

[McPhillips et al., 2009] MC PHILLIPS, T. et al. **Scientific workflow design for mere mortals**. Future Generation Computer Systems 25, mai. 2009, pp 541–551.

[Moreau et al., 2010] MOREAU, L. et al. **The Open Provenance Model core specification (v1.1)**. Future Generation Computer Systems 27, n. 6, 2011, pp 743-756.

[Muscle, 2011] **Muscle**. Disponível em: <http://www.drive5.com/muscle/>. Acesso em: 13 abr. 2011.

[myGrid, 2011] **myGrid Project**. Disponível em: <http://www.mygrid.org.uk/>.

[MySQL, 2011] **MySQL Server**, 2011. Disponível em: <http://www.mysql.com/>.

[NCBI, 2011] **National Center for Biotechnology Information**. Disponível em: <http://www.ncbi.nlm.nih.gov/>. Acesso em: 13 abr. 2011.

[Noronha, 2006] NORONHA, M. **Controle da execução e disponibilização de dados para aplicativos sobre seqüências biológicas: o caso BLAST**. Dissertação (Mestrado).Departamento de Informática, PUC-Rio, 2006.

[Oinn et al., 2006] OINN, T. et al. **Taverna: lessons in creating a workflow environment for the life sciences.** Concurrency and Computation: Practice and Experience, vol. 18, iss. 10, pp. 1067-1100, 2006.

[Oliveira et al., 2010] OLIVEIRA, D. et al. **GExpLine: A Tool for Supporting Experiment Composition**. In: Proc. International Provenance and Annotation Workshop (IPAW'2010), LNCS 6378, pp. 251-259, 2010.

[OMG, 2011] OMG. **Business Process Model and Notation (BPMN)**. Versão 2.0. Release Date: Jan, 2011. Disponível em: <http://www.omg.org/spec/BPMN/2.0/>. Acesso em: 13 abr. 2011.

[Ooms, 2009] OOMS, M. **Provenance Management in Practice**. Master's thesis. University of Twente, Enschede, the Netherlands, 2009.

[Phylip] **PHYLIP**. Disponível em: <http://evolution.genetics.washington.edu/phylip.html>. Acesso em: 13 abr. 2011.

[ProvChallenges, 2011] **The Provenance Challenges wiki**. Disponível em <http://twiki.ipaw.info/bin/view/Challenge/>. Acesso em: 13 abr. 2011.

[Ptolemy, 2011] **Ptolemy Project**. Disponível em: <http://ptolemy.eecs.berkeley.edu/>. Acesso em: 13 abr. 2011.

[Rivest, 1992] RIVEST, R., **The MD5 Message-Digest Algorithm**. RFC-1321, MIT LCS and RSA Data Security, Inc., April 1992.

[Scheidegger et al., 2008] SCHEIDEGGER, C. et al. **Querying and Re-Using Workflows with VisTrails**. In Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 1251-1254, 2008.

[Shoshani e Rotem, 2009] SHOSHANI, A.; ROTEM, D. **Scientific Data Management**. Challenges, Technology, and deployment. Chapman & Hall/CRC Computational Science, 2009.

[Simmhan et al., 2005] SIMMHAN, Y.; PLALE, B.; GANNON, D. **A survey of data provenance in e-science**. SIGMOD Rec. 34, n. 3 (2005): 31-36.

[Soiland-Reyes, 2010] SOILAND-REYES, S., taverna-users Mailing List. Tópico: **Provenance Model**. Mensagem enviada em 03 nov. 2010. Disponível em: <http://taverna-users.markmail.org/search/?q=provenance+model#query:provenance%20model+page:1+mid:bcz2w2i7pgtgcd4u+state:results>.

[SQLite, 2011] **SQLite**. Disponível em: <http://www.sqlite.org/>. Acesso em: 13 abr. 2011.

[SVN, 2011] **SVN SubVersion**. Disponível em: <http://subversion.tigris.org/>.

[Taverna External Tool, 2011] **Calling External Tools from Taverna**. Disponível em <http://www.mygrid.org.uk/dev/wiki/display/developer/Calling+external+commands+from+Taverna>. Acesso em: 13 abr. 2011.

[Taverna Query, 2011] **Taverna Provenance Query Language**. Disponível em: <http://www.mygrid.org.uk/dev/wiki/display/provenance/Provenance+Query+Language>.

[Taverna Schema, 2011] **Taverna Provenance Schema in 2.2.0**. Disponível em: <http://www.mygrid.org.uk/dev/wiki/display/developer/Provenance+schema+in+2.2.0>.

[Taylor et al., 2007] TAYLOR, I. et al. **The Triana workflow environment: architecture and applications**. Workflows for e-Science, Springer, pp.320–339, 2007.

[Taylor et al., 2007] TAYLOR, I. et al. **Workflows for e-Science: Scientific Workflows for Grids**. 1 ed. Springer, 2007.

[Vistrails User Guide, 2011] **Vistrails User Guide**, versão 1.6.2, 5 abr. 2011. Disponível em: <http://www.vistrails.org/index.php/Users_Guide>. Acesso em: 13 abr. 2011.

[Vistrails FAQ, 2011] **Vistrails Frequently Asked Questions**. Disponível em: <http://www.vistrails.org/index.php/FAQ#How_do_I_access_the_information_in_the_execution_log.3F>. Acesso em: 13 abr. 2011.

[Wassink et al., 2009] WASSINK, I. et al. **Analysing Scientific Workflows: Why Workflows Not Only Connect Web Services**. In Proceedings of the 2009 Congress on Services - I, 314–321. Washington, DC, USA: IEEE Computer Society, 2009.

[Wassink, 2010] WASSINK, I. **Workflows in life science**. PhD thesis, University of Twente, Enschede, the Netherlands, 2010.

[Wassink et al., 2010] WASSINK, I. et al. **e-BioFlow - Improving Practical Use of Workflow Systems in Bioinformatics**. In Proc Information Technology in Bio- and Medical Informatics, ITBAM 2010. Lecture Notes in Computer Science, 2010, Vol 6266, pp 1-15.

[WinWorkflow, 2011] **Windows Workflow Foundation**. Disponível em: <http://www.windowsworkflowfoundation.eu/>. Acesso em: 13 abr. 2011.

[Zhao et al., 2006] ZHAO, Y.; WILDE, M.; FOSTER, I. **Applying the Virtual Data Provenance Model**. In: Proc. International Provenance and Annotation Workshop (IPAW'2006), Chicago, Illinois, USA, 2006. LNCS 4145, pp. 148-161.

Anexo 1.
Descrição das tabelas

As figuras 41 e 42 apresentam a descrição das tabelas obtidas a partir do mapeamento do esquema entidade-relacionamento descrito na seção 3.1 para o modelo relacional. A tabela InputValue_Artifact foi derivada do relacionamento 'is related to'. A implementação do esquema conceitual fez-se necessária para demonstrar com exemplos práticos a contribuição da modelagem proposta.

```
Workflow (workflowId, workflowVersion, workflowDef, workflowName, workflowDefHash, user)

Activity (activityId, activityVersion, activityDescription,
          activityCode, activityHash, programId)
   programId refers to ExternalResource

Parameter (parameterName, parameterRole, activityId, activityVersion)
   (activityId, activityVersion) refers to Activity

BioInput (bioData, bioModel, bioFormat, parameterName)
   parameterName refers to Parameter

PrimitiveParameter (type, defaultValue, parameterName)
   parameterName refers to Parameter

ExternalResource (programId, programName, programProvider, programVersion, programType)

Step(stepId, workflowId, workflowVersion, activityId, activityVersion)
   (activityId, activityVersion) refers to Activity
   (workflowId, workflowVersion) refers to Workflow

PortLink (portLinkId, from_stepId, from_workflowId, from_workflowVersion,
          to_stepId, to_workflowId, to_workflowVersion, from_activityId,
          from_activityVersion, from_parameterName, to_activityId,
                to_activityVersion, to_parameterName)
   (from_stepId, from_workflowId, from_workflowVersion) refers to Step
   (to_stepId, to_workflowId, to_workflowVersion) refers to Step
   (from_activityId, from_activityVersion, from_parameterName) refers to Parameter
   (to_activityId, to_activityVersion, to_parameterName) refers to Parameter

InputValue(parameterValueId, parameterValue, parameterName,  activityId,
           activityVersion, stepId, workflowId, workflowVersion)
   (parameterName, activityId, activityVersion) refers to Parameter
   (stepId, workflowId, workflowVersion) refers to Step

OPMArtifact(OPMartifactId, artifactValue, artifactShortValue, artifactMD5,
           artifactPath, opmProcessId, role)
    opmProcessId refers to Process
```

Figura 41 – Descrição das tabelas – parte 1

```
OPMProcess(OPMProcessId, started, finished, stepId, workflowId, workflowVersion,
           OPMGraphId)
   (stepId, workflowId, workflowVersion) refers to Step
   OPMGraphId refers to OPMGraph

OPMGraph (OPMGraphId, workflowId)
    workflowId refers to Workflow

OPMUsed (OPMProcessId, OPMArtifactId, OPMRole)
   OPMProcessId refers to OPMProcess
   OPMArtifactId refers to OPMArtifact

OPMWasGeneratedBy (OPMProcessId, OPMArtifactId, OPMRole)
   OPMProcessId refers to OPMProcess
   OPMArtifactId refers to OPMArtifact

OPMWasDerivedFrom (effectOPMArtifactId, causeOPMArtifactId, OPMRole)
   effectOPMArtifactId refers to OPMArtifact
   causeOPMArtifactId refers to OPMArtifact

OPMWasTriggeredBy (effectOPMProcessId, causeOPMProcessId)
   effectOPMProcessId refers to OPMProcess
   causeOPMProcessId refers to OPMProcess

Annotation (annotationId, property, value, OPMProcessId)
   OPMProcessId refers to OPMProcess

InputValue_Artifact (parameterValueId, OPMartifactId)
```

Figura 42 – Descrição das tabelas – parte 2

Anexo 2.
Exemplos de arquivos de especificação

Na seção 4.3 foram omitidos os arquivos descritores de workflows e atividades nas instâncias apresentadas para facilitar a exibição das instâncias. A seguir apresentamos na Figura 43 o arquivo de definição do workflow de Geração de Árvore Filogenética, e na Figura 44 o arquivo de definição da atividade Clustalw, ambos gerados pelo BioSide. Os arquivos foram reduzidos para facilitar a visualização de sua estrutura.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE scenario PUBLIC "-//ENSTBr//DTD Bioside Scenario 2.0//EN">
<scenario>
    <id>ACE90CE589F1073296C0200007870000000074001...2416C69436C7573745F526566696E655F4E4A</id>
    <name>AliClust_Refine_NJ</name>
    <user>bioside</user>
    <inputs>
        <input type="file" x="25" y="250" name="" id="0">
            <infile>f0-1</infile>
        </input>
    </inputs>
    <programs>
        <program idref="bob___muscle___3.7" x="250" y="300" info="bob muscle 3.7" id="muscle_alignment-1">
            <parameter idref="guidetree">
                <data />
            </parameter>
            <parameter idref="alignment1">
                <data>Clustalw-1.A_fasta_outfile</data>
            </parameter>
            ...
        </program>
        <program idref="phylip___seqboot___3.65" x="500" y="75" info="phylip seqboot 3.65" id="seqboot-1">
            <parameter idref="infile">
                <data>muscle_alignment-1.phyiout</data>
            </parameter>
            ...
        </program>
        .
        .
        .
    </programs>
</scenario>
```

Figura 43 – Arquivo descritor do workflow de Geração de Árvore Filogenética

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE program_description PUBLIC "-//Telecom Bretagne/DTD XML Praxis Program Description 3.0//EN">

<program_description export_date="2010-01-12 15:54:56" modification_date="2009-11-10 10:12:04.0">
  <program provider="ght" name="clustalw" version="2.0.11" />
  <from desc_id="S60" status="test" />
  <description>Clustalw2 by Gibson Higgins Thompson</description>
  <parameters>
    <parameter id="clustalw2" type="command" ismandatory="1">
      <name>clustalw2</name>
      <description>Clustalw2 by Gibson Higgins Thompson</description>
    </parameter>
    <parameter id="SA_infile" ishidden="0" type="input">
      <name>sequences  (GCG/MSF, EMBL, clustalw format, NBRF-PIR, fasta)</name>
      ...
    </parameter>
    <parameter id="A_fasta_outfile" ishidden="0" type="output">
      <name>alignment  (fasta)</name>
      ...
    </parameter>
    <parameter id="IHM_operation" ishidden="0" type="enum">
      <name>operation type</name>
      <vlist>
        <item id="T1">
          <description>multiple sequences alignment</description>
          <code />
        </item>
        <item id="T4">
          <description>phylogenetic tree</description>
          <code>4%n4%n</code>
        </item>
        ...
    </parameter>
  </parameters>
</program_description>
```

Figura 44 – Arquivo descritor da atividade Clustalw