Transparência em Experimentos Científicos Apoiados Em Proveniência: Uma Perspectiva para Workflows Científicos Transparentes

André Luiz de Castro Leal[1], Sergio Manuel Serra da Cruz[1,2,3]

[1] ICE/DEMAT – Universidade Federal Rural do Rio de Janeiro
Seropédica, RJ, Brasil

[2] Programa de Pós Graduação em Modelagem Matemática e Computacional

[3] PET-SI - Programa de Educação Tutorial – Sistemas de Informação

andrecastr@gmail.com, serra@ufrrj.br

Resumo. *Este trabalho aplica o SIG de Transparência no domínio de experimentos científicos in silico baseados em workflows científicos (WfC). A demanda da aplicação está inserida no contexto da relação humana uma vez que workflows são fruto de uma construção social, com diferentes participantes e suas expectativas, onde entende-se que informações e processos científicos necessitam ser transparentes para que possam potencializar as características do contexto social. Para isso sugere-se nesse artigo um SIG de Transparência aderente ao contexto de WfC para experimentos científicos in silico com a expansão do softgoal de Rastreabilidade a partir da decomposição de características de Proveniência.*

Abstract. *This work applies the GIS Transparency in the field of science in silico-based scientific workflows (WFC) experiments. The demand of the application is embedded in the context of human relationship, since workflows are the result of a social construction, with different participants and their expectations, which means that scientific information and processes need to be transparent so they can leverage the features of the social context. For this, it is suggested in this paper a SIG of Transparency adherent to the context of WFC for scientific in silico with the expansion of softgoal traceability from the decomposition characteristics of Provenance.*

1. Introdução

Workflows têm uma definição peculiar que segundo o *Workflow Management Coalition*[1] (WFMC) perpassa pela automação de um processo de negócio, total ou parcial, na qual documentos, informações ou tarefas são transferidos entre participantes de acordo com um conjunto de regras geralmente aplicadas aos fluxos de processos do ambiente corporativo. Apesar de sua inicial aplicação em ambientes corporativos, os *workflows* atualmente são amplamente utilizados ambientes científicos, tais como laboratórios de pesquisa de universidades ou mesmo de grandes complexos industriais.

Muitos dos processos experimentais escalaram de isolados para distribuídos devido à complexidade imposta pelos problemas a serem analisados, ao crescente volume de dados manipulados e necessidade elaboração de times de pesquisa

[1] http://www.wfmc.org/

geograficamente dispersos. Este processo denomina-se ciência em larga escala (*e-Science*) e os novos experimentos mediados por computadores baseados em aparatos tecnológicos sofisticados passaram a se chamar experimentos científicos *in silico*. Os *workflows* científicos (*WfC*) são ferramentas importantes da *e-Science,* eles têm por objetivo representar processos encadeados (também conhecidos como atividades) e uma combinação de dados consumidos ou produzidos por estes. Eles representam uma alternativa para o tratamento do problema do sequenciamento estruturado de passos necessários ao desenvolvimento de experimentos científicos, reforçando ainda mais as necessidades de garantias de reprodutibilidade e confiabilidade desses experimentos [Altintas et al. 2006]. Atualmente, os *WfC* compartilham das mesmas características dos *workflows* de negócio. Os científicos possuem aspectos de especificação e implementação que lhes são peculiares, como suas etapas de composição, execução e análise [Mattoso e Cruz 2008] [Cruz et al. 2009] que podem levar meses para serem finalizadas e gerar resultados previstos ou inesperados. Tais resultados necessitam ser registrados para efeito de avaliação dos experimentos; compartilhamento de dados; garantia de reprodução ou mesmo para fins de auditoria e segurança. As ferramentas de maior uso para construção *WfC* são chamadas de Sistemas de Gerência de *Workflow* Científico (*SGWfC*), dão suporte a: construção dos fluxos (composição), captura de metadados, como também no suporte à execução, coleta, análise e tratamento da proveniência dos dados e processos [Simmhan et al. 2005].

Fortemente relacionado aos *WfC* está o conceito de proveniência [Altintas et al. 2006] [Mattoso e Cruz 2008]. Em experimentos científicos são gerados registros sobre a computação dos fluxos que procuram caracterizar o processo científico ou os dados que são manipulados ao longo do processo. A importância da proveniência tem crescido nos últimos anos e tem sido utilizada por pesquisadores de diversas áreas para manter descritores sobre os rastros de dados e processos, auxiliando-os a desvendar como foram produzidos e também assegurar a qualidade dos procedimentos que os produziram [Miles 2007].

O presente trabalho analisa o alinhamento entre os conceitos de Transparência e Proveniência no contexto da pesquisa científica apoiada por *WfC* para potencializar o uso do que se pode denominar *Workflows Científicos Transparentes* (*WfCT*). Está subdividido com segue: na seção 2 são apresentados os objetivos da pesquisa; na seção 3, as contribuições científicas; na seção 4, os resultados iniciais já alcançados; e na seção 5 são apresentadas as conclusões e perspectivas de trabalhos futuros.

2. Objetivos da Pesquisa

A análise da literatura científica com relação a duas qualidades, Transparência e Proveniência, no contexto de *WfC*, sugere que há uma forte interseção entre suas características. A importância da caracterização de Transparência inserida no contexto científico se dá principalmente pela presença de diferentes atores envolvidos na gestão de *WfC*, atores esses que possuem diferentes perfis (*skills*). Bourdieu (1976) cita que *a sociologia da ciência repousa no postulado de que a verdade do produto – mesmo em se tratando desse produto particular que é a verdade científica – reside numa espécie particular de condição social de produção*.

No contexto da Transparência de Processos Organizacionais [Cappelli 2009], o termo Rastreabilidade aparece como uma decomposição de Auditabilidade. A Transparência é realizada através da Auditabilidade, ela é a capacidade de identificar

através da aferição de práticas que implementam características de Explicação, Rastreabilidade, Verificabilidade, Validade e Controlabilidade [Cappelli 2009]. Como são metas a serem aferidas com características de qualidade, as mesmas são denominadas no SIG como metas-flexíveis (*softgoals*). Atrelado ao conceito de rastreabilidade (*traceability*) está o conceito de Proveniência (*provenance*), amplamente utilizado na área de Banco de Dados [Mattoso e Cruz 2008] [Simmhan et al. 2005] [Miles 2007] e [Moreau 2010]. Duas de suas características, *derivação* e *anotação*, são um indicativo da aderência de características de proveniência com transparência.

A *derivação* consiste em registrar o processo (ou atividade do *WfC*) pelo qual os dados passam para gerar um determinado resultado, o que pode incluir, por exemplo, parâmetros de entrada, arquivos e bancos de dados utilizados, queries, ou mesmo a própria concepção de um *WfC Abstrato* [Cruz et al 2012]. As *anotações* consistem em informações descritivas sobre objetos persistidos, como data de criação, data de atualização, data de execução, nomes e formatos de arquivos, assim como anotações mais estruturadas que descrevem, por exemplo, a semântica do objeto no domínio ao qual ele pertence [Cruz et al 2012] e [Moreau 2010]. Outras classificações de Proveniência comum às áreas de *WfC* são as *proveniências prospectiva e a retrospectiva* [Zhao et al. 2006]. A *prospectiva* corresponde à definição dos procedimentos de alto nível adotados pelo cientista para conceber um experimento mapeado como um *WfC* concreto; a *retrospectiva* corresponde a todas as informações sobre a materialização do *WfC Abstrato* em *WfC Concreto* e das execuções dos processos em ambientes computacionais, correlacionando-os com a geração do resultado (*WfC* concreto e suas atividades), tais como parâmetros, dados de entrada e saída e outras informações referentes à execução do processo propriamente dita.

3. Contribuições Esperadas

Entendemos que, por haver o envolvimento de diferentes atores no ciclo de vida de um experimento científico baseado em *WfC* [Cruz et al 2012], tais atores em muitas ocasiões podem desempenhar papéis e ter expectativas com forte demanda por Transparência da Informação.

A contribuição desse trabalho de pesquisa é fomentar a Transparência nos experimentos científicos *in silico* em *e-Science* a partir da caraterização do SIG de Transparência [Cappelli 2009] com características de Proveniência. Dessa forma, este trabalho propõe que *WfC* sejam concebidos sob a ótica de Transparência, permitindo que se torne em um *workflow* científico transparente (*WfCT).* A nova caracterização pretende expor à comunidade científica que ao incorporar a Transparência será possível aumentar significativamente não só qualidade dos resultados científicos baseados em *WfC,* mas também ampliar o grau de confiabilidade de experimentos e procedimentos experimentais, expandir sua reprodutibilidade e também oferecer novos argumentos apoiados em métricas comuns à Engenharia de Requisitos que ampliem a confiança (*trust*) em relação aos dados, redução da propagação de erros experimentais e incrementos na segurança dos experimentos. Uma abordagem utilizada nessa pesquisa é explicitar os elos de relação entre Proveniência e demais *softgoals* de Transparência, bem como expandir o *softgoal* de Rastreabilidade a partir da decomposição das características de Proveniência com o objetivo de tornar o SIG aderente ao seu tópico principal, *WfC in silico*.

4. Primeiros Resultados

Associar o SIG de Proveniência [Leal et al. 2014] ao presente trabalho é um ponto importante da pesquisa e em [Leal et al. 2014] os autores relacionam os *softgoals* de Transparência, Proveniência e Confiança em um modelo único para demonstrar a interseção entre diferentes modelos para sistemas de informação que necessitem trabalhar conjuntamente com tais características. Para uma primeira análise isolada é apresentado o SIG de Proveniência [Leal et al. 2014] na Figura 1. Nela há a uma decomposição, baseada no *framework* NFR [Chung 2000], de características de proveniência que foram elaboradas a partir da análise documental de literatura científica baseada em [Simmhan et al. 2005] [Cruz et al 2012] e [Moreau 2010].

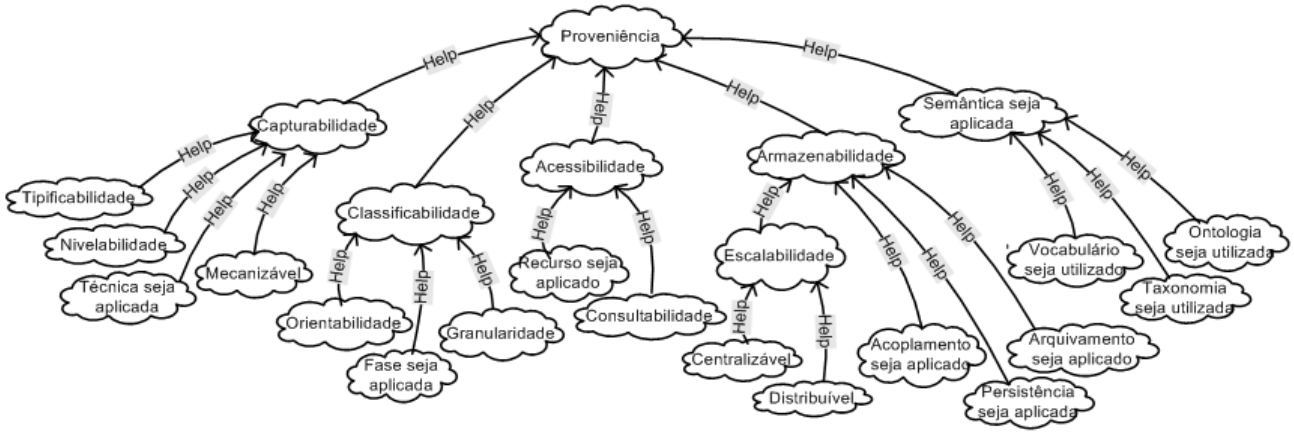

Figura 1. SIG de Proveniência proposto em [Leal et al. 2014].

Nesta proposta, defendemos que, do ponto do *softgoal* de Rastreabilidade do SIG de Transparência a expansão a partir de características de Proveniência torna o SIG de Transparência aplicado ao domínio de *WfC*, conforme apresentado na Figura 2. Em [Chung 2000], o reuso do conhecimento sobre requisitos não funcionais a partir de SIG pode ser feito e é orientação indicar por tópico, rótulo entre colchetes, o contexto ao qual o SIG é aplicado, no caso dessa pesquisa o tópico refere-se à: WfC para experimentos científicos *in silico*. O *softgoal* com um círculo de linhas tracejadas representa o ponto de expansão de Rastreabilidade valorada pelo tópico Proveniência.

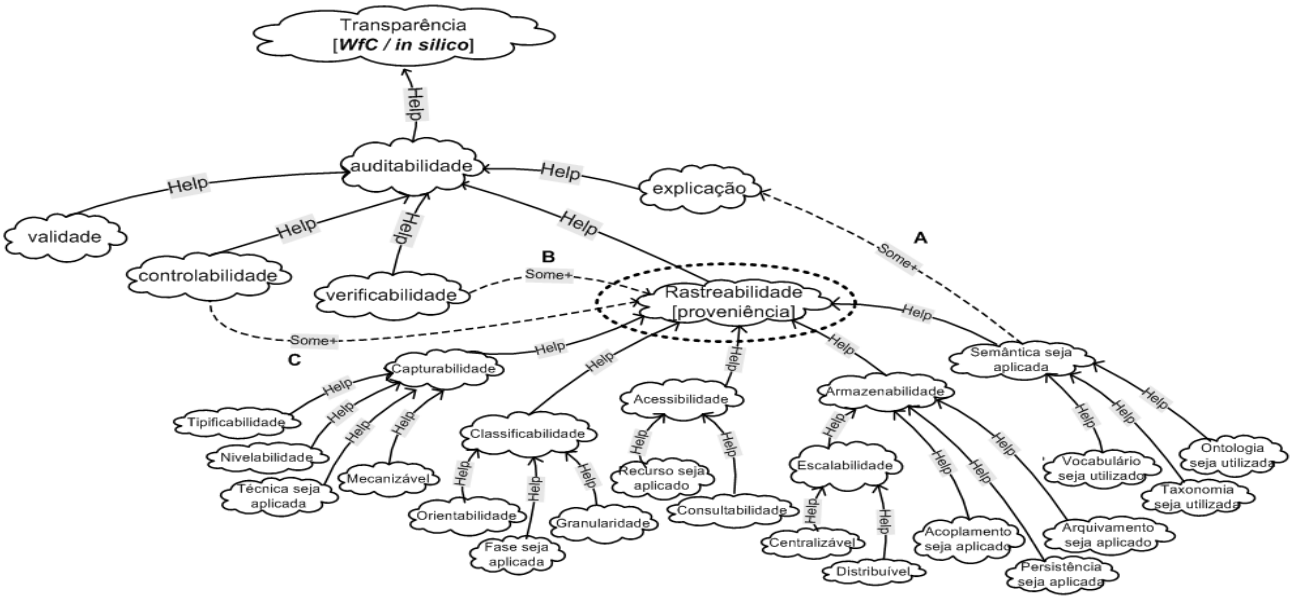

Figura 2. SIG de Transparência com tópico em WfC.

Através do SIG de Transparência inicialmente elaborado em [Cappelli 2009] para processos organizacionais passa a ter em seu *softgoal* Rastreabilidade em uma estrutura mais adequada ao contexto dos *WfC*. A Rastreabilidade com suporte de Proveniência potencializa como elo de contribuição *HELP* o *softgoal* de Auditabilidade e consequentemente a Transparência. Essa é uma análise vertical no sentido da Transparência, mas há também análises horizontais, ou seja, entre *softgoals*, que também estão presentes no modelo representados a partir dos elos de relação positivo do tipo *SOME+*. As contribuições positivas em análises horizontais possuem potencialização de um *softgoal* da ramificação de Transparência e consequentemente também potencializa a Transparência. É importante deixar claro que nesse momento da pesquisa estamos apenas focando no ramo de Auditabilidade, mas pelas leituras iniciais foram identificadas outras contribuições de *softgoals* de outras ramificações do SIG de Transparência que não serão detalhados nesse trabalho.

Outras relações podem ser vistas no grafo (linhas tracejadas: A, B e C), ora com origem em proveniência para algum *softgoal* já presente no catálogo de Transparência, ora com proveniência sendo o destino de um elo de relação. Ressalta-se que o importante é perceber que há uma potencialização bilateral e, que ao serem aplicadas as características de Transparência estas estarão contribuindo positivamente para Rastreabilidade dos processos de concepção, materialização e execução de *WfCT* e vice-versa.

5. Conclusão e Trabalhos Futuros

O presente trabalho apresenta características de Transparência e Proveniência no contexto de *WfC* para experimentos científicos *in silico*. Pretende-se chamar a atenção para a adoção de características de Transparência em *WfC* sob a denominação de *WfCT*.

A Transparência é um importante fator que desponta atualmente nas relações humanas sendo intrínseca aos novos ambientes de pesquisa, entendemos que tornar transparente os mecanismos de concepção, elaboração, execução e análise de resultados, processos, experimentos e outros elementos envolvidos no contexto da *e-Science* poderá aumentar significativamente a qualidade intrínseca das pesquisas científicas que

manipulam grandes volumes de dados e podem executar seus processos em ambientes heterogêneos e distribuídos. Entendemos que pesquisas científicas apoiadas por experimentos computacionais são estruturadas, mas muitas vezes pouco documentadas, a aplicação das características de proveniência podem facilitar a análise dos processos científicos e consequentemente na divulgação dos resultados, que poderiam receber contribuições de um estudo de Transparência.

Como trabalho futuro serão investigadas operacionalizações [Chung 2000] no SIG *WfCT* para demonstrar como devem ser construídos *WfC in silico* que possam ter Transparentes. Outra perspectiva de trabalho futuro é estabelecer, a partir de critérios técnicos, novas relações de correlação positiva entre *softgoals* do modelo, bem como explorar correlações negativas que possam anular as positivas. Dessa forma, explicitar obstáculos para atingir as metas definidas.

Agradecimentos

À FAPERJ pelos financiamentos (E-26/112.588/2012) e (E-26/110.928/2013) e ao FNDE/MEC/SeSU.

Referências

Altintas, I. O.; Barney, O.; Jaeger-Frank, E. Provenance collection support in the kepler scientific workflow system. In: Proc. Int. Provenance and Annotation Workshop (IPAW'2006), Chicago, Illinois, USA. LNCS 4145, pp. 118--132 (2006).

Mattoso, M.; Cruz, S.. Gerência de workflows científicos: oportunidades de pesquisa em bancos dedados. In Proc. of the 23rd Brazilian Symposium on Databases, 313–314. SBBD'08. Porto Alegre, Brazil (2008).

Simmhan, Y. L., Plale, B., Gannon, D.. A survey of data provenance in e-science. SIGMOD Record, v. 34, n. 3, pp. 31--36 (2005).

Miles, S.; Groth, P.; Branco, M.; Moreau, L. The requirements of using provenance in e-science experiments, Journal of Grid Computing, 5, 1–25 (2007).

Cappelli, C. Uma Abordagem para Transparência em Processos Organizacionais Utilizando Aspectos. Rio de Janeiro. 328 p. Tese de Doutorado – Departamento de Informática, PUC-Rio (2009).

Zhao, Y.; Wilde, M.; Foster, I. Applying the Virtual Data Provenance Model. In: Proc. Int. Provenance and Annotation Workshop (IPAW), Chicago, Illinois, USA, 2006. LNCS 4145, pp. 148--161 (2006).

Chung, L.; Nixon, B. A.; Yu, E., and MYLOPOULOS, J. Non-Functional Requirements in Software Engineering. Springer (2000).

Leal, A. L. C.; Sousa, H. P.; Leite, J. C. S. P., Modelo orientado à meta para estabelecer relações de contribuição mútua entre Proveniência, Transparência e Confiança. In: XVII Ibero-American Conference on Software Engineering & XVII Workshop on Requirements Engineering, 2014, Pucón, Chile. Proceedings of XVII Cibse & XVII Wer (2014).

Cruz, S. M. S., Campos, M. L. M., Mattoso, M., Towards a Taxonomy of provenance in Scientific Workflow Management Systems". In: Proceedings of the SERVICES '09 Congress on Services - I, pp. 259--266. Los Angeles, Califórnia, Jul (2009).

Moreau. L. The foundations for provenance on the web. Foundations and Trends in Web Science, 2(2-3):99-241, November (2010).

Bourdieu, P. Le Champ Scientifique. Actes de Ia Recherche en Sciences, Sociales, n. 2/3, Jun, pp. 88--04. Tradução Paula Montero (1976).