

## E-SECO ProVersion: Manutenção e Evolução de Experimentos Científicos

**Tassio F. M. Sirqueira<sup>1,2</sup>, Humberto L. O. Dalpra<sup>1</sup>, Regina Braga<sup>1</sup>,  
Marco A. P. Araújo<sup>1,2</sup>, José Maria N. David<sup>1</sup>, Fernanda Campos<sup>1</sup>**

<sup>1</sup>Programa de Pós-graduação em Ciência da Computação –  
Departamento de Ciência da Computação – Universidade Federal de Juiz de Fora  
(UFJF) – Juiz de Fora – MG – Brasil

<sup>2</sup>Instituto Federal de Educação, Ciência e Tecnologia do Sudeste de Minas Gerais  
(IF Sudeste MG) – Campus Juiz de Fora – Juiz de Fora – MG – Brasil

{tassio.sirqueira, marco.araujo}@ice.ufjf.br, humbertodalpra@gmail.com  
{regina.braga, jose.david, fernanda.campos}@ufjf.edu.br

***Abstract.** This paper discusses some characteristics of a scientific software ecosystem and the life cycle of an experiment, with emphasis on maintenance and evolution aspects. The E-SECO ProVersion is presented, that aim to support maintenance and evolution of experiments using data provenance. An extension of PROV model is presented together with an ontology, named PROV-OEXT. The article also presents existing workflows repositories and discuss the support that they provide to evolution and maintenance.*

**Resumo.** Este trabalho discute algumas características de um ecossistema de software científico e do ciclo de vida de um experimento, com ênfase em manutenção e evolução dos experimentos, usando dados de proveniência. Uma extensão do modelo PROV é apresentada conjuntamente com uma ontologia, PROV-OEXT. O artigo também apresenta repositórios de *workflows* existentes e discute o suporte provido por esses no que tange a manutenção e evolução.

### 1. Introdução

Um experimento científico pode ser definido como um conjunto de atividades (análises) interligadas entre si [GOBLE *et al.*, 2010]. O ciclo de vida de um experimento é composto por diversas etapas, desde a concepção do problema até a obtenção dos resultados conforme proposto por Belloum *et al.* (2011). Todas essas etapas do experimento devem ser registradas de forma a manter o registro da modelagem e execução do mesmo.

Um *workflow* científico é um modelo ou *template* que representa a sequência de atividades implementadas por ferramentas, programas ou serviços [DEELMAN *et al.*, 2009]. Um ou mais *workflows* científicos podem ser utilizados para a execução de um experimento. *Workflows* científicos são interpretados e executados por Sistemas Gerenciadores de *Workflows* Científicos (SGWfC). No geral, os SGWfC limitam-se a gerenciar a execução de *workflows* científicos de forma isolada ao experimento do qual fazem parte. Assim, de acordo com Hasan *et al.* (2007), faz-se necessário o uso de ferramentas independentes do SGWfC para apoiar o desenvolvimento do experimento

científico, sendo necessário o registro das variações dos *workflows* associados ao experimento, devido as modificações no decorrer da pesquisa [MATTOSO *et al.*, 2009].

Neste contexto, podemos considerar que todas as informações a respeito do experimento científico fazem parte de sua gerência de configuração [MATTOSO *et al.*, 2009]. Assim, avanços e mudanças no experimento devem ser registrados, criando uma base de conhecimento sobre a pesquisa. O registro dessas informações pode se tornar ainda mais útil se aplicado ao conceito de laboratórios colaborativos [VAZ *et al.*, 2012], onde pesquisadores, geograficamente dispersos, estão trabalhando em um mesmo experimento e as atividades e dados de análises devem ser registrados e compartilhados com os demais pesquisadores do grupo, evitando a perda de conhecimento e do controle sobre os dados do experimento. Para isso, é necessário o armazenamento tanto dos dados quanto dos processos que os geraram. Uma das abordagens para se realizar esse registro é o uso de modelos de proveniência [MATTOSO *et al.*, 2009].

Alguns SGWfC como o Taverna [OINN *et al.*, 2007], Kepler [ALTINTAS *et al.*, 2004] e Pegasus [GIL *et al.*, 2007], permitem capturar os passos do *workflow* (processo e dados) durante sua execução. Esses sistemas, em geral, adotam modelos proprietários para capturar os traços de proveniência gerados nas execuções. Com a ausência da padronização entre os SGWfC é difícil a interoperabilidade dos dados, bem como a consulta e a análise pelo pesquisador. Faz-se necessário um modelo de proveniência de dados que permita a captura de proveniência retrospectiva e prospectiva, mantendo o padrão para qualquer SGWfC. Proveniência de dados é o registro da história da derivação dos dados, que possibilita a reprodutibilidade, interpretação dos resultados e diagnóstico de problemas [LIM *et al.*, 2010]. Estas informações podem ser usadas para identificar métodos, regras, auxiliar os usuários na criação de *workflows* semelhantes, na compreensão de correlações de dados e na experiência para futuros experimentos [MOREAU *et al.*, 2011]. Neste trabalho é utilizado o modelo de proveniência de dados PROV [MOREAU e MISSIER, 2013], padronizado pela W3C.

Todos os dados referentes ao experimento fazem parte de sua gerência de configuração, a qual deve acompanhar todos os passos da pesquisa, utilizando-se da proveniência de dados. Além desta, informações como derivações, manutenções e evolução do experimento e *workflow* vinculados, contribuem para a compreensão do mesmo. Para apoio ao respectivo controle, pode-se utilizar, como arcabouço, as técnicas de manutenção e evolução de software.

Assim, propõe-se o desenvolvimento de uma arquitetura para suporte a manutenção e evolução de experimentos científicos, fornecendo, de maneira automatizada, informações estratégicas relacionadas a evolução e manutenção do experimento, de forma que os cientistas possam tomar decisões ou obter maior conhecimento em relação ao mesmo. Esta arquitetura denomina-se E-SECO ProVersion, e é parte da abordagem E-SECO [FREITAS *et al.*, 2015]. O E-SECO ProVersion busca adicionar funcionalidades específicas para a gerência de configuração de experimentos científicos no contexto de ecossistemas de software científicos e de laboratórios colaborativos

O artigo é composto por 3 seções, além da introdução. A seção 2 apresenta os trabalhos relacionados. Na seção 3 é apresentada a proposta do E-SECO ProVersion,

sua arquitetura, seu desenvolvimento e uma breve análise junto a base do myExperiment. Por fim, a seção 4 apresenta as conclusões e os trabalhos futuros.

## 2. Trabalhos Relacionados

O SimiFlow é uma arquitetura para comparação e agrupamento de *workflows* pré-existentes, por similaridade, visando a construção de vários experimentos por meio de abordagem ascendente [SILVA *et al.*, 2010]. O CollabCumulus é um portal com conteúdo de repositórios de proveniência que realiza a análise dos dados gerados ou consumidos [MIRANDA *et al.*, 2014]. O PBASE [CUEVAS-VICENTTÍN *et al.*, 2014] é uma extensão do modelo de proveniência PROV destinada a *workflows* científicos, permitindo análise e replicação de experimentos.

O myExperiment é um ambiente colaborativo de compartilhamento e publicação de *workflows* [GOBLE *et al.*, 2010]. O CrowDLabs é um repositório similar que foi desenvolvido pelo grupo do SGWfC VisTrails [CALLAHAN *et al.*, 2006]. Permite a execução de *workflows*, importação de dados e reutilização por terceiros, porém não permite a análise detalhada de seus dados.

O diferencial da proposta do E-SECO ProVersion em relação a estes trabalhos é o suporte explícito a manutenção e evolução de experimentos científicos. Para isso utiliza uma base de dados de proveniência, modelada segundo o padrão PROV, que garante a interoperabilidade dos dados entre experimentos. Por meio de uma ontologia, apoio de regras de inferência para a descoberta de informações implícitas, e o acesso à repositórios de *workflows*, objetiva-se a formação de uma base histórica de dados de experimentos, com vistas a um melhor suporte a manutenção e evolução dos mesmos.

## 3. E-SECO ProVersion

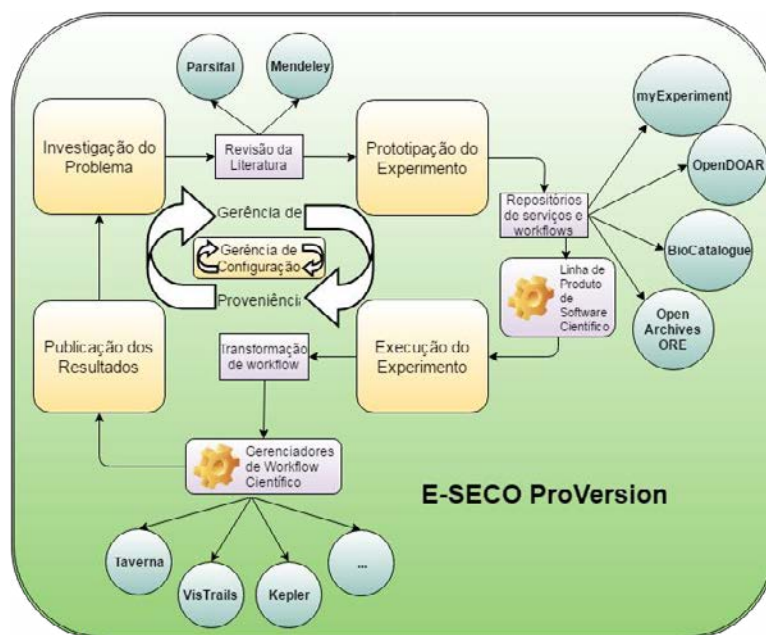
O uso de *workflows* científicos é uma abordagem bastante utilizada no contexto de e-Science e existem muitas pesquisas voltadas para o gerenciamento e execução de experimentos baseados em *workflows*. No entanto, experimentos complexos envolvem interações entre pesquisadores geograficamente distribuídos, podendo caracterizar-se como laboratórios colaborativos, e que demandam a utilização de grandes volumes de dados, serviços e recursos computacionais distribuídos. Este cenário categoriza um ecossistema de experimentação científica [FREITAS *et al.*, 2015].

Neste contexto, foi proposta uma abordagem baseada em ecossistemas de software [BOSCH, 2009], denominada E-SECO (E-Science Software ECOsystem) [FREITAS *et al.*, 2015]. O E-SECO possui um ciclo de vida baseado na proposta de Belloum *et al.* (2011). Este ciclo de experimentação foi expandido para englobar a abordagem E-SECO ProVersion, a fim de viabilizar o suporte a gerência de configuração de experimentos, conforme Figura 1.

A gerência de configuração deve ser entendida como uma etapa de suma importância para o ciclo de vida de um experimento e dos *workflows* vinculados. Através desta pode-se determinar o estado do experimento em um determinado momento, o que, considerando o contexto de um laboratório colaborativo, pode derivar informações importantes para acertos futuros, tais como resultados parciais, erros encontrados ou gerados, entre outras informações, durante a execução do experimento.

Além disso, permite-se prever o comportamento do experimento no futuro e, no caso dos *workflows*, como os mesmos devem ser mantidos e evoluídos.

Conforme detalhado por Mattoso *et al.* (2009), o apoio a reutilização e a gerência de configuração deve ser tratado com um item importante a ser explorado, de forma a diminuir o retrabalho e apoiar o aumento de produtividade e de qualidade dos experimentos dos pesquisadores. São necessárias soluções mais abrangentes, que permitam tratar as manutenções e evoluções dos *workflows* ao longo do seu ciclo de experimentação e permitir a reutilização do próprio experimento.



**Figura 1. Ciclo de vida de um experimento científico no E-SECO ProVersion.**

Na proposta E-SECO ProVersion, a gerência de configuração de um experimento está dividida em duas etapas, Gerência de Proveniência e Gerência de Manutenção e Evolução. A Figura 2 apresenta a arquitetura do E-SECO ProVersion, integrando os módulos relacionados à gerência de configuração. O gerente de manutenção e evolução é responsável pelo controle de dados de proveniência prospectiva e retrospectiva de um experimento, com o objetivo de controlar a evolução e as manutenções existentes no experimento e *workflows* a ele vinculado, além disso, considera que o experimento pode ser composto por múltiplos *workflows*, e para cada um são geradas diversas versões que devem ser controladas. Além das informações do experimento a qual faz parte, informações acerca do comportamento do *workflow*, resultados de execução e controle dos dados consumidos e gerados no experimento também são controlados pelo módulo. O módulo de proveniência, também pertencente ao gerente de configuração, é responsável pela coleta e armazenamento dos dados capturados durante a execução dos *workflows*, que são utilizados pelo módulo de manutenção e evolução. Como o E-SECO ProVersion está inserido no contexto de um ecossistema e a interoperabilidade dos dados é um fator importante, o uso de um modelo de proveniência padrão, amplamente aceito pela comunidade, é importante. Neste sentido, o modelo PROV (MOREAU & MISSIER, 2013) foi o escolhido para ser utilizado na proposta deste trabalho, compondo a gerência de proveniência. A modelagem do banco, a qual segue as regras do modelo PROV, pode ser acessada no

seguinte endereço <http://goo.gl/LvP8Aq>. Este modelo de proveniência permite acompanhar as etapas do experimento, segundo o modelo de ciclo de vida definido, criando uma base de conhecimento sobre a pesquisa. Além da coleta e o armazenamento em uma base de dados, uma ontologia integrada ao módulo é utilizada, juntamente com os dados do experimento, para auxiliar na extração do conhecimento por meio de inferências, possibilitando a identificação de *workflows* com tarefas similares, fluxos de trabalho próximos e características de manutenção e evolução.

O modelo PROV já conta com o suporte de uma ontologia denominada PROV-O [LEBO *et al.*, 2013], que é pública e disponibilizada pela W3C. No entanto, a PROV-O não expressa todo o conhecimento necessário para o suporte a evolução e manutenção de *workflows* e experimentos. O trabalho de expansão da ontologia PROV-O, através da criação da ontologia PROV-OEXT, permite a descoberta de novas informações, tanto para a captura da proveniência prospectiva quanto retrospectiva. Algumas das relações causais da ontologia PROV-OEXT são exibidas na Figura 3. Com essas informações, o E-SECO ProVersion pode sugerir mudanças estratégicas tanto na modelagem quanto na execução do experimento ou *workflows* associados.

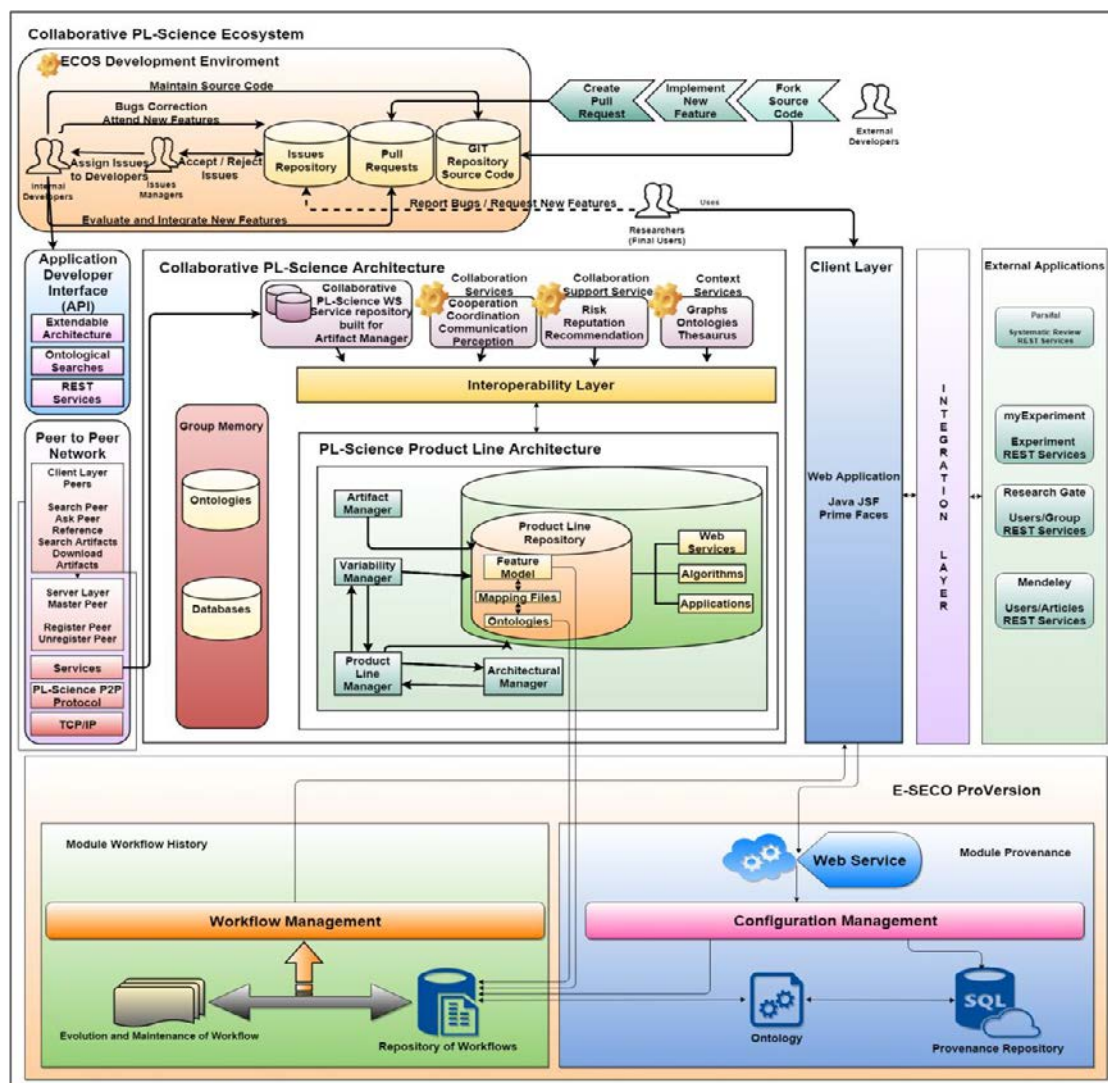


Figura 2. Arquitetura do E-SECO ProVersion.



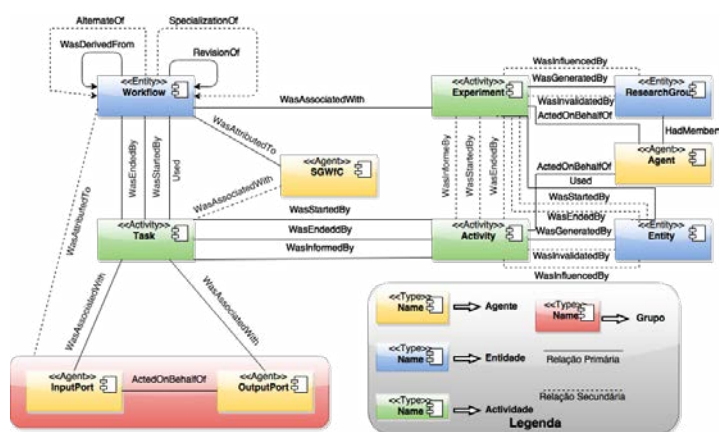


Figura 3. Relações causais da ontologia PROV-OEXT.

Além disso, no contexto de manutenção e evolução de software, uma fonte importante de informação são os dados históricos. A falta deste tipo de informação dificulta a análise dos resultados do experimento, o que pode impedir ou atrapalhar a reutilização dos *workflows* e dos experimentos, visto que não se conhece claramente a origem dos mesmos. Com o objetivo de mapear dados históricos, extraídos de repositórios existentes, foi realizado um estudo nos repositórios de *workflows* disponíveis, a fim de analisar a manutenção e evolução dos *workflows*. O referido estudo foi realizado junto aos repositórios myExperiment e CrowDLabs, ao longo do mês de outubro de 2015. No repositório CrowDLabs não foi possível extrair nenhum dado, devido a necessidade de permissão de acesso a base. No myExperiment constatou-se que entre os 3692 *workflows* disponíveis na base, 1571 utilizam os SGWfCs Taverna, Kepler ou VisTrails, os quais são os principais SGWfC, representando mais de 40% do total. Destes, 1520 são *workflows* desenvolvidos no SGWfC Taverna, 47 no SGWfC Kepler e 4 no VisTrails, destacando o Taverna como o principal em utilização entre os membros do repositório. Apesar de ser um número considerável, a quantidade é inferior a metade do total disponível no repositório, o que pode enviesar o estudo. Com relação ao histórico de versão dos *workflows*, observou-se que entre os 1571 *workflows* analisados, apenas 29% possuem dados de versionamento, o que reitera a falta de informações sobre o ciclo de vida do *workflow*, dificultando a sua reutilização.

Foi também realizada uma análise para a verificação das tarefas mais utilizadas entre os *workflows* estudados. Essa análise é útil por apresentar a quantidade de *workflows* que são afetados, caso uma tarefa seja modificada, o que irá impactar diretamente nos experimentos a que o *workflow* está vinculado. A lista das 10 tarefas mais utilizadas pode ser vista na Figura 4. Considerando a mineração de dados históricos, no repositório do myExperiment não existem informações sobre a proveniência dos *workflows* disponibilizados e os experimentos em que foram aplicados. Do ponto de vista da manutenção e evolução, tais informações se tornam essenciais para entender como os mesmos foram mantidos e evoluídos ao longo de um ciclo de experimentação.

A falta de dados históricos adequados dificulta a análise dos *workflows* e consequentemente dos experimentos pelo E-SECO ProVersion. Entretanto foi desenvolvido no contexto do E-SECO ProVersion um meta-repositório já preparado para extração e mineração destes dados de forma a contribuir com as pesquisas em e-

Science. O objetivo é acessar repositórios de *workflows* existentes e complementar com informações necessárias para o controle de manutenção e evolução.

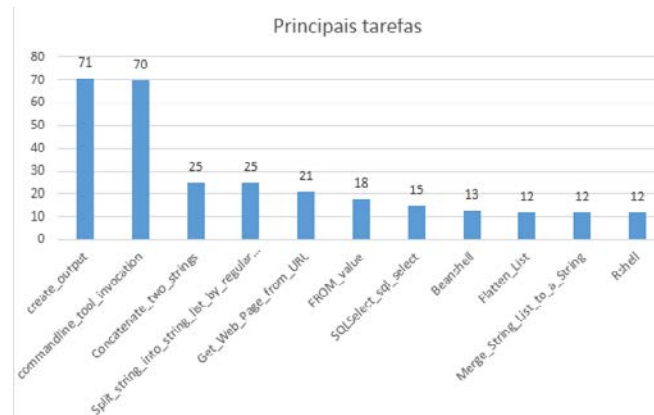


Figura 4. Lista das tarefas mais utilizadas nos *workflows*.

#### 4. Conclusão

Atualmente, o E-SECO ProVersion permite ao pesquisador capturar os dados do *workflow* por meio de um serviço Web disponível na ferramenta. Os dados coletados junto aos *workflows* são referentes às atividades executadas, suas portas de comunicação, valores de entrada e saída, tempo de execução e informações sobre falhas na execução, além das relações causais entre as atividades. Estes dados alimentam a ontologia PROV-OEXT, que por meio de regras específicas do domínio detectam informações sobre a evolução e manutenção em *workflows* e experimentos, sendo essas informações disponibilizadas ao pesquisador por meio de uma interface Web.

A disposição destas informações ao pesquisador, de forma automatizada, facilita a identificação da origem das falhas e conseqüentemente a busca pela solução. Também contribuem para formação de uma base de conhecimento sobre o experimento, fortalecendo o uso de laboratórios colaborativos e trazendo ganhos ainda maiores no que tange a gerência de configuração dos experimentos.

#### Agradecimentos

Os autores agradecem o apoio da Fapemig, CAPES e CNPq.

#### Referências

- Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludascher, B. and Mock, S. (2004). Kepler: an extensible system for design and execution of scientific workflows. In Scientific and Statistical Database Management. 16<sup>th</sup> International Conference on (p. 423-424). IEEE.
- Belloum, A., Inda, M. A., Vasunin, D., Korkhov, V., Zhao, Z., Rauwerda, H. and Hertzberger, L. O. (2011). Collaborative e-science experiments and scientific workflows. Internet Computing, IEEE, 15(4), p. 39-47.
- Bosch, J. (2009, August). From software product lines to software ecosystems. In Proceedings of the 13th international software product line conference (pp. 111-119). Carnegie Mellon University.
- Callahan, S. P., Freire, J., Santos, E., Scheidegger, C. E., Silva, C. T. and Vo, H. T. (2006). Managing the evolution of dataflows with VisTrails. In Data Engineering Workshops, 2006. Proceedings. 22<sup>nd</sup> International Conference on (pp. 71-71). IEEE.

- Cuevas-Vicentín, V., Kianmajd, P., Ludäscher, B., Missier, P., Chirigati, F., Wei, Y. and Dey, S. (2014). The PBase scientific workflow provenance repository. *International Journal of Digital Curation*, 9(2), 28-38.
- Deelman, E., Gannon, D., Shields, M. and Taylor, I. (2009). Workflows and e-Science: An overview of workflow system features and capabilities. *Future Generation Computer Systems*, 25(5), p. 528-540.
- Freitas, V.; David, J. M. N.; Braga, R. M. and Campos, Fernanda. (2015). An architecture for a Scientific Ecosystem. In: 9<sup>th</sup> Workshop on Distributed Software Development, Software Ecosystems and Systems-of-Systems (WDES 2015), Belo Horizonte, Brazil, SBC, p. 41-48.
- Gil, Y., Ratnakar, V., Deelman, E., Mehta, G. and Kim, J. (2007, July). Wings for pegasus: Creating large-scale scientific applications using semantic representations of computational workflows. In *Proceedings of the National Conference on Artificial Intelligence (Vol. 22, No. 2, p. 1767)*.
- Goble, C. A., Bhagat, J., Aleksejevs, S., Cruickshank, D., Michaelides, D., Newman, D. and De Roure, D. (2010). myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic acids research*, 38(suppl 2), p. 677-682.
- Hasan, R., Sion, R. and Winslett, M. (2007, October). Introducing secure provenance: problems and challenges. In *Proceedings of the 2007 ACM workshop on Storage security and survivability (p. 13-18)*. ACM.
- Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D. and Zhao, J. (2013). PROV-O: The PROV Ontology [www Document]. URL <http://www.w3.org/TR/prov-o/>
- Lim, C., Lu, S., Chebotko, A., & Fotouhi, F. (2010, July). Prospective and retrospective provenance collection in scientific workflow environments. In *Services Computing (SCC), 2010 IEEE International Conference on (pp. 449-456)*. IEEE.
- Mattoso, M., Werner, C., Travassos, G., Braganholo, V., Murta, L., Ogasawara, E. and Martinho, W. (2009). Desafios no apoio à composição de experimentos científicos em larga escala. *Seminário Integrado de Software e Hardware, SEMISH*, 9, 36.
- Miranda, G; De Souza, J. A.; Braganholo, V.; Oliveira, D. de. (2014). CollabCumulus: Uma Ferramenta de Apoio à Análise Colaborativa de Proveniência em Workflows Científicos. SBSC.
- Moreau, L. and Missier, P. (2013). PROV-DM: The PROV Data Model [www Document]. URL <http://www.w3.org/TR/prov-dm/>
- Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P. and Plale, B. (2011). The open provenance model core specification (v1.1). *Future generation computer systems*, 27(6), 743-756.
- Oinn, T., Li, P., Kell, D. B., Goble, C., Goderis, A., Greenwood, M. and Zhao, J. (2007). Taverna/myGrid: aligning a workflow system with the life sciences community. In *Workflows for e-Science (p. 300-319)*. Springer London.
- Silva, V., Chirigati, F., Maia, K., Ogasawara, E., Oliveira, D., Braganholo, V. and Mattoso, M. (2010). SimiFlow: Uma Arquitetura para Agrupamento de Workflows por Similaridade. *IV e-Science*, 1-8.
- Vaz, G. J., Giachetto, P. F., Torres, T. Z. and Massruhá, S. M. (2012). Um Modelo de Estrutura Organizacional em Plataformas de E-Science. In *Anais do Congresso da Sociedade Brasileira de Computação (Vol. 32)*.