

# RFlow: Uma Abordagem de Reutilização de *Workflows* Estatísticos Legados

José Antônio Pires do Nascimento<sup>1,2</sup>, Sérgio Manuel Serra da Cruz<sup>1</sup>

<sup>1</sup> Programa de Pós-Graduação em Matemática Computacional - PPGMMC/UFRRJ

<sup>2</sup> Embrapa Agrobiologia - Núcleo de Tecnologia da Informação (NTI)

joseantonio.pires@embrapa.br, serra@ufrrj.br

**Abstract.** *This paper presents an approach named RFlow, it aims to facilitate the management of statistical workflows and to mitigate limitations of statistical packages with regards to the management of provenance. RFlow is an approach that allows scientists to use legacy R scripts as meta-workflows enhancing their reuse, sharing, execution control with support of provenance metadata about the executions of such workflows.*

**Resumo.** *Este trabalho apresenta a abordagem RFlow que facilita o gerenciamento dos workflows estatísticos e mitiga limitações dos sistemas estatísticos no que diz respeito ao gerenciamento da proveniência. RFlow é uma solução que permite que os cientistas (re)utilizem scripts R legados sob a forma de meta-workflows facilitando o reuso, compartilhamento, controle de execução com apoio da coleta de proveniência sobre as execuções individualizadas de cada workflow.*

## 1. Introdução

O entrelaçamento entre o avanço do processamento de alto desempenho, da manipulação de crescentes volumes de dados científicos e a necessidade de gerência da proveniência em experimentos científicos *in silico* tem contribuído para o aumento sistemático do uso de *workflows* científicos [Mattoso *et al.* 2009]. A utilização de *workflows* científicos com foco no processamento estatístico também vem crescendo. Estes *workflows* se caracterizam pela manipulação de grandes volumes de dados e por executarem sofisticadas análises estatísticas através da incorporação de recursos (funções, algoritmos e métodos) disponíveis em sistemas estatísticos (SPSS, SAS, Statistica, Mapple, MathLab, Weka, R, entre outros [Mair e de Leeuw 2010]). Esses sistemas possuem características peculiares; podem ser de código aberto ou proprietários ou disponibilizarem recursos estatísticos e gráficos com distintos graus de sofisticação. No entanto, como característica comum, sua utilização requer sólidos conhecimentos em Estatística por parte dos usuários.

O sistema R é um dos mais difundidos na atualidade, sendo amplamente utilizado tanto nas áreas comerciais e científicas. Ele é capaz de executar desde comandos *online* até longos e sofisticados *scripts* desenvolvidos de modo *ad hoc* [Crawley 2002]. No entanto, apresenta uma limitação, ele ainda não dispõe de recursos de captura de proveniência sobre os processamentos e análises estatísticas realizadas pelos cientistas [Runnalls 2013].

O objetivo deste trabalho é apresentar uma abordagem denominada RFlow, que permite o gerenciamento dos *workflows* estatísticos e mitiga parte das limitações dos sistemas estatísticos no que diz respeito a ausência de gerenciamento da proveniência. RFlow encapsula *scripts* R legados sob a forma de meta-*workflows* que proporcionam o reuso, compartilhamento e controle de execução com apoio da coleta de proveniência sobre as

execuções individualizadas de *scripts* R legados. O conceito *meta-workflow* indica uma abstração. Ele se materializa sob a forma de um *workflow* concreto passível de ser executado por um sistema gerenciador de *workflows* científicos (SGWfC) [Kumar e Wainer 2005]. A abordagem RFlow auxilia aqueles cientistas que: (i) utilizam *scripts* R legados como parte dos seus experimentos *in silico*; (ii) possuem limitados conhecimentos sobre como desenvolver novos *workflows* científicos ou não dispõem de recursos nem tempo para refatorar seus *scripts* R e; (iii) necessitam dos descritores de proveniência para controlar seus experimentos *in silico*. A abordagem não requer recodificação ou alteração de códigos-fonte dos *scripts* R legados, ela oferece uma alternativa que possibilita a coleta de descritores de proveniência produzidos durante a execução de *scripts* R.

Este artigo está organizado da seguinte forma: a seção 2 descreve a fundamentação teórica (características e limitações do sistema R e apresentação dos *workflows* estatísticos). A seção 3 descreve a abordagem e a arquitetura RFlow. A seção 4 apresenta um protótipo do RFlow, um *meta-workflow* concebido para ser utilizado no sistema Kepler [Ludäscher et al, 2006] e um *schema* de dados OPM-compatível utilizado pelo mecanismo de coleta de proveniência. A seção 5 apresenta um primeiro estudo de caso na área de Agrobiologia e discute os resultados iniciais. A seção 6 contextualiza a abordagem em relação aos trabalhos relacionados. A seção 7 conclui este artigo e apresenta os trabalhos futuros.

## 2. Fundamentação Teórica

Inúmeros ramos das Ciências Naturais manipulam crescentes volumes de dados. Neste sentido, diversas áreas necessitam utilizar de modo integrado sistemas estatísticos e *workflows* científicos para se beneficiarem da gestão da proveniência dos experimentos científicos.

### 2.1 Sistema Estatístico R

O sistema R é um ambiente interativo de programação maduro e largamente utilizado, que permite a codificação de *scripts* capazes de executar sofisticados processamentos estatísticos [Crawley 2002 e Chambers 2008]. R possui código aberto e oferece uma ampla gama de recursos estatísticos (funções e métodos para executar modelagem linear e não linear, análises multivariadas, análises de séries temporais, entre outros), além de uma biblioteca de funções gráficas. O sistema é facilmente conectado a diversos tipos de sistemas gerenciadores de bancos de dados (SGBD), entretanto, apesar do potencial, o sistema ainda não possui facilidades que apoiem a coleta de proveniência [Runnalls 2013].

R pode ser associado com diversos SGWfC. Por exemplo, a associação entre o R e o Taverna dá-se pelo componente *RShell*. No Kepler, ela ocorre mediante a construção de *workflows* que incorporam atores e diretores R-específicos (por exemplo, *ReadTable*, *RandomNormal*, *ANOVA*, *Correlation*, *LinearModel*, *RMean*, *Rmedian*, *Rquantile*, *Summary*, *Barplot*, *Boxplot*, *RExpression*, *Scatterplot*)[Ludäscher et al 2006]. No VisTrails não estão disponíveis módulos R-específicos, a associação se dá através de codificação direta de módulos em *workflows* que invocam os recursos estatísticos do sistema R.

### 2.2. Workflows Estatísticos

Segundo Kumar e Wainer (2005) e Ranabahu *et al.* (2011), os *workflows* estatísticos são utilizados para analisar grandes volumes de dados. Os *workflows* estatísticos são codificados através de linguagens próprias dos sistemas estatísticos sob a forma de complexos *scripts*. Os *workflows* estatísticos são geralmente desenvolvidos por cientistas que possuem grande experiência em estatística e limitada capacidade de programação. Por esse motivo, possuem

um alto custo de desenvolvimento, são domínio-específicos e difíceis de serem reutilizados e compartilhados por terceiros [Kirchkamp 2011]. Atualmente não estão disponíveis na literatura trabalhos que relacionam proveniência, SGWfC e sistemas estatísticos através de *meta-workflows* reutilizáveis que encapsulam *scripts* legados sem a necessidade de alteração em seus códigos-fonte. Na próxima seção, apresentamos uma contribuição denominada abordagem RFlow para ajudar a sanar esta ausência.

### 3. A Abordagem RFlow

A abordagem RFlow é caracterizada, segundo os critérios classificatórios da taxonomia proposta por Cruz et al (2009), como uma arquitetura de captura semi-automática de proveniência que utiliza mecanismos externos que atuam no nível do *workflow* por meio de técnicas de anotação. Neste caso, os descritores de proveniência são capturados pelos sistemas de proveniência acoplados ao SGWfC que controlam a execução do *workflow* estatístico. O benefício de trabalhar neste nível reside no fato de que os *script* R são preservados, não necessitam ser adaptados ou (re)codificados para se beneficiarem de suporte do mecanismo de captura de proveniência do SGWfC.

A proveniência capturada é de natureza prospectiva e retrospectiva, de baixa granulosidade e orientada a dados. As vantagens de operar nesse nível são inúmeras. É possível coletar descritores detalhados sobre as execuções individualizadas de cada instância do *workflow*. Os descritores são armazenados de modo centralizado em um repositório que está desacoplado do sistema de coleta. Estas características facilitam a substituição de módulos da arquitetura.

#### 3.1. Arquitetura RFlow

A Figura 1 apresenta uma representação conceitual da arquitetura, das suas camadas e componentes da RFlow. Através da interface o cientista selecionará o *script* legado na camada de sistema estatístico, este será posteriormente encapsulado na camada do SGWfC por um *meta-workflow* genérico que passará a se beneficiar da coleta de proveniência associado ao SGWfC. Essa etapa é controlada pelos serviços de configuração do RFlow que acessam o *script*, fazem a associação ao *meta-workflow* correspondente e associam o mecanismo de captura de proveniência. Durante a fase de configuração, os descritores de proveniência produzidos pelo usuário, parâmetros, arquivos de entrada e saída, descritores do experimento (proveniência prospectiva) são capturadas pelos módulos publicadores de proveniência e posteriormente armazenados em um repositório de proveniência na camada de persistência representada pelo SGBD.

Após as configurações iniciais, o cientista executa o *workflow*. Os descritores de proveniência retrospectiva relacionados com a execução do *script* R são coletadas pelo sistema de proveniência acoplado ao SGWfC e transferidos para o repositório. Ao término da execução do experimento, o cientista possuirá um repositório integrado de proveniências que poderá ser analisado *à posteriori*.

Ao considerarmos todo o processo, ao menos duas questões relacionadas aos *meta-workflows* merecem ser discutidas. A primeira diz respeito à necessidade da escolha condicionada do *meta-workflow*. Ele deve ser previamente codificado de acordo com o SGWfC adotado pelo cientista, uma vez que cada SGWfC possui características operacionais peculiares. A outra questão é a decisão de utilizar (ou não) atividades R-específicas nos *meta-workflows*. Mais uma vez, a escolha depende do SGWfC adotado. Alguns SGWfC oferecem

suporte nativo ao R diretamente na especificação das atividades concretas (Kepler), outros trabalham apenas com suporte a módulos genéricos, serviços Web ou executáveis.

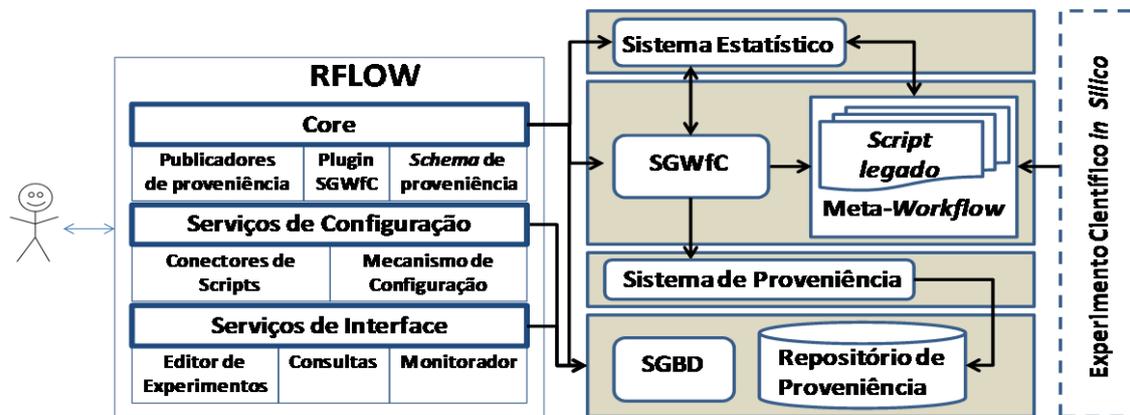


Figura 1 – Arquitetura RFlow, camadas e componentes.

### 3.2. Meta-Workflows.

Neste trabalho, o *meta-workflow* representa uma solução genérica usada para encapsular *scripts* R legados. Ele atua como um *wrapper* que encapsula as funções estatísticas do *script* sob a forma de uma, ou mais, atividades de *workflow*, permitindo sua execução em um SGWfC com todos os benefícios inerente ao sistema além da captura de proveniência. Os *meta-workflows* permitem a (re)utilização dos *scripts* R legados sob a forma de *workflows* científicos sem que seja necessário refatorar ou recodificá-los. A execução do *meta-workflow* é de responsabilidade do SGWfC, assim, se tira proveito de facilidades que estão ausentes no sistema R, tais como controle de execução, rastreabilidade, reprodutibilidade e primordialmente a coleta de proveniência sobre cada execução.

### 3.3. Módulos da RFlow

A Figura 1 apresenta a arquitetura do RFlow de acordo com o estilo camadas [Qin *et al.* 2008]. A arquitetura é definida em três camadas: (i) *Core* - contém os componentes básicos para o funcionamento da arquitetura. O *Core* permite a inserção *plugins* que são responsáveis pela configuração da tríade *meta-workflow*, SGWfC e repositório de proveniência. O componente *publicador de proveniência prospectiva* representa os serviços relacionados à publicação de dados de proveniência prospectiva. Analogamente, o componente *publicador de proveniência retrospectiva* representa os serviços relacionados à publicação de proveniência retrospectiva. Estes componentes são necessários devido às naturezas distintas da proveniência que operam no nível de *workflow* e que são coletadas inicialmente durante a configuração e posteriormente durante a execução dos *workflows*. O componente *schema de proveniência*, define o modelo de dados compatível com especificações de proveniências (OPM, PROV, etc), que será utilizada pela abordagem para armazenar os descritores de proveniência. (ii) *Serviços de configuração* – contém os componentes que são utilizados para viabilizar a configuração do RFlow. O componente *mecanismo de configuração* é responsável por definir qual o SGWfC e o sistema de coleta de proveniência que serão utilizados no experimento. Os componentes *conectores de scripts* são utilizados para facilitar o processo de carga dos *scripts*, e conexão dos sistemas estatísticos e os *meta-workflows*. (iii) *Serviços de Interface* – oferecem ao cientista uma interface Web. Utiliza componentes de *editor de experimentos*, permitindo que cientistas cadastrem seus *scripts* R, *meta-workflows* e também descrições dos experimentos. Os serviços de interface, opcionalmente, poderão incluir

atividades de monitoramento e de consulta de proveniência. As *atividades de monitoramento* permitem que os cientistas acompanhem a execução dos seus experimentos. As *atividades de consulta* realizam consultas sobre os descritores de proveniência capturados ao longo da configuração e execução do experimento. Além destes componentes, outros podem ser acrescentados na interface de acordo com a necessidade.

#### 4. O Protótipo da Arquitetura RFlow e seus módulos principais

O protótipo RFlow, em desenvolvimento na forma de aplicação Web, utiliza a tecnologia Java EE. O estilo arquitetural utilizado foi o MVC, pois permite melhor separação entre a lógica de negócio (camadas *Core* e de *serviços de configuração*) e a lógica de apresentação (*camada de interface*). Além disso, foi utilizado o *framework* JSF para desenvolver a aplicação segundo a tecnologia Java EE e o servidor de aplicação Apache TomCat para executá-la. Com relação à camada de persistência, o protótipo persiste os descritores de proveniência prospectiva e retrospectiva em um banco de dados MySQL EE (versão 5.5.27). Como SGWfC adotou-se o Kepler (versão 2.3) devido a sua estabilidade e principalmente a facilidade de integração com o sistema R através da utilização de atores genéricos. Adicionalmente, o Kepler permite acoplar soluções de coleta de proveniência (COMAD, Provenance Recorder ou RWS). O sistema estatístico utilizado foi o R (versão 2.15.3).

##### 4.1. Meta-Workflow ExecScript

O meta-workflow ExecScript (Figura 2) é um exemplo simplificado da implementação do conceito no SGWfC Kepler. Ele é composto por atores genéricos do tipo “External Execution”, que são os responsáveis por se comunicarem tanto com os componentes da interface do RFlow (ChoiceScript.jar) quanto com os demais atores (“Constant” e “File Reader”). Os atores tipo “External Execution” encapsulam e permitem o reuso dos *scripts* R legados sem que seja necessário alterar seus códigos-fonte.

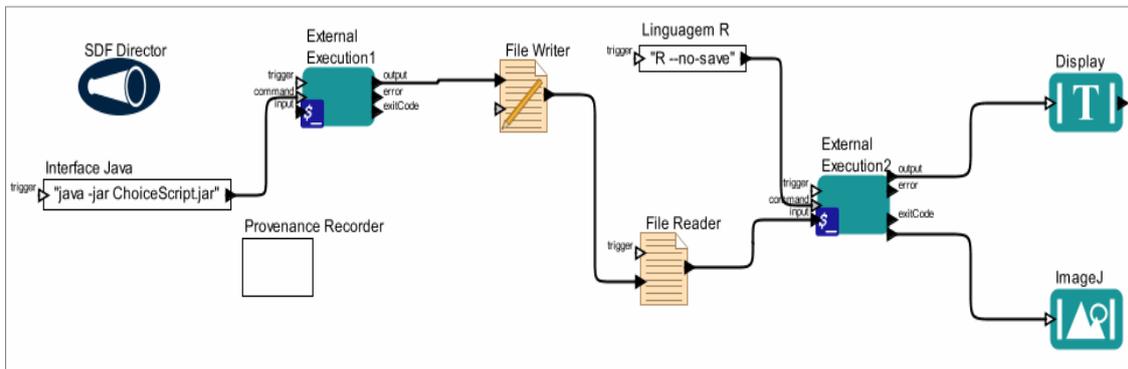


Figura 2 – Meta-workflow ExecScript codificado para o SGWfC Kepler.

##### 4.2. Modelo de proveniência da RFlow

O coletor de proveniência utilizado neste protótipo foi o *Provenance Recorder* [Altintas, et al, 2006]. Sua escolha se deve aos seguintes motivos: (i) mecanismo de simples utilização (representado pelo ator “Provenance Recorder” adicionado ao meta-workflow); (ii) capacidade de registrar a proveniência dos dados de entrada, saída e intermediários, além das definições do *workflow* (diretores, atores, sub-workflows, portas, tokens, *script* R utilizado) e informações de *timestamp* relacionadas a sua execução. (iii) capacidade para armazenar proveniência de baixa granulosidade e oferecer facilidades de conectividade tanto com

arquivos XML ou Oracle, MySQL, HSQL); (iv) por fim, o sistema é compatível com a especificação OPM<sup>1</sup>.

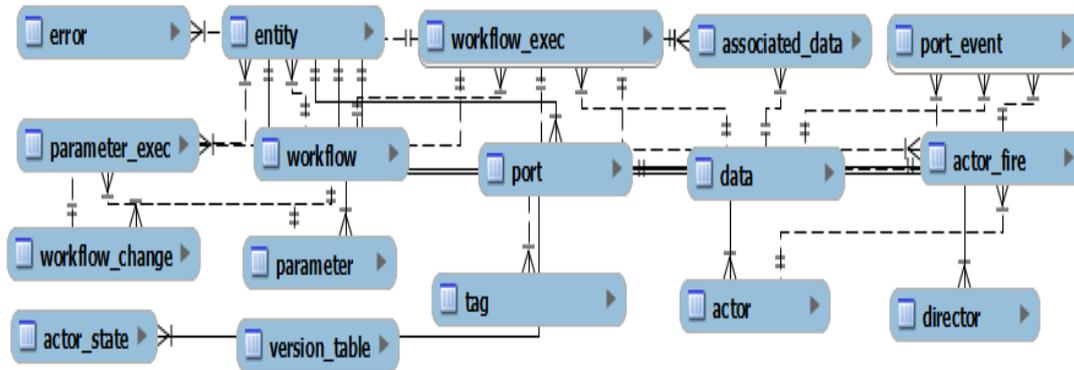


Figura 3 – Modelo de dados da RFlow, adaptado de [Altintas, et al, 2006].

O armazenamento da proveniência é local e centralizado (Figura 3). O modelo é OPM-compatível e suas relações mapeiam os principais componentes dos *workflows* estatísticos (atores, diretores, portas, parâmetros e o contexto de execução). Os descritores de proveniência prospectiva e retrospectiva podem ser consultados através de simples consultas SQL através dos serviços de interface. O esquema representa três tipos de informação: o conteúdo ou especificação de *workflows*, como essas especificações mudam ao longo do tempo, e os eventos que ocorrem durante a execução do *workflow*. A partir da captura de proveniência de dados e eventos que é possível gerar a re-execução dos *workflows* executados. Nas tabelas *workflow\_change*, *workflow* e *workflow\_exec* que são armazenados os descritores de proveniência que indicam a proveniência prospectiva gerada quando há alteração de parâmetros, portas, componentes, dados de arquivo, entre outros durante a fase de configuração.

## 5. Estudo de Caso - *Workflows* estatísticos em Agrobiologia

O estudo de caso adotado é da área de Agrobiologia<sup>2</sup> e foi avaliado durante Novembro de 2012 até Março de 2013. Como etapa preliminar, investigou-se um conjunto de *workflows* estatísticos do sistema R (n=50). A investigação contou com entrevistas técnicas com cientistas e desenvolvedores, além de inspeção dos códigos-fonte dos *scripts* R legados.

O *script* R encapsulado pelo *ExecScript* diz respeito à análise de fertilidade do solo. Essas análises são de importância econômica e ecológica, pois é a partir das análises que serão feitas as recomendações técnicas da adubação/calagem nas áreas a serem cultivadas. O *script* utiliza massas de dados sob a forma de alto número de arquivos de entrada (o número de registros varia em função do tamanho da área avaliada) e emprega uma ampla variedade de funções estatísticas e gráficas do R (regressão linear, análise multivariada, variância, amostragem, entre outras). O meta-*workflow* produziu resultados textuais e gráficos do *script* R legado. Os dados utilizados no estudo de caso são de natureza sintética e representam valores residuais de amostras de material orgânico coletados em várias amostras de solos

<sup>1</sup> Este trabalho adotou a especificação OPM em vez da PROV ([www.w3.org/TR/prov-dm/](http://www.w3.org/TR/prov-dm/)) devido às limitações técnicas impostas pelo Provenance Recorder.

<sup>2</sup> Até o momento da publicação deste trabalho, os *scripts* R e dados reais dos experimentos em Agrobiologia se encontravam em processo revisão/publicação, por esse motivo não foram expostos nesta seção.

dispersas por todo o estado do Rio de Janeiro durante um período de 12 meses. Os arquivos de dados de entrada são unicamente identificados e formatados em multi-colunas que representam as variáveis numéricas dos materiais orgânicos e o tipo de tratamento aplicado. O experimento foi executado 30 vezes (com variações nas quantidades e dados dos arquivos de entrada e parâmetros de configuração) em um equipamento com 8 GB de RAM e processador Intel Core i3.

Graças aos descritores de proveniência coletados durante as (re)execuções, elaborou-se um pequeno conjunto de questões de proveniência que até então impossíveis de serem respondidas caso os *scripts* fossem executados somente no sistema R. Neste protótipo as questões de proveniência são materializadas sob a forma de consultas SQL que operam sobre o modelo de proveniência descrito na Figura 3. Apesar de ser um protótipo e ainda estar em fase de desenvolvimento RFlow é capaz de responder as seguintes questões: (i) Quais os nomes dos *scripts* executados pelo pesquisador *PI*? (ii) Quais foram os dados de entrada, intermediários e finais produzidos pela execução *workflow* de nome *Wf1* que encapsula o script *Sc1* que ocorreu no dia 15/03/2013? (iii) Quais as execuções do script *Sc1* foram finalizadas com mensagens de erro do tipo *file not found*? (iv) Quem executou o script *Sc1* e qual foi a sua duração?, entre outras.

## 6. Trabalhos Relacionados

Atualmente existem poucos trabalhos que relacionam proveniência, *workflows* científicos e sistemas estatísticos através do uso de meta-*workflows* reutilizáveis que preservam os *scripts* legados sem a necessidade de alteração em seus códigos-fonte. Pelo contrário, as alternativas apontam soluções na direção oposta [Higgins 2007]. Por exemplo, o Kepler oferece um conjunto de atores R-específicos que precisam ser explicitamente modelados sob a forma de atividades concretas no *workflow* para que invoquem os recursos do sistema R. Essa abordagem não é tecnologicamente neutra, exige: razoáveis esforços de programação por partes dos cientistas e abandono ou substituição dos *scripts*.

Uma alternativa que vem ganhando corpo nos últimos anos é a incorporação de recursos de proveniência aos sistemas estatísticos. Silles e Runnalls (2010) e Runnalls (2013) propõe a refatoração do código do sistema R para que incorpore recursos de proveniência no seu motor de execução. Eles apresentaram uma variante do R denominada CXXR. O sistema já oferece algum tipo de coleta de proveniência retrospectiva sob a forma de *logs* de execução. No entanto, ainda está em desenvolvimento, opera desacoplado de SGWfC e permanece compatível com a OPM.. Além disso, ainda não oferece soluções de conectividade via SGBD para repositórios de proveniência. No que se refere ao encapsulamento, Marinho *et al.* (2012) propuseram uma abordagem de gerência de proveniência que captura, armazena a análise, de modo integrado, os descritores de proveniência de *workflows* executados em ambientes heterogêneos e distribuídos. Os autores não consideram a presença de meta-*workflows* nem apresentam comentários sobre a utilização/integração de sistemas estatísticos.

## 7. Conclusão e Trabalhos Futuros

Este trabalho apresentou uma abordagem capaz de reutilizar *workflows* estatísticos previamente modelados sob a forma de *scripts* e ao mesmo tempo tirar proveito da associação entre sistemas estatísticos e SGWfC. A abordagem oferece abstração na definição de experimentos e novas facilidades na condução destes experimentos graças aos mecanismos de coleta de proveniência até então indisponível no *Sistema R*. As contribuições deste trabalho são a definição de uma abordagem denominada RFlow e seu protótipo, um meta-workflow

genérico para o SGWfC Kepler capaz de encapsular *scripts* R legados e um *schema* de dados OPM-compatível cujo propósito é armazenar os diferentes tipos de proveniência. Como trabalho futuro se amadurecerá a arquitetura, dando continuidade ao desenvolvimento dos módulos, em especial aos *plugins* do *Core* e atualização do *schema*, tornando-a compatível com a especificação PROV recentemente homologada pelo W3C.

## Agradecimentos

À Embrapa Agrobiologia e à FAPERJ pela aprovação do APQ1 e apoio financeiro (E-26/110.840/2012 e E-26/112.588/2012).

## Referências

- Altintas, I. et al (2006), “Provenance Collection Support in the Kepler Scientific Workflow System”, Proc. of IPAW2006, 118-132.
- Chambers, J. R (2008) Software Data Analysis Programming with R Software. Springer. 1st edition.
- Crawley, M.J. (2002) Statistical Computing to Data Analysis using S-plus. Wiley. 1st edition.
- Cruz, S. M. S. , Campos, M L M., Mattoso (2009) “Towards a Taxonomy of Provenance in Scientific Workflow Management Systems”,In: Services, .2009. pp. 259 – 266.
- Higgins, D. (2007) ,“Using R in Kepler”, [ptolemy.eecs.berkeley.edu/conferences/05/presentations/higginsRSystem.pdf](http://ptolemy.eecs.berkeley.edu/conferences/05/presentations/higginsRSystem.pdf).
- Kirchkamp, O. (2011) “Workflow of statistical data analysis”. <http://www.kirchkamp.de/oekonometrie/pdf/wf-screen2.pdf>.
- Kumar, A., Wainer, J. (2005) “Meta-workflows as a control and coordination mechanism for exception handling in workflow systems”. Decision Support Systems. v. 40 pp. 89-105.
- Ludäscher, B. et al. (2006) "Scientific workflow management and the Kepler system: Research Articles". Concurrency and Computation: Practice & Experience, v. 18, n. 10, pp. 1039-1065.
- Mair, P., de Leeuw, J. (2010). “A general framework for multivariate analysis with optimal scaling: The R package aspect”. Journal of Statistical Software, 32(9), pp. 1-12.
- Marinho et al (2012) “ProvManager: a provenance management system for scientific workflows” Conc. and Comp.: Practice & Experience. v. 24, n. 13, pp. 1513-1530.
- Mattoso, M. et al., (2009), "Desafios no apoio à composição de experimentos científicos em larga escala". In: Seminário Integrado de Software e Hardware (XXXVI SEMISH), pp. 307-321.
- Runnalls, A. (2013) “CXXR: an extensible R interpreter “In: Wiley Interdisciplinary Reviews: Computational Statistics. DOI: 10.1002/wics.1251.
- Qin, Z., Xing, J., Zheng, X., (2008), Software architecture. Springer. 1st edition.
- Ranabahu, A., Anderson, P. Sheth, A. P. (2011) “The Cloud Agnostic e-Science Analysis Platform”. IEEE Internet Computing v. 15.pp. 85-89.
- Silles, C. A., Runnalls, A. (2010) “Provenance-Awareness in R”. LNCS 6378. Springer, pp. 64-72.