7th International Digital Curation Conference
December 2011

Trends in Use of Scientific Workflows: Insights from a Public Repository and Recommendations for Best Practices

Richard Littauer,
DataONE Project,
University of New Mexico

Karthik Ram,
Environmental Science, Policy, and Management,
University of California, Berkeley

Bertram Ludäscher,
Dept. of Computer Science & Genome Center,
University of California, Davis

William Michener,
Professor and Director e-Research Program,
University Libraries, University of New Mexico

Rebecca Koskela,
Executive Director of DataONE,
University of New Mexico

December 2011

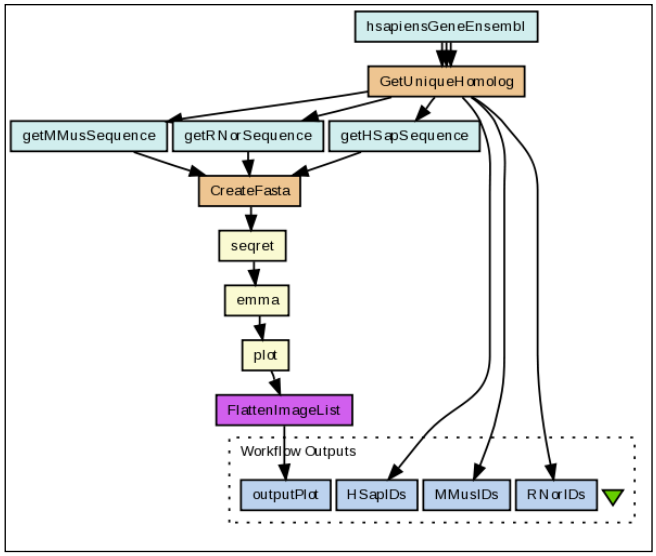
Abstract

Scientific workflows are typically used to automate the processing, analysis, and management of scientific data. Most scientific workflow programs provide a user-friendly graphical user interface that enables scientists to more easily create and visualize complex workflows that may be comprised of dozens of processing and analytical steps. Furthermore, many workflows provide mechanisms for tracing provenance and methodologies that foster reproducible science. Despite their potential for enabling science, few studies have examined how the process of creating, executing, and sharing workflows can be improved. In order to promote open discourse and access to scientific methods as well as data, we analyzed a wide variety of workflow systems and publicly-available workflows on the public repository myExperiment. It is hoped that understanding the usage of workflows and developing a set of recommended best practices will lead to increased contribution of workflows to the public domain.

Background

Scientists need to process, analyze, visualize, and manage increasing amounts of scientific data. For complex analyses, scientists often must combine multiple processing steps into larger "analysis pipelines" that can involve a number of custom algorithms, specialized tools (e.g., statistics packages, geographic information systems), local and remote databases, and web services. Such *scientific workflows* (Figures 1,2) are typically executed repeatedly, with different combinations of inputs and parameters. Most current research efforts lack any form of workflow automation or partially accomplish such goals via a loose collection of programming scripts that link various computational steps.

In recent years, scientific workflow systems such as Kepler[1] (Altintas et al., 2004), Taverna[2] (Hull et al., 2006; Oinn et al., 2006), VisTrails[3] (Callahan et al., 2006), and many others have emerged as a promising technology to further simplify the creation, execution, and sharing of computational workflows among scientists and laboratories. Scientific workflow systems often support visual workflow design, execution monitoring, fault-tolerance and recovery, and the use of distributed and parallel computing resources. Perhaps the most powerful feature of state-of-the-art workflow systems is the ability to record data lineage and other provenance information during execution, thus allowing scientists to "replay" processing steps, study data dependencies, and the datasets and parameters specified (or not used!) during any workflow run. Thus, scientific workflow systems may also foster more transparent and reproducible science. In several cases, workflows have already been either used in diagnosis of inconsistent methodology (Coombes et al., 2007), or in eliminating years of software development for researchers (Fisher et al., 2007).

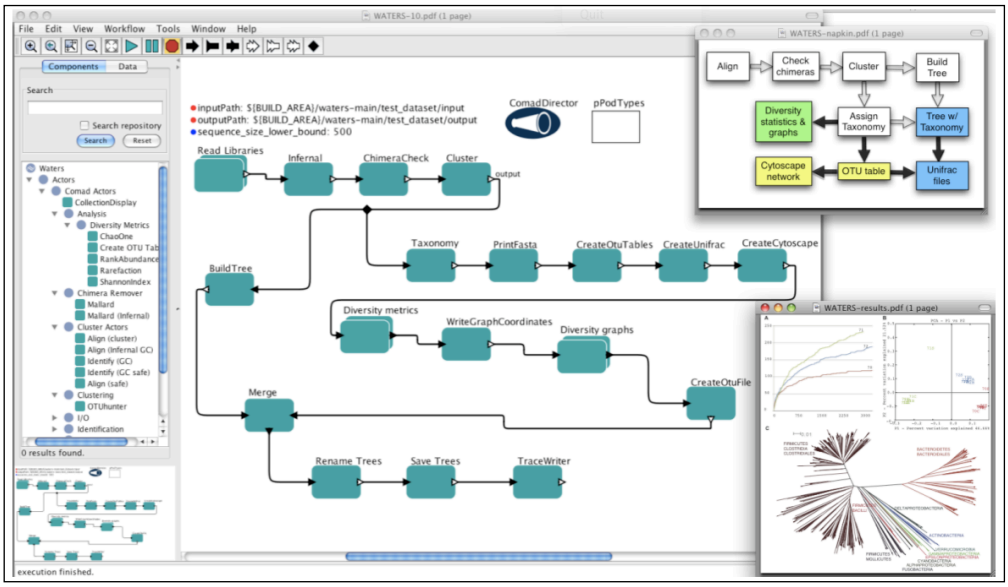


[Caption] Figure 1 Taverna workflow `BiomartAndEMBOSSAnalysis` in myExperiment for retreiving sequences from different species, aligning them, and plotting the result. © 2009 Alan Williams.[4]

Scientific workflows are usually represented as directed graphs, whose nodes

[1] Kepler Project. http://www.kepler-project.org
[2] Taverna. http://www.taverna.org.uk
[3] Vistrails. http://www.vistrails.org
[4] BiomartAndEMBOSSAnalysis. http://www.myexperiment.org/workflows/158.html

represent workflow steps that are linked via dataflow edges, thus prescribing serial or parallel execution of nodes. Fig.1, e.g., shows a Taverna workflow; Fig.2 depicts a Kepler metagenomics workflow (Hartman et al., 2010).



[Caption] Figure 2 Kepler workflow for the Alignment, Taxonomy, and Ecology of Ribosomal Sequences (Hartman et al., 2010).

Workflow Repository Study

Increasingly, scientists are sharing not only their data, but also software tools and scientific workflows. myExperiment.org is a public workflow repository currently containing over 2,000 workflows with more than 5,000 registered users[5] (Goble et al., 2010). myExperiment allows one to discover workflows of interest, which can then be used or adapted for specific requirements. myExperiment also provides social tools that foster development of virtual communities around topics of interest. A feedback and attribution mechanism allows workflow developers to gain credibility and recognition from their peers.

In order to understand patterns of workflow use, we studied various attributes of workflows that were deposited in myExperiment since it became available in 2007. Specifically, we sought to identify characteristics of workflows that were related to their degree of use (e.g., number of downloads). Based on our findings, we recommend a set of best practices that may increase the usability and amount of re-use of workflows.

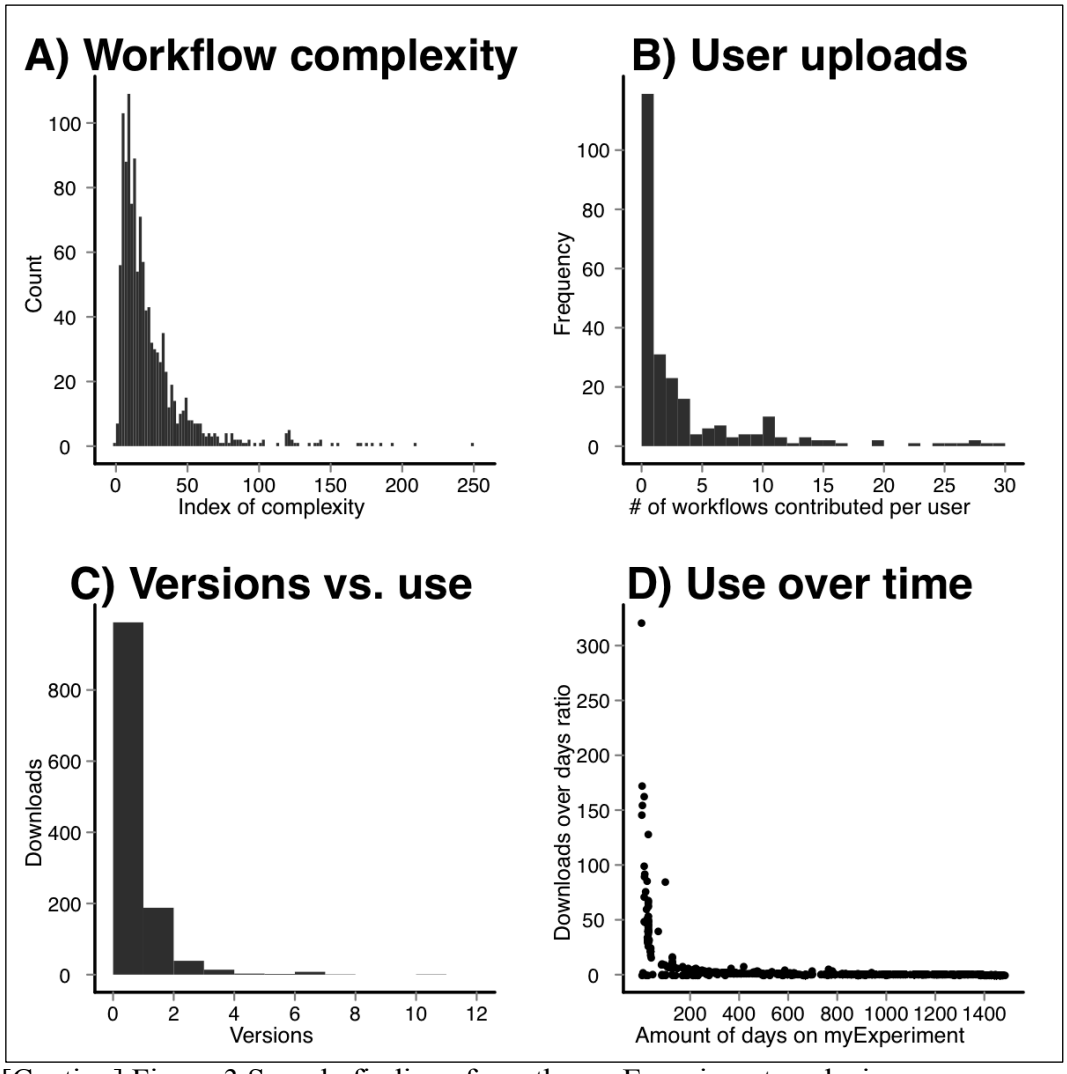
The study focused on publicly available information gathered from myExperiment both through the SPARQL endpoint that accessed the central RDF triplestore in the myExperiment ontology[6] and through the HTML source code on the public site. The information on the HTML pages for workflows, packs, files, groups, statistics, and users was harvested using Python; all of the code used in this research has been publicly uploaded to GitHub.[7]

This harvesting, along with calibration and development of the scripts, unfortunately resulted in some of the data being affected by the research: amount of views and downloads for each workflow were marginally affected. Noise was already

[5] myExperiment. http://www.myexperiment.org
[6] myExperiment ontology. http://rdf.myexperiment.org/
[7] DataONE Workflow Project GitHub. https://github.com/RichardLitt/Understanding-Workflows

present to some extent in the myExperiment system; for instance, the first workflow that appears by ranking (the default setting) when accessing the workflow repository on myExperiment has been viewed around 10,000 times and downloaded 4,000 times. The disparity between this high amount of usage and the average use of other workflows on myExperiment can be accounted for by it's prominent location on myExperiment where all beginner users would likely view it as a first step when exploring the site.

myExperiment can store scientific workflows from several different workflow plartforms, including Kepler, Taverna, VisTrails, Knime, and Rapid Miner. Of the workflows on the site at the time of this analysis, Rapid Miner had 153 workflows, Taverna 1 had 479 workflows, and Taverna 2 had 684 workflows. Other workflow platforms represent a combined total of 281 workflows. As Taverna workflows utilize the XML language SCUFL (Simplified Conceptual Uniform Flow Language), myExperiment stores information about Taverna workflows internal properties, whereas the other workflows are generally stored as inscrutable objects. While Rapid Miner also has some internal property information stored on the site, this study focused on Taverna workflows when studying workflow-internal metrics. This practice is not unprecedented; Wassink et al. (2009) did a similar study on the 415 Taverna workflows then on the site.



[Caption] Figure 3 Sample findings from the myExperiment analysis.

Findings

Figure 3 shows some of the findings from our initial study of the myExperiment repository, based on data publicly available from the website. The numbers of views and downloads were used as a measure of workflow popularity and as a proxy for workflow use.

The distribution of workflow *complexity* is shown in Fig. 3A. It indicates that a large percentage of workflows consist of few components. The complexity of a workflow was determined by aggregating the number of beanshells, processors, inputs, coordinations, datalinks, and operators. The number of tasks included in Taverna workflows ranged from 1 to 250. The average workflow supports 24.3 tasks; with a standard deviation of 26.6 tasks. The total number of tasks in all Taverna workflows together is 28411 (not including 76 workflows for which this information was unavailable). This is a large increase on the Wassink et al. (2009) study, which found that: "The number of tasks per workflow ranges from 1 to 70 tasks. The average workflow size is 8.8 tasks; the standard deviation is 11.7 tasks. The total number of tasks in all workflows together is 3660."

Many workflows have other workflows embedded within them, indicating that workflows may be shared and and reused. The total number of embedded workflows is 694; the average number of embedded workflows is 0.6; the standard deviation is 1.7 embedded workflows. In addition, many workflows access web services; the average amount of web service tasks (e.g., WSDL, Biomart, SOAP, or XML services) in a workflow is 2; the standard deviation is 3.5. The total amount is 2,368 such tasks.

A large percentage of workflow components are *shims* (components that are used to make output from one step conform to the format expected by a subsequent step), indicating the value of workflows to "glue" pre-existing components together. Shims are hard to define clearly using the SCUFL information on myExperiment, as many different sorts of components may be used for data conversaion. Using a loose aggregate of Beanshells and Processors that do not access online databases (which may themselves perform shimming processes), the average amount of shim components per workflow is 6.3, with a standard deviation of 9.3. The total aggregate comes to 7,405 components; indicating that approximately 38% of workflow components are shims. The shimming problem has received some attention in the literature (Cui et al., 2009); this figure suggests that it remains a significant problem for workflow developers.

Most workflow contributors submit a *single* workflow, whereas there are a smaller number of developers that contribute many workflows to myExperiment. (Fig. 3B) Only 13 users have uploaded more than 30 workflows; of these, only two uploaded more than 100, with one user having uploaded 145 workflows, and the other 255. (These users have not been included in the 3B graph, for reasons of scale.) Just over 5% of the users on myExperiment have uploaded workflows; only 346 users in total.

Complex workflows that perform many tasks are downloaded more frequently than simpler workflows. Mature workflows (i.e., where several versions have been released) are more frequently downloaded than "single-edition" workflows, similar to the trend one sees for published textbooks. (Fig. 3C)

Workflow downloads occur more frequently for registered myExperiment members (as opposed to anonymous users), and they occur via external applications (e.g. Taverna) that allow one to execute those workflows. Workflow use also declined significantly thirty days after initial upload. (Fig. 3D)

Use of workflows (e.g., numbers of views and downloads) does not seem to be related to the volume of documentation associated with the workflow nor the number

of tags (e.g., searchable keywords) assigned to the workflow, but is related to the degree of community engagement with the workflow as exhibited by number of citations, comments, ratings, and reviews. Similarly, the frequency with which a workflow is viewed is correlated with the number of downloads by users.

Recommended Best Practices

Based on our study of the myExperiment workflows and usage patterns, we provide a number of suggestions for workflow developers and users of workflow repositories, that may increase the usability and amount of re-use of workflows. Note that our suggestions focus on aspects that can be derived from automatically harvesting information from a workflow repository, such as myExperiment and the workflows therein. In particular we did not assess general software engineering and usuability aspects of workflows directly, although these play a major role in workflow adoption and re-use.

1. It is important to consider workflows as evolving entities that are updated in response to user feedback, engagement, and improvements in methodology. Results suggest that frequently updated workflows receive greater use than ones that are shared, but never revised or improved through subsequent versions.
2. Social annotation tools such as user contributed tags can play an important role in making workflows accessible. However, we found that workflows annotated with superfluous tags are not necessarily accessed more frequently; a discrepancy which may be due to the nature of freely chosen tags. The benefits of tagging may be increased by normalization, e.g., through a controlled vocabulary of workflow tags.
3. Workflow reuse may be significantly increased by fostering greater community awareness. This may be accomplished by citing the workflow in publications, sharing the workflow with colleagues working on similar scientific problems, and "advertising" the workflow through various social media to relevant communities.
4. Although our initial study revealed no strong relationship between workflow use and the amount of associated documentation or metadata, we recommend that workflow developers provide sufficient descriptions of their workflows so that potential users may more readily discover and understand workflows that can potentially meet their needs, especially if it is clear that a workflow saves time and increases productivity.
5. Workflow re-use could benefit significantly from the assignment of stable identifiers, e.g. Digital Object Identifiers (DOI) or similar persistent identifiers. Without stable identifiers, workflow URLs are impermanent and prone to loss over time. On the other hand, stable identifiers, in particular versioned ones, can prolong workflow longevity long after publication.
6. One size does not fit all. It is important for developers and repositories to create and support libraries of smaller workflows (i.e., components) that can do simple tasks and be integrated into more complex workflows with greater functionality. The more complex workflows, on the other hand, can serve as show-cases for how to solve larger "end-to-end" problems, encouraging new users to build their own solutions using component libraries and ideas from other end-to-end workflows.

7. Increased usage of workflows and workflow repositories will likely be related to the degree that education and outreach opportunities are provided to scientists through professional society meetings, online courses, and incorporation into academic training (e.g. undergraduate and graduate courses). We anticipate that communities of practice will emerge through such efforts, greatly enhancing scientific productivity and supporting reproducibility of scientific results.

An interesting subject for future studies is the question of how communities of practice form and what the main technical drivers are. For example, the availability of workflow repositories such as myExperiment has led to increased awareness and use of scientific workflows. Scientific workflow system such as Kepler and Taverna and repositories such as myExperiment are generic with respect to their application domain. On the other hand, domain-centric systems such as Galaxy (Goecks et al., 2010) for the biomedical community, are widely used by a specific community due to the presence of a rich set of commonly used, interoperable software components for the target users.

Potential Impact on Science

Following good practices in using workflow tools like Taverna and Kepler and deposition of workflows in public domain repositories like myExperiment.org can make science much more efficient by allowing scientists to re-use previously created and peer-reviewed workflows and can promote reproducible science (Reichmann, Jones, & Schildhauer, 2011). In addition, the United States National Science Foundation considers patents, copyrights and software systems (such as workflows) to be valuable contributions that can be listed alongside publications. Thus, publishing workflows in open repositories, provides further venues for cataloging an individual's research contributions. Further, workflows that are shared and identified with permanent identifiers (such as DOIs) also satisfy outreach activities that are necessary to satisfy broader impact requirements of many sponsors.

Many scientists spend much of their time performing relatively mundane data management tasks such as transforming data from one format to another, re-running analyses with updated data, and reviewing results of quality assurance and quality control procedures. Scientific workflow packages and public workflow repositories provide mechanisms that allow scientists to re-use workflow components or to repeat entire analyses without having to re-create the entire series of analytical steps. Importantly, well-documented and functional workflows allow one to trace the provenance of data and to more readily reproduce scientific results. Such capabilities allow for more thorough peer-review and will likely increase the pace of science as complex analyses will not need to be recreated from scratch.

Acknowledgments

This work was supported by: (1) INTEROP: Creation of an International Virtual Data Center for the Biodiversity, Ecological and Environmental Sciences, US National Science Foundation (NSF), award #0753138; and (2) Data Observation Network for Earth (DataONE), NSF award #0830944 under a Cooperative Agreement.

References

[proceedings]Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludäscher, B., & Mock, S. (2004). Kepler: An Extensible System for Design and Execution of Scientific Workflows. In *Proceedings of the The Future of Grid Data Environments, Global Grid Forum 10*. Retrieved November 11, 2011, from http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1311241

[proceedings]Callahan, S. P., Freire, J., Santos, E., Scheidegger, C. E., Silva, C. T., & Vo, H. T. (2006). VisTrails: visualization meets data management. *Proceedings of the 2006 ACM SIGMOD international conference on Management of data* (pp. 745–747). ACM. Retrieved November 11, 2011, from http://portal.acm.org/citation.cfm?id=1142574

[journal article]Coombes, K. R., Wang, J. & Baggerly, K. A. (2007). Microarrays: retracing steps. *Nature Med. 13,* (pp. 1276–1277). Retrieved November 11, 2011, from http://www.nature.com/nm/journal/v13/n11/full/nm1107-1276b.html

[proceedings]Cui, L., Shiyong, L., Xubo, F., Darshan, P., & Jing, H. (2009). A Task Abstraction and Mapping Approach to the Shimming Problem in Scientific Workflows. In *Proceedings of the 2009 IEEE International Conference on Services Computing (SCC '09),* (pp. 284-291). Washington DC: IEEE Computer Society. Retrieved November 11, 2011, from http://dx.doi.org/10.1109/SCC.2009.77

[journal article]Fisher, P., Hedeler, C., Wolstencroft, K., Hulme, H., Noyes, H., Kemp, S., Stevens, R., & Brass, A. (2007). A systematic strategy for large- scale analysis of genotype phenotype correlations: Identification of candidate genes involved in African trypanosomiasis. *Nucleic Acids Research 35*, (pp. 5625–5633). Retrieved November 11, 2011 from http://nar.oxfordjournals.org/content/35/16/5625.full

[journal article]Goble, C.A., Bhagat, J., Aleksejevs, S., Cruickshank, D., Michaelides, D., Newman, D., Borkum, M., Bechhofer, S., Roos, M., Li, P., & De Roure, D. (2010). myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Research 38 (suppl 2),* (pp. W677-W682). Retrieved November 11, 2011, from http://nar.oxfordjournals.org/content/38/suppl_2/W677

[journal article]Goecks, J., Nekrutenko, A., Taylor, J., & Galaxy Team, T. (2010). Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology 11 (8)*, (pp. R86). Retrieved November 11, 2011, from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2945788/?tool=pmcentrez

[journal article]Hartman, A. L., Riddle, S., McPhillips, T., Ludäscher, B., & Eisen, J. A. (2010). Introducing W.A.T.E.R.S.: a workflow for the alignment, taxonomy, and ecology of ribosomal sequences. *BMC bioinformatics 11 (1),* (pp. 317). Retrieved November 11, 2011, from http://www.biomedcentral.com/1471-2105/11/317

[journal article]Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M. Li, P. & Oinn, T. (2006). Taverna: a tool for building and running workflows of services. *Nucleic Acids Research 34, (Web Server issue),* (pp. 729-732). Retrieved November 11, 2011, from http://nar.oxfordjournals.org/content/34/suppl_2/W729.full

[journal article]Oinn, T., Greenwood, M., Addis, M., Alpdemir, N., Ferris, J., Glover, K., Goble, C., Goderis, A., Hull, D., Marvin, D., Li, P., Lord, P., Pocock, M. Senger, M., Stevens, R., Wipat, A. & Wroe, C. (2006). Taverna: lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience 18, (10),* (pp. 1067-1100). Retrieved November 11, 2011, from http://eprints.ecs.soton.ac.uk/10908/

[journal article]Reichman, O.J., Jones, M.B., & Schildhauer, M.P. (2011). Challenges and Opportunities of Open Data in Ecology. *Science 331*, (pp. 703-705). Retrieved November 11, 2011, from http://www.sciencemag.org/content/331/6018/703.full

[journal article]Wassink, I., Vet, P. E. V. D., Wolstencroft, K., Neerincx, P. B. T., Roos, M., Rauwerda, H., & Breit, T. M. (2009). Analysing Scientific Workflows: Why Workflows Not Only Connect Web Services. *2009 Congress on Services - I*, (pp. 314-321). Retrieved November 11, 2011, from http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5190670