

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

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

Rebecca Koskela
Executive Director of DataONE
University of New Mexico

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

Select Intern

Richard Littauer
University of Edinburgh/Saarland

OpenNotebook

Research and Review

myExperiment.org

Public workflow repository:

- 1,900+ workflows
- 4,850+ registered users
- 4 years old

Fosters:

- Discovery and use
- Scientific feedback
- Networking and development

RDF mining

Screen Scraping

SQL

Statistics in R

Scientific Workflows

Scientists need to process, analyze, visualize, and manage increasing amounts of scientific data.

For complex analyses, scientists often have to combine multiple processing steps into larger "analysis pipelines" that can involve a number of custom algorithms, specialized tools (e.g., statistics packages), databases and web services.

Such *scientific workflows* 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 computational steps.

Workflow Workbenches

Scientific workflow systems such as Kepler, Taverna, Vistrails, and others have emerged as a promising technology to further simplify the creation, execution, and sharing of computational workflows among scientists and laboratories. Another powerful feature of workflow systems is the ability to record data lineage and provenance information during execution, thus allowing scientists to "replay" processing steps and study data dependencies.

Literature Mining

Mendeley Group with
500~ Documents on
Workflows

Suggested Practices:

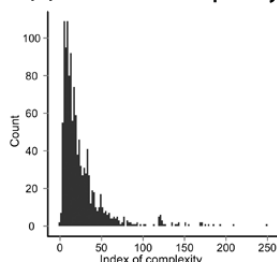
- Updating and continually managing workflows is important.
- A controlled workflow tagging vocabulary should be adopted.
- Workflow re-use could benefit significantly from the assignment of stable identifiers such as Digital Object Identifiers (DOI).
- Smaller workflows are just as important as complex ones, as they can be embedded within each other.

- Workflow reuse may be increased significantly 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
- 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).

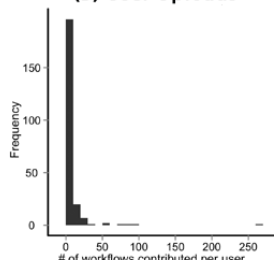
Main Findings:

- The distribution of workflow *complexity* (number of components) indicates that a large percentage of workflows consist of few components (Fig. 2A)
- Most workflow contributors submit a single workflow, whereas there are a few developers who contribute many workflows (Fig. 2B)
- 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 workflows. Workflow use also declined significantly thirty days after initial upload (Fig. 2C)
- 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 (Fig. 2D).
- 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.
- Use of workflows does not seem to be strongly related to the volume of documentation or 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.

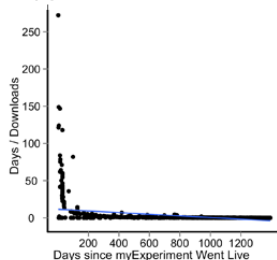
(A) Workflow Complexity



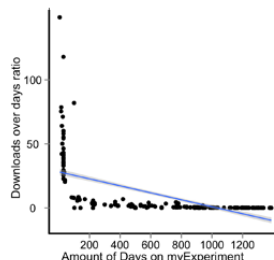
(B) User Uploads



(C) Workflow use over time



(D) Versions vs Downloads



Future Work:

- Cluster analysis of data to ascertain different classes of workflow based on use

- Mining Groups, Packs, and User Relations of myExperiment.org
- Publicizing suggested Best Practices and encouraging their adoption

Contact:

<http://Notebooks.dataone.org/workflows>
<http://www.mendeley.com/groups/1189721>
Richard.littauer@gmail.com
[@richlitt](http://richlitt)

References:

Kepler Project. <http://www.kepler-project.org>
 Taverna. <http://www.taverna.org.uk/>
 Vistrails <http://www.vistrails.org>
 myExperiment. <http://www.myexperiment.org>