Mining Abstractions in Scientific Workflows

Daniel Garijo

dgarijo@isi.edu

1 Motivation and Objectives

Scientific workflows have been adopted in the last decade to represent the computational methods used in scientific publications [11]. Scientific workflows define the set of computational tasks and dependencies needed to carry out *in silico* experiments, and besides enabling the execution of existent methods, they also facilitate visualizing methods, debugging experiments, teaching students and saving time when re-executing previous work [4].

Scientific workflows are relevant products of research, and "*key enablers for reproducibility of experiments involving large-scope computations*" [6]. However, reusing a workflow or any of its parts may become a daunting task. Scientific workflows can become large and complex heterogeneous structures, and the lack of documentation and examples often increases the difficulty in understanding their main goals [2]. In addition, there are large amounts of workflows in existing workflow repositories. A repository may contain hundreds or thousands of different workflows [10], and finding which ones are relevant to the problem at hand might become a hurdle for a researcher. In this regard, the creation of abstractions that group different workflows by certain criteria (e.g., common general functionality, shared workflow steps, etc.) is needed to improve workflow understandability.

In this work we address the issue of reusability and abstraction in scientific workflows. Our main goal is to find out whether *scientific workflow repositories can be automatically analyzed to extract commonly occurring patterns and abstractions that are useful for workflow developers.* This goal presents several challenges: 1) the *workflow representation heterogeneity* in the domain, as there is still no standard model for representing workflows and their metadata; 2) the *inadequate level of workflow abstraction* currently found in workflows, since they may contain several scientifically significant analysis steps combined with data preparation steps that are auxiliary and not central to the main purpose of the experiment; 3) the *current difficulty for workflow reuse*, as some scientists tend to reuse fragments from other workflows, but it is challenging to automatically detect which existing fragments may be useful for reuse; and 4) the *lack of support for workflow annotation*, having currently authors as the main source of documentation of the inputs, outputs and metadata of a workflow.

We tackle these challenges by describing our approach and results in Section 2, a summary of the impact and limitations of our work in Section 3 and lines of future work in Section 4.

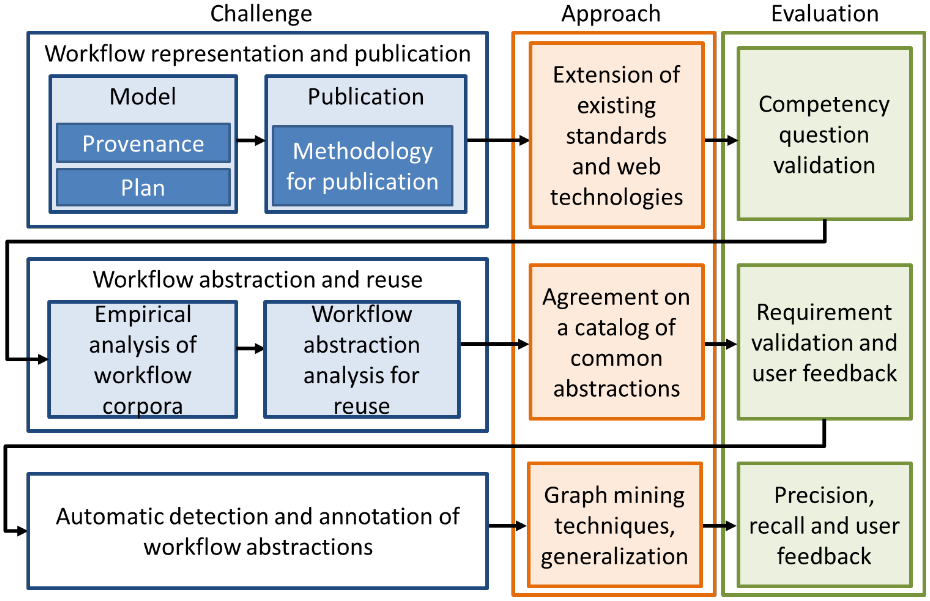
Challenge
Approach
Evaluation
Workflow representation and publication
Model
Provenance
Plan
Publication
Methodology for publication
Extension of existing standards and web technologies
Competency question validation
Workflow abstraction and reuse
Empirical analysis of workflow corpora
Workflow abstraction analysis for reuse
Agreement on a catalog of common abstractions
Requirement validation and user feedback
Automatic detection and annotation of workflow abstractions
Graph mining techniques, generalization
Precision, recall and user feedback

Fig. 1. Roadmap of the thesis work, organized by the different problems, the approach followed to tackle each one and its proposed evaluation.

2 Research Methodology, Results and Evaluation

Figure 1 shows a roadmap of the thesis work divided in three layers. Each layer represents on the left one or more research challenges described in the previous section, on the center our approach to tackle them and, on the right, the evaluation followed to validate such approach. The arrows indicate the order in which the different layers were undertaken, as each layer depends on the outcome of the previous ones. More information on the layers is provided below:

2.1 Workflow Representation and Publication:

On the top layer in Figure 1, we identified the requirements and developed the models to represent workflow templates and their associated executions. The requirements were gathered by looking at the main workflow operations available in data intensive workflow management systems, and selecting the minimum common group among all of them. The development was guided by the reuse of existing standards like the W3C Provenance Model [7]. This resulted in our **first contribution**, *the OPMW model*, a lightweight vocabulary created to represent workflow templates and workflow executions from different workflow systems, along with their relationships and metadata.

The developed model was then validated through competency questions and used to adapt an existing methodology for publishing content in the web [9] to the scientific workflow domain. The adapted methodology, our **second contribution**, consists on five steps (specification, modeling, generation, publication and exploitation), which enable publishing and consuming all the workflow related resources following the Linked Data paradigm. As a result we produced a public corpus of workflows, accessible by both humans and machines, which allowed validating the requirements established for the models, helping to refine them accordingly.

2.2 Workflow Abstraction and Reuse:

As described on the second layer of Figure 1, next we performed a manual analysis on the published workflow corpus, expanding it with the workflows available in online repositories like myExperiment CrowdLabs and Galaxy. The analyses aimed at determining whether workflows in different domains and workflow systems share common functionality among their steps, producing as an outcome our **third contribution**: *a catalog of common domain independent conceptual abstractions*, which we refer to as *common workflow motifs*. The catalog was validated and refined with each new analyzed workflow, until no further abstractions could be obtained. The most common motif is data preparation (i.e., filtering, merging, reformatting, etc.), highlighting that an important effort in the creation of workflows is dedicated to data integration, filtering and reformatting activities. The analysis also revealed that reuse is a common practice among scientists publishing their workflows.

We then analyzed the current practices of users in workflows (and their fragments) in terms of workflow reuse, and we related the results to the abstractions obtained in the catalog. This allowed us to focus on the automatic detection of those abstractions that are relevant for workflow reuse. These analyses constitute our **fourth contribution**, and are unique on measuring sub-workflow reuse in a workflow corpus.

2.3 Automatic Detection and Annotation of Workflow Abstractions:

As described on the bottom layer of Figure 1, we addressed next the gap between a corpus of workflows and the catalog of abstractions defined in the previous layer. Our **fifth contribution** consists on a novel approach for the automated detection and filtering of workflow fragments using existing graph mining techniques, which we refer to as *FragFlow*. Our approach is able to detect successfully commonly occurring workflow fragments and filter the resultant fragments to simplify the results. In order to assess our approach we defined metrics for evaluating the usefulness of a fragment, based on precision and recall, in workflow corpora created by users from two different workflow systems. Additionally, we validated our results with further feedback from three domain experts.

All of our detected abstractions are supported by a novel framework for annotating them in workflows. Our **sixth contribution** consists on the definition of two vocabularies for this purpose. The first one is the workflow motifs vocabulary (wf-motifs), which defines relationships to annotate workflows and workflow steps with any of the motifs described on the catalog. The second one is the workflow fragment description vocabulary (Wf-fd), which provides the means to expose the links found between commonly occurring workflow fragments and the original workflows where they were found.

3 Significance and Limitations

The contributions of this thesis have already started to be used in recent work by other researchers. The OPMW model has been used for exposing automatically annotated bio-informatic workflows[1] [3] and for documenting the results of scientific research.[2] Our catalog of motifs has been adopted and extended by analyzing workflows in distributed environments [8] and summarizing workflows [1]. Our other contributions influence areas such as workflow exploration, visualization, discovery and understanding, as we have shown in recent work [5].

However, our work is not without limitations. Regarding workflow representation and publication, our OPMW model aims for simplicity by capturing the minimum set of workflow operations common to most workflow systems, representing as labeled directed acyclic graphs. More complex workflow structures such as sub-workflows, loops and conditionals are not supported, as they are not common in data-intensive computational experiments.

Our catalog of workflow motifs is the result of an analysis on workflows from different domains and platforms, but it could be expanded with other types of workflows, such as those executed in distributed systems.

Finally, regarding the automatic detection and annotation of workflow abstractions, we have focused on this thesis on finding those abstractions typically useful for reuse. The detection of other types of abstractions, such as the ones included in our workflow motif catalog, remains as an open research challenge.

4 Future Work

In this thesis we have addressed the challenges associated to the automatic detection common reusable abstractions in scientific workflows. Our work may be extended to address three main open lines of research, further described below.

The first line is related to workflow generalization. Our approach can exploit existing taxonomies of components that describe workflow steps at different levels of abstraction (e.g., quick sort and merge sort are both a type of sorting algorithm). Devising a method to learn these taxonomies automatically from domain knowledge (e.g, scientific publications), instead of domain experts, would help linking different workflows to each other.

A second challenging line of is, given a workflow as input, to recognize and annotate all its motifs automatically. This would help understand the workflow

[1] https://openscience.adaptcentre.ie/publications/2017/eswc/

[2] http://linked-research.270a.info/linked-research.html

better, as in many cases there is not enough documentation to determine what a particular step of a workflow does unless we execute it.

The third line of work refers to facilitating workflow reuse. Common fragments can be compared and ranked. A recommendation based on similar workflow fragments may provide similar or better results than other approaches based on tags and descriptions.

For all three lines of work, we believe that the role of semantics is critical to represent, expose and interoperate among different levels of abstraction in scientific workflows.

References

1. Alper, P., Belhajjame, K., Goble, C., Karagoz, P.: Small is beautiful: Summarizing scientific workflows using semantic annotations. In: Big Data (BigData Congress), 2013 IEEE International Congress on. pp. 318–325 (June 2013)
2. Belhajjame, K., Roos, M., Garcia-Cuesta, E., Klyne, G., Zhao, J., De Roure, D., Goble, C., Gomez-Perez, J.M., Hettne, K., Garrido, A.: Why workflows break: Understanding and combating decay in Taverna workflows. In: Proceedings of the 2012 IEEE 8th International Conference on E-Science (e-Science). pp. 1–9. Washington, DC, USA (2012)
3. García-Jiménez, B., Wilkinson, M.D.: Automatic annotation of bioinformatics workflows with biomedical ontologies. In: Leveraging Applications of Formal Methods, Verification and Validation. Specialized Techniques and Applications, Lecture Notes in Computer Science, vol. 8803, pp. 464–478. Springer Berlin (2014)
4. Garijo, D., Corcho, O., Gil, Y., Braskie, M.N., Hibar, D., Hua, X., Jahanshad, N., Thompson, P., W.Toga, A.: Workflow reuse in practice: A study of neuroimaging pipeline users. In: 10th IEEE International Conference on eScience 2014 (2014)
5. Garijo, D., Gil, Y., Corcho, O.: Abstract, link, publish, exploit: An end to end framework for workflow sharing. Future Generation Computer Systems pp. – (2017)
6. Gil, Y., Deelman, E., Ellisman, M., Fahringer, T., Fox, G., Gannon, D., Goble, C., Livny, M., Moreau, L., Myers, J.: Examining the challenges of scientific workflows. Computer 40(12), 24–32 (Dec 2007)
7. Moreau, L., Missier, P., Belhajjame, K., BF́ar, R., Cheney, J., Coppens, S., Cresswell, S., Gil, Y., Groth, P., Klyne, G., Lebo, T., McCusker, J., Miles, S., Myers, J., Sahoo, S., Tilmes, C.: PROV-DM: The PROV Data Model. W3C Recommendation. Tech. rep., WWW Consortium (2013)
8. Olabarriaga, S.D., Jaghoori, M.M., Korkhov, V., van Schaik, B., van Kampen, A.: Understanding workflows for distributed computing: Nitty-gritty details. In: Proceedings of the 8th Workshop on Workflows in Support of Large-Scale Science. pp. 68–76. WORKS '13, ACM, New York, NY, USA (2013)
9. Radulovic, F., Poveda-Villalón, M., Vila-Suero, D., Rodríguez-Doncel, V., García-Castro, R., Gómez-Pérez, A.: Guidelines for Linked Data generation and publication: An example in building energy consumption. Automation in Construction 57, 178 – 187 (2015)
10. Roure, D.D., Goble, C.A., Stevens, R.: The design and realisation of the myExperiment virtual research environment for social sharing of workflows. Future Generation Comp. Syst. 25(5), 561–567 (2009)
11. Taylor, I.J., Deelman, E., Gannon, D.B., Shields, M.: Workflows for e-Science: Scientific Workflows for Grids. Springer-Verlag New York, Inc. (2006)