Mining Abstractions in Scientific Workflows

Daniel Garijo

dgarijo@isi.edu

1 Motivation and Objectives

Scientific workflows capture the set of computations and data dependencies that are needed to carry out computational experiments. Scientific workflows have been adopted during the last decade to represent the computational methods used in scientific publications [19], in domains like astronomy [18], brain image analysis [3] and bio-informatics [20] among others. Besides enabling the execution of existing methods, scientific workflows have proved to facilitate debugging experiments, teaching students, visualizing the steps that produced a result and saving time when re-executing previous work [6] [12].

Scientific workflows are important products of research and "*key enablers for reproducibility of experiments involving large-scope computations*" [11]. Workflows also facilitate incorporating methods developed by other scientists into new experiments [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 [17]. A repository may contain hundreds or thousands of different workflows, and finding which ones are relevant to the problem at hand might become a hurdle for a researcher. In this regard, abstractions that group different workflows by certain criteria (e.g., common general functionality, shared workflow steps, etc.) are needed to improve workflow understandability.

This thesis proposes a novel approach to automatically analyze workflow repositories for extracting commonly occurring patterns and abstractions that facilitate workflow reuse. This goal presents three major challenges: 1) *workflow representation heterogeneity*, as there is 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 a workflow and 3) *identifying useful workflow fragments*, as scientists tend to reuse fragments from other workflows, but it is challenging to detect which existing fragments may be useful. Our approach has been evaluated with a large collection of workflows developed by scientists.

We describe our approach and results in Section 2, summarizing the significance and limitations of our work in Section 3 and outlining future lines of work in Section 4.

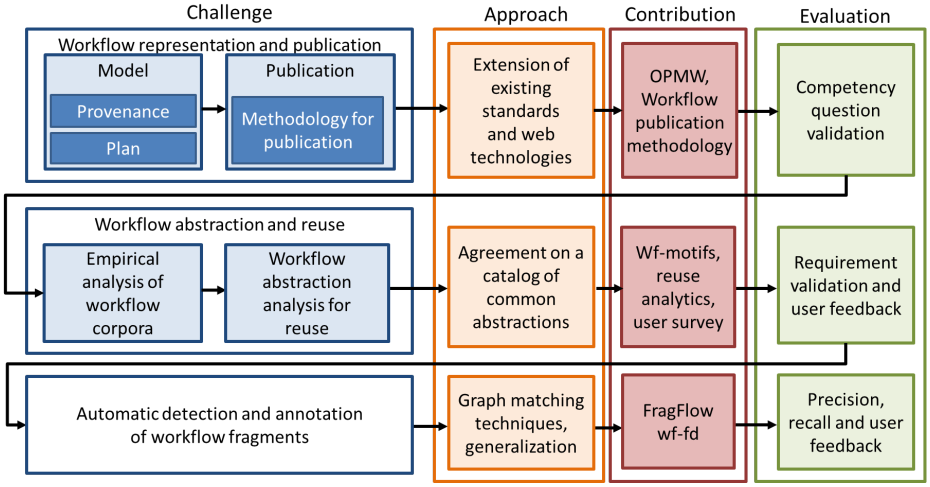
Challenge
Workflow representation and publication
Model
Provenance
Plan
Publication
Methodology for publication
Approach
Extension of existing standards and web technologies
Contribution
OPMW, Workflow publication methodology
Evaluation
Competency question validation
Workflow abstraction and reuse
Empirical analysis of workflow corpora
Workflow abstraction analysis for reuse
Agreement on a catalog of common abstractions
Wf-motifs, reuse analytics, user survey
Requirement validation and user feedback
Automatic detection and annotation of workflow fragments
Graph matching techniques, generalization
FragFlow wf-fd
Precision, recall and user feedback

Fig. 1. Roadmap of the thesis work, organized by the different problems, the approach followed to tackle each one, their associated contributions and their evaluation criteria.

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 followed by their corresponding contribution and, on the right, the evaluation followed to validate each of them. 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 each of the three layers is provided in the rest of this section.

2.1 Workflow Representation and Publication:

On the first layer we identified requirements and developed vocabularies to represent workflow templates and their associated executions. The requirements were gathered by analyzing the most common workflow features available in data intensive workflow management systems (e.g., representation of the workflow inputs, steps, parameters, etc.). The development was guided by extending existing standards like the W3C Provenance Model [14] and commonly used vocabularies like the Open Provenance Model [13]. This resulted in our **first contribution**, *the OPMW ontology* [8], a lightweight vocabulary created to represent workflow templates and workflow executions from different workflow systems, along with their relationships and metadata. OPMW was validated through competency questions and used to represent workflows from different workflows systems [9]

Next, we adapted an existing methodology for publishing content in the web [16] to the scientific workflow domain. This *workflow publication 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. We used this approach to exchange workflows across different tools for generating, executing, viewing and mining workflows [10]; and we produced a public corpus of workflows, accessible by both humans and machines, which is open and available to the community.

2.2 Workflow Abstraction and Reuse:

Next we performed a manual analysis on the published workflow corpus, expanding it with the workflows available in repositories such as myExperiment, CrowdLabs and Galaxy [5]. The analysis 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 *workflow motifs*. This motif 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.), demonstrating empirically that 60-80% of the steps in workflows are dedicated to data pre-processing activities. The analysis also revealed that reuse is a common practice among scientists publishing their workflows, with two main motifs related to reuse: *internal macros*, which refers to groups of steps reused within a single workflow, and *external macros*, i.e., workflows that are reused as part of another workflow.

In order to understand how useful were these two motifs, we surveyed more than 20 neuro-image scientists who were familiar with scientific workflows [6]. We also analyzed their current practices in terms of reuse in three corpora of workflows [7]. The results of our survey showed that authors consider useful to reuse both workflows and workflow fragments created by others, and hence, automatically detecting external and internal macros may facilitate their work. Both the survey and analysis constitute our **fourth contribution**, and are unique on measuring sub-workflow reuse and usefulness in a workflow corpus.

2.3 Automatic Detection and Annotation of Workflow Abstractions:

We addressed next the automated extraction of both internal and external macros in a corpus of workflows. Our **fifth contribution** consists on a novel approach for the automated detection and filtering of frequent workflow fragments, which we refer to as *FragFlow* [7]. FragFlow relies on existing exact and inexact graph matching techniques to extract frequent workflow fragments, using different heuristics for inexact matching (based on the size of a graph or the minimum description length needed to encode it) and filtering techniques for removing redundant fragments. Furthermore, FragFlow allows generalizing workflows to find fragments at higher levels of abstraction, based on their common functionality (e.g., a *quick sort* step and a *merge sort* step in a workflow may be generalized to a *sort* step, which is more general). FragFlow also enables configuring how fragments are mined, based on the minimum frequency of the fragment, its maximum size or the type of algorithm selected to search. In order to assess our approach, we defined metrics for evaluating the usefulness of a fragment by comparing our results to workflow corpora from two different workflow systems. The workflow corpora contains more than 940 different real world neuro-imaging workflows which were created by scientists in their day-to-day work. Additionally, we evaluated the usefulness of our fragments with three neuro-image scientists, obtaining positive results.

Our approach supports annotating motifs and common workflow fragments, linking them back to the original workflows where they appear. Our **sixth contribution** consists on the definition of two vocabularies for this purpose. The first one is the workflow motifs vocabulary (wf-motifs[1]), which defines relationships to annotate workflows and workflow steps with any of the motifs described in the catalog. The second one is the workflow fragment description vocabulary (Wf-fd[2]), 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

This thesis presents one of the first works on mining scientific workflows to find reusable fragments, and it is the first to have been evaluated on more than 900 workflows from a single domain. 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[3] [4] and for documenting the results of scientific research.[4] Our catalog of motifs has been adopted and extended by analyzing workflows in cloud environments [15] and summarizing workflows [1].

However, our work is not without limitations. Regarding workflow representation and publication, our OPMW model aims for simplicity by representing workflows as labeled directed acyclic graphs. More complex workflow structures such as loops and conditionals are not supported, as they are not common in data-intensive computational experiments. In addition, 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 cloud based environments. Finally, regarding the automatic detection and annotation of workflow abstractions, we have focused on this thesis on finding abstractions 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. The graph matching algorithms used present some limitations as well. For example, when setting a low value for the frequency of fragments

[1] http://purl.org/net/wf-motifs

[2] http://purl.org/net/wf-fd

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

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

to be mined, the system may find too many potential candidates and take time to produce an answer. A similar case happens when the target fragments to be mined are large, due to the amount of fragment candidates to be explored.

4 Future Work

This thesis addressed the challenges associated to the automatic detection common reusable fragments 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. Devising a method to learn these taxonomies automatically from domain knowledge (e.g, scientific publications), instead of domain experts, would help generalizing different workflows based on their common functionality.

A second challenging line of future work is, given a workflow as input, to recognize and annotate all its motifs automatically. This would help understand the workflow better, as in many cases there is not enough documentation to determine what a particular step of a workflow does.

The third line of work refers to facilitating workflow reuse. Common fragments can be compared and ranked. A recommendation to the user based on similar workflow fragments may provide 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. Dinov, I.D., Horn, J.D.V., Lozev, K.M., Magsipoc, R., Petrosyan, P., Liu, Z., MacKenzie-Graham, A., Eggert, P., Parker, D.S., Toga, A.W.: Efficient, distributed and interactive neuroimaging data analysis using the LONI Pipeline. In: Frontiers in Neuroinformatics. vol. 3 (2009)
4. 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)
5. Garijo, D., Alper, P., Belhajjame, K., Corcho, O., Gil, Y., Goble, C.: Common motifs in scientific workflows: An empirical analysis. Future Generation Computer Systems 36, 338 – 351 (2014)

6. 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)
7. Garijo, D., Corcho, O., Gil, Y., Gutman, B.A., Dinov, I.D., Thompson, P., Toga, A.W.: Fragflow: Automated fragment detection in scientific workflows. In: Proceedings of the 2014 IEEE 10th International Conference on e-Science - Volume 01. pp. 281–289. E-SCIENCE '14, IEEE Computer Society, Washington, DC, USA (2014)
8. Garijo, D., Gil, Y.: A new approach for publishing workflows: Abstractions, standards, and Linked Data. In: Proceedings of the 6th workshop on Workflows in support of large-scale science. pp. 47–56. ACM, Seattle (2011)
9. Garijo, D., Gil, Y., Corcho, O.: Towards workflow ecosystems through semantic and standard representations. In: Proceedings of the 9th Workshop on Workflows in Support of Large-Scale Science. pp. 94–104. WORKS '14, IEEE Press, Piscataway, NJ, USA (2014)
10. Garijo, D., Gil, Y., Corcho, O.: Abstract, link, publish, exploit: An end to end framework for workflow sharing. Future Generation Computer Systems pp. – (2017)
11. 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)
12. Goderis, A.: Workflow re-use and discovery in Bioinformatics. Ph.D. thesis, School of Computer Science, The University of Manchester (2008)
13. Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E., den Bussche., J.V.: The open provenance model core specification (v1.1). Future Generation Computer Systems, 27(6). (2011)
14. 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)
15. 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)
16. 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)
17. 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)
18. Ruiz, J., Garrido, J., Santander-Vela, J., Sánchez-Expósito, S., Verdes-Montenegro, L.: AstroTaverna: Building workflows with Virtual Observatory services. Astronomy and Computing 7–8, 3 – 11 (2014)
19. Taylor, I.J., Deelman, E., Gannon, D.B., Shields, M.: Workflows for e-Science: Scientific Workflows for Grids. Springer-Verlag New York, Inc. (2006)
20. Wolstencroft, K., Haines, R., Fellows, D., Williams, A., Withers, D., Owen, S., Soiland-Reyes, S., Dunlop, I., Nenadic, A., Fisher, P., Bhagat, J., Belhajjame, K., Bacall, F., Hardisty, A., de la Hidalga, A.N., Vargas, M.P.B., Sufi, S., Goble, C.: The Taverna workflow suite: designing and executing workflows of web services on the desktop, web or in the cloud. Nucleic Acids Research (2013)