Future Generation Computer Systems 36 (2014) 338–351

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Common motifs in scientific workflows: An empirical analysis

Daniel Garijo [a,*,1], Pinar Alper [b,1], Khalid Belhajjame [b], Oscar Corcho [a], Yolanda Gil [c], Carole Goble [b]

[a] Ontology Engineering Group, Universidad Politécnica de Madrid, Spain

[b] School of Computer Science, University of Manchester, United Kingdom

[c] Information Sciences Institute, Department of Computer Science, University of Southern California, United States

HIGHLIGHTS

- We present an empirical analysis performed over 260 scientific workflow descriptions.
- We define a catalog of domain independent abstractions for workflows.
- We discuss the distribution of the abstractions across different workflow systems.
- Different workflow systems share a common core of workflow abstractions.
- Data preparation is an obstacle for workflow understandability.

ARTICLE INFO

Article history:
Received 1 February 2013
Received in revised form
2 August 2013
Accepted 5 September 2013
Available online 21 September 2013

Keywords:
Scientific workflows
Workflow motif
Workflow pattern
Taverna
Wings
Galaxy
Vistrails

ABSTRACT

Workflow technology continues to play an important role as a means for specifying and enacting computational experiments in modern science. Reusing and re-purposing workflows allow scientists to do new experiments faster, since the workflows capture useful expertise from others. As workflow libraries grow, scientists face the challenge of finding workflows appropriate for their task, understanding what each workflow does, and reusing relevant portions of a given workflow. We believe that workflows would be easier to understand and reuse if high-level views (abstractions) of their activities were available in workflow libraries. As a first step towards obtaining these abstractions, we report in this paper on the results of a manual analysis performed over a set of real-world scientific workflows from Taverna, Wings, Galaxy and Vistrails. Our analysis has resulted in a set of *scientific workflow motifs* that outline (i) the kinds of data-intensive activities that are observed in workflows (*Data-Operation motifs*), and (ii) the different manners in which activities are implemented within workflows (*Workflow-Oriented motifs*). These motifs are helpful to identify the functionality of the steps in a given workflow, to develop best practices for workflow design, and to develop approaches for automated generation of workflow abstractions.

1. Introduction

A scientific workflow is a template defining the set of tasks needed to carry out a computational experiment [1]. Scientific workflows have been increasingly used in the last decade as an instrument for data intensive science. Workflows serve a dual function: first, as detailed documentation of the scientific method used for an experiment (i.e. the input sources and processing steps taken for the derivation of a certain data item), and second, as re-usable, executable artifacts for data-intensive analysis. Scientific workflows are composed of a variety of data manipulation activities such as data movement, data transformation, data analysis and data visualization to serve the goals of the scientific study. The composition is done through the constructs made available by the workflow system used, and is largely shaped by the function undertaken by the workflow and the environment in which the system operates.

A variety of workflow systems, both open source (e.g. Taverna [2], Wings [3], Galaxy [4], Vistrails [5], Kepler [6], ASKALON [7]) and commercial (e.g. Pipeline Pilot[2]) are in use in a variety of scientific disciplines such as genomics, astronomy, cheminformatics, etc. A workflow is a software artifact, and once developed and tested, it

* Corresponding author. Tel.: +34 667949892.
E-mail addresses: dgarijo@fi.upm.es (D. Garijo), alperp@cs.manchester.ac.uk (P. Alper), khalidb@cs.manchester.ac.uk (K. Belhajjame), ocorcho@fi.upm.es (O. Corcho), gil@isi.edu (Y. Gil), carole.goble@cs.manchester.ac.uk (C. Goble).

[1] The first and second authors have contributed equally to the work presented in this paper.

[2] http://accelrys.com/products/pipeline-pilot/.

0167-739X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.future.2013.09.018

can be shared and exchanged between scientists. Other scientists can then reuse existing workflows in their experiments, e.g., as sub-workflows [8].

Workflow reuse presents several advantages [9]: allowing for principled attribution of established methods, improving quality through incremental/evolutionary workflow development (by leveraging the expertise of previous users), and making scientific processes more efficient. Users can also re-purpose existing workflows to adapt them to their needs [9]. Emerging workflow repositories such as myExperiment [10] and CrowdLabs [11] have made publishing and finding workflows easier, but scientists still face the challenges of understanding and reusing the available workflows.

A major difficulty in understanding workflows is their complex nature. A workflow may contain several scientifically-significant analysis steps, combined with other data preparation or result delivery activities, and in different implementation styles depending on the environment and context in which the workflow is executed. This difficulty in understanding stands in the way of reusing workflows.

Through an analysis of the current practices in scientific workflow development, we pursue the following objectives:

1. To reverse-engineer the set of current practices in workflow development through an empirical analysis.
2. To identify workflow abstractions that would facilitate understandability and therefore effective reuse.
3. To detect potential information sources that can be used to inform the development of tools for creating workflow abstractions.

In this paper we present the result of an empirical analysis performed over 260 workflow descriptions from Taverna [2], Wings [3], Galaxy [4] and Vistrails [5]. Based on this analysis, we propose a catalog of domain independent conceptual abstractions for workflow steps that we call *scientific workflow motifs*. Motifs are provided through (i) a characterization of the kinds of data-operation activities that are carried out within workflows, which we refer to as *Data-Operation motifs*, and (ii) a characterization of the different manners in which those activity motifs are realized/implemented within workflows, which we refer to as *Workflow-Oriented motifs*.

This paper extends our previous work [12], which performed an analysis of 177 workflows from Wings and Taverna. The new contributions reported on in this paper are an extension of the related work in Section 2, the addition and extension of scientific domains from Wings and Taverna workflows (Social Network Analysis, Astronomy and Domain Independent) in Sections 3 and 5; and the analysis of workflows from the Galaxy and Vistrails systems among different domains (Genomics, Text Mining, Domain Independent and Medical Informatics). Finally, we have also revisited the motif catalog (Section 4), our previous results (Section 5) and conclusions (Section 7) according to our new findings.

2. Related work

Our motifs can be seen as higher-level patterns observed in scientific workflows. "Workflow patterns" have been extensively studied [13], where inventories of possible patterns are developed based on workflow constructs that are possible in different languages, along with the ways to combine those constructs. Scientific workflows typically use a dataflow paradigm rather than a control flow paradigm that is more typical of business workflows [14], and generic data-intensive usage patterns[3] are described in [15]. Other classifications are based on the intrinsic properties of the workflows (size, structure, branching factor, resource usage, etc.) [16,17] and their environmental characteristics (makespan, speed-up, success rate, etc.) [17]. Rather than specifying what is theoretically possible with the given constructs, our work is instead based on an empirical analysis to detect similar data-intensive activities that recur in workflows across different domains and workflow systems. In addition, our work offers a complementary perspective in that we aim to understand groupings of workflow steps that form a meaningful high-level data manipulation operation.

In Software Engineering, the term "pattern" refers to established best practices to solve recurring problems. In this regard patterns represent good and exemplary practice. In [18] the authors outline anti-patterns in scientific workflows, namely redundancy and structural conflicts. The authors go on to provide a solution to address the redundancy anti-pattern. Particularly due to this perception of the term "pattern", in this paper we opted to use the term "motif" for our classification of tasks. Our objective is to take a snapshot of the existing set of activities in workflows, rather than try to prescribe a best practice.

Our Data-Operation motifs can be seen as a domain-independent classification of tasks within scientific workflows. Similar analyses have been done in a domain-specific manner in areas such as bioinformatics, based on user studies [19]. Combined with such-domain specific classifications, motifs can make way for specification of abstract workflow templates, which can be elaborated to concrete workflows prior to their execution [20].

Another work, somewhat closer to our study in spirit, is an automated analysis of workflow scripts from the Life Science domain [21]. This work aims to deduce the frequency of different kinds of technical ways of realizing workflow steps (e.g. service invocations, local "scientist-developed" scripting, local "ready-made" scripts, etc.). [21] also drills down into the category of local ready-made scripts, to outline a functional breakdown of their activity categories such as data access or data transformation. While this provides an insight into the kind of activities undertaken in workflows, it focuses on characterizing local task types. Our approach is different from this work as we focus on detecting multi-step activities with many realizations (not just individual steps).

[22] extends the categories defined in [21] identifying subcategories at a processor level by analyzing 898 workflows in myExperiment. The main difference with our analysis is that some of the proposed categories are based on technological features of the processors (i.e., the type of script) for highlighting workflow reuse among the dataset, while our catalog relies on their functional characteristics.

Finally, Problem Solving Methods (PSMs) is another area of related work. PSMs describe the reasoning process to achieve the goal of a task in an implementation and domain-independent manner [23]. Some libraries aim to model the common processes in scientific domains [24], which could be further refined with the motifs proposed in this work.

3. Analysis setup

For the purposes of the analysis, we used workflows from Taverna [2], Wings [3], Galaxy [4] and Vistrails [5]. These systems have different features:

- Taverna [2] can operate in different execution environments and provides several possibilities of deployment. Taverna is available as a workbench,[4] which embodies a desktop design

[3] http://www.workflowpatterns.com/patterns/data/.

[4] Taverna Workbench http://www.taverna.org.uk/download/workbench/.