Accepted Manuscript

Using ontologies for verification and validation of workflow-based experiments

Tomasz Miksa, Andreas Rauber

PII: S1570-8268(17)30011-2
DOI: http://dx.doi.org/10.1016/j.websem.2017.01.002
Reference: WEBSEM 427

To appear in: *Web Semantics: Science, Services and Agents on the World Wide Web*

Received date: 12 May 2016
Revised date: 13 October 2016
Accepted date: 17 January 2017

Please cite this article as: T. Miksa, A. Rauber, Using ontologies for verification and validation of workflow-based experiments, *Web Semantics: Science, Services and Agents on the World Wide Web* (2017), http://dx.doi.org/10.1016/j.websem.2017.01.002

Using ontologies for verification and validation of workflow-based experiments

Tomasz Miksa

SBA Research, Wien, Austria

Andreas Rauber

Vienna University of Technology, Wien, Austria

Abstract

Scientific experiments performed in the eScience domain require special tooling, software, and workflows that allow researchers to link, transform, visualize and interpret data. Recent studies report that such experiments often cannot be replicated due to differences in the underlying infrastructure. The provenance collection mechanisms were built into workflow engines to increase research replicability. However, the traces do not contain the execution context that consists of software, hardware and external services used to produce the result which may change between executions.

The problem thus remains on how to identify such context and how to store such data. To address this challenge we propose the context model that integrates ontologies which describe workflow and its environment. It includes not only high level description of workflow steps and services but also low level technical details on infrastructure, including hardware, software, and files. In this paper we discuss which ontologies that compose the context model must be instantiated to enable verification of a workflow re-execution. We use a tool that monitors a workflow execution and automatically creates the context model. We also authored the VPlan ontology that enables modelling validation requirements. It contains a controlled vocabulary of metrics that can be used for quantification of requirements. We evaluate the proposed ontologies on five Taverna workflows that differ in the degree on which they depend on additional software and services.

The results show that the proposed ontologies are necessary and can be used for verification and validation of scientific workflows re-executions in different environments without the necessity of accessing the original environment at the same time. Thus the scientists can state whether the scientific experiment is replicable.

Keywords: verification, validation, workflow, reproducibility, context model

1. Introduction

In many natural science disciplines, complex and data driven experiments form the basis of research [1]. Scientific breakthroughs would not be possible without special tooling, software and processes that allow researchers to link, transform, visualise and interpret the data [2]. The low maturity of tools and the possible lack of scientific scrupulousness [3] led to a low reproducibility and replicability of experiments [4], [5]. Many problems can be attributed to the fact that the software is not available any more [4]. This may appear to be a pure management issue that can be overcome by imposing better policies, but recent findings show that also the context in which the software is run, that is the infrastructure and the third party dependencies, can have a crucial impact on the final results delivered by a computational experiment. In [6] the authors demonstrate that a different version of the operating system used for neuroimaging analysis in clinical research produces different visualisations. This implies that in order to replicate the same result, not only the same data must be used, but also it must be run on an equivalent software stack.

Workflow engines were proposed in order to bring some standardisation, as well as to hide complexity of the underlying infrastructure. Workflow engines like VisTrails[7], Kepler [8] or Taverna [9] have become popular in research areas like Astronomy, Bioinformatics or Clinical Research. They enable researchers to graphically represent their experiments in form of workflows that can be built using pre-defined elements. These elements range from dedicated data parsers to in-