ARTICLE IN PRESS

Web Semantics: Science, Services and Agents on the World Wide Web ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Web Semantics: Science, Services and Agents on the World Wide Web

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

Using a suite of ontologies for preserving workflow-centric research objects

Khalid Belhajjame [a,*], Jun Zhao [b], Daniel Garijo [c], Matthew Gamble [d], Kristina Hettne [e], Raul Palma [f], Eleni Mina [e], Oscar Corcho [c], José Manuel Gómez-Pérez [g], Sean Bechhofer [d], Graham Klyne [h], Carole Goble [d]

[a] PSL, Université Paris Dauphine, LAMSADE, France
[b] School of Computing and Communications, Lancaster University, UK
[c] Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
[d] School of Computer Science, University of Manchester, UK
[e] Leiden University Medical Center, Leiden, The Netherlands
[f] Poznan Supercomputing and Networking Center, Poznan, Poland
[g] iSOCO, Madrid, Spain
[h] Department of Zoology, University of Oxford, UK

ARTICLE INFO

Article history:
Received 28 August 2013
Received in revised form
11 December 2014
Accepted 26 January 2015
Available online xxxx

Keywords:
Research object
Scientific workflow
Preservation
Annotation
Ontologies
Provenance

ABSTRACT

Scientific workflows are a popular mechanism for specifying and automating data-driven *in silico* experiments. A significant aspect of their value lies in their potential to be reused. Once shared, workflows become useful building blocks that can be combined or modified for developing new experiments. However, previous studies have shown that storing workflow specifications alone is not sufficient to ensure that they can be successfully reused, without being able to understand what the workflows aim to achieve or to re-enact them. To gain an understanding of the workflow, and how it may be used and repurposed for their needs, scientists require access to additional resources such as annotations describing the workflow, datasets used and produced by the workflow, and provenance traces recording workflow executions.

In this article, we present a novel approach to the preservation of scientific workflows through the application of *research objects*—aggregations of data and metadata that enrich the workflow specifications. Our approach is realised as a suite of ontologies that support the creation of workflow-centric research objects. Their design was guided by requirements elicited from previous empirical analyses of workflow decay and repair. The ontologies developed make use of and extend existing well known ontologies, namely the Object Reuse and Exchange (ORE) vocabulary, the Annotation Ontology (AO) and the W3C PROV ontology (PROVO). We illustrate the application of the ontologies for building Workflow Research Objects with a case-study that investigates Huntington's disease, performed in collaboration with a team from the Leiden University Medial Centre (HG-LUMC). Finally we present a number of tools developed for creating and managing workflow-centric research objects.

1. Introduction

As science becomes increasingly data driven, many scientists have adopted *workflows* as a means to specify and automate repetitive experiments that retrieve, integrate, and analyse datasets using distributed resources [1]. Using a workflow, an experiment can be defined as a graph where the nodes represent analysis operations, which can be supplied locally or accessible remotely, and edges specify dependencies between the operations.

* Corresponding author.
E-mail addresses: Khalid.Belhajjame@dauphine.fr (K. Belhajjame), j.zhao5@lancaster.ac.uk (J. Zhao), dgarijo@fi.upm.es (D. Garijo), matthew.gamble@cs.man.ac.uk (M. Gamble), k.m.hettne@lumc.nl (K. Hettne), rpalma@man.poznan.pl (R. Palma), ocorcho@fi.upm.es (O. Corcho), jmgomez@isoco.com (J.M. Gómez-Pérez), sean.bechhofer@cs.man.ac.uk (S. Bechhofer), Graham.Klyne.@zoo.ox.ac.uk (G. Klyne), carole.goble@cs.man.ac.uk (C. Goble).

http://dx.doi.org/10.1016/j.websem.2015.01.003
1570-8268/

Please cite this article in press as: K. Belhajjame, et al., Using a suite of ontologies for preserving workflow-centric research objects, Web Semantics: Science, Services and Agents on the World Wide Web (2015), http://dx.doi.org/10.1016/j.websem.2015.01.003

The value of a workflow definition is not limited to its original author, or indeed to the original study for which it was created. Once specified, a workflow can be re-used or repurposed by other scientists. This reuse can be as a means of understanding an experimental process, replicating a previous experimental result, or even using the workflow as a building-block in the design of new workflow-based experiments. To support this potential for reuse, public repositories such as myExperiment [2] and CrowdLabs [3] can be used by scientists to publish workflow definitions and share them over the web.

However, sharing just the workflow specifications is not always sufficient to guarantee successfully reuse. Previous empirical analysis of 92 workflows from myExperiment [4] has demonstrated that nearly 80% of the workflows suffered from *decay* in the sense that they could not be understood, or executed when downloaded. These failures were shown to be a result of one or more of the following issues:

(i) **Insufficient documentation**. The user was unable to grasp the analysis or experiment implemented by the workflow due to the lack of descriptions of its inputs, intermediate steps, and outputs.
(ii) **Missing example data**. Even in situations where the users were able to understand the overall analysis implemented by the workflow, it was difficult to determine what kind of data values to use as inputs to successfully execute that workflow.
(iii) **Volatile third-party resources**. Many workflows could not be run because the third party resources they rely on were no longer available (e.g., web services implementing their steps). For example, the SOAP web services provided by KEGG[1] to query its databases have been replaced by Rest Web Services. As a result, a large number of the workflows that use the SOAP services in myExperiment could not be run.
(iv) **Execution environment**. In certain cases, the execution of the workflow required some specific software infrastructures to be installed locally, e.g., the R statistical tool.

It is clear that in order to ensure the successful *preservation* of workflows, there is a need to change how we make them. Specifically we understand successful workflow preservation to be *the immediate and continued ability to understand, run, and reuse the experimental process described by a workflow*.

Issues 1, 2, and 4 above are all introduced at the point of the workflow's publication, through the omission of necessary supporting data or metadata. Issue 3 is instead a consequence of using 3rd party services as part of a workflow, and is a relevant issue in *workflow decay* [4]. Whilst the loss of 3rd party services is out of the control of original authors, there are a number of approaches to remedy this type of workflow decay by making use of metadata – such as additional semantic descriptions about the services used [5], or provenance information [6–8] – all of which can be either provided by the author of the workflow or automatically tracked and computed.

In light of this we propose a novel approach to workflow preservation where workflow specifications are not published in isolation, but are instead accompanied by auxiliary resources and additional metadata. Specifically we have chosen to adopt and extend the *Research Object* approach proposed in [9].

The Research Object approach defines an extendable model of data aggregation, and semantic annotation. At its core, the model allows us to describe aggregations of data and enrich that aggregation with supporting metadata. This aggregation can then be published and exchanged as a single artifact. Using this approach we have built a unit of publication that combines the workflow specification along with the supporting data and metadata required to improve preservation and the potential for reproducibility. Our implementation of workflow-centric research objects is realised as a series of ontologies that support both a core model of aggregation and the domain specific workflow preservation requirements.

In this paper we make the following contributions:

- We present a series of requirements for the data and metadata needed to accompany workflow specifications to support workflow preservation.
- We outline four ontologies that we have developed in response to those requirements, that can be used to describe Workflow-Centric Research Objects.
- We present a collection of tools that make use of those ontologies in the support and management of Workflow Research Objects.
- Finally, we present a series of competency queries that demonstrate how Workflow Research Objects support workflow preservation.

The remainder of this paper is organised as follows. We present the main requirements that guided the ontology development in Section 2. We present a case study from a Huntington's disease investigation for illustrating how the ontologies can be used (in Section 3). We present the ontologies in Section 4. We go on to present the tools we developed around them, and competency queries that can be answered using Workflow Research Objects[2] (in Section 5). We present and compare related work with ours in Section 6. Finally, we present our conclusions and future work in Section 7. The resources used in the paper are available online,[3] and the ontologies are documented online [10].

2. Requirements

Our previous work [4] has identified a need to preserve more than just the workflow specifications in order to preserve their understandability, reusability and reproducibility. Related literature on supporting preservation of software [11,12] and best practice recommendations on supporting scientific reproducibility and computing [13–15] has further confirmed the need to preserve software, data and methods in aggregate. We present 5 requirements in detail that serve to establish the type of data and metadata that we need to support workflow preservation.

R_1 *Example data inputs should be provided.* Of the 92 workflows analysed in [4], 15% of them could no longer be run because they were not accompanied with any data examples. Even when inputs were textually described, it was difficult to establish input data values to be used for their execution. Without input data, both experiment reproducibility and the ability to understand the function the workflow is inhibited.

1 http://www.kegg.jp.
2 Note that in this paper we use the terms Workflow Research Object and Research Object interchangeably.
3 http://purl.org/net/jwsRO.

Please cite this article in press as: K. Belhajjame, et al., Using a suite of ontologies for preserving workflow-centric research objects, Web Semantics: Science, Services and Agents on the World Wide Web (2015), http://dx.doi.org/10.1016/j.websem.2015.01.003