Chapter 20
Workflow Discovery: Requirements from E–Science and a Graph–Based Solution

Antoon Goderis
University of Manchester, UK

Peter Li
University of Manchester, UK

Carole Goble
University of Manchester, UK

ABSTRACT
Much has been written on the promise of Web service discovery and (semi-) automated composition. In this discussion, the value to practitioners of discovering and reusing existing service compositions, captured in workflows, is mostly ignored. We present the case for workflows and workflow discovery in science and develop one discovery solution. Through a survey with 21 scientists and developers from the myGrid/Taverna workflow environment, workflow discovery requirements are elicited. Through a user experiment with 13 scientists, an attempt is made to build a benchmark for workflow ranking. Through the design and implementation of a workflow discovery tool, a mechanism for ranking workflow fragments is provided based on graph sub-isomorphism detection. The tool evaluation, drawing on a corpus of 89 public workflows and the results of the user experiment, finds that, for a simple showcase, the average human ranking can largely be reproduced.
INTRODUCTION

Problem Statement

As more scientific resources become available on the World Wide Web, scientists increasingly rely on Web and Grid services for performing in silico (i.e., computerised) experiments. Bioinformatics, for example, has seen a spectacular rise in the availability of distributed services—the myGrid/Taverna workbench offers access to over 3000 of these. A popular example of a bioinformatics Web service is BLAST (Basic Local Alignment Search Tool), a service for finding regions of genome sequence similarity.

Distributed service composition is difficult, be it manual or automatic. In this light, and to promote cross-disciplinary scientific collaborations, research councils worldwide are building a supporting infrastructure under the banner of e-Science. Exemplar initiatives include the Open Middleware Infrastructure Institute in the United Kingdom, D-Grid in Germany, and the Kepler project in the USA.

Workflow technology has been widely adopted in e-Science as the mechanism for orchestrating both distributed and local resources from within one environment. It potentially allows the e-Scientist to describe and enact her experimental processes in a structured, repeatable and verifiable way. Figure 1a displays a simple bioinformatics workflow loaded up in the myGrid/Taverna workbench (Oinn, Greenwood, Addis, Alpdemir, Ferris, Glover, Goble, Goderis, Hull, Marvin, Li, Lord, Pocock, Senger, Stevens, Wipat, & Wroe, 2006) on the left hand side, while a list of available services and workflows is shown on the right.

With the adoption of workflow environments by scientists comes a boom in the number of freely available scientific workflows. This motivates the problem of workflow discovery. Much has been written on the promise of Web service discovery and (semi-)automated composition. In this discussion, the value to practitioners of discovering and reusing existing service compositions, captured in workflows, is mostly ignored. Scientists have a need for finding workflows, either to draw from previous work investigating hypotheses similar to theirs or to discover fixes for broken workflows.

Article Contribution and Structure

Based on a case study from bioinformatics, the article (i) motivates the use of workflows in science, (ii) gathers requirements for workflow discovery, (iii) attempts building a benchmark to evaluate workflow discovery, (iii) develops a discovery technique and (iv) evaluates the technique. We expand on the work described in Goderis, Li, and Goble (2006) by presenting the case for workflows in bioinformatics, the role of the BPEL language, workflow discovery across workflow systems, and by covering related work in more detail.

In the next section we present the case for workflows in science. The following section contrasts Web service reuse with workflow reuse. Next, we consider why BPEL is not the workflow language of choice in e-science and present open issues for workflow discovery across systems. The subsequent section presents workflow discovery use cases, leading to workflow discovery requirements collected from 21 scientists and developers from the myGrid/Taverna workflow environment. Through a user experiment with 13 scientists, the next section attempts to build a benchmark for evaluating workflow rankings, with mixed results. Then, through the design and implementation of a workflow discovery tool, a mechanism for ranking workflow fragments is provided based on graph sub-isomorphism detection. The tool evaluation of the following section draws on a corpus of 89 public workflows to combine the results of the next two sections. It finds that, for a simple showcase, the average human ranking can largely be reproduced. We relate our work to the literature in the following section and the final section concludes.