Computers & Geosciences 79 (2015) 58–68

Contents lists available at ScienceDirect

Computers & Geosciences

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

An agent-based approach for capturing and linking provenance in geoscience workflows

Tom Narock [a,*], Victoria Yoon [b]

[a] Department of Information Technology, Marymount University, Arlington, VA, USA
[b] Department of Information Systems, Virginia Commonwealth University, Richmond, VA, USA

ARTICLE INFO

Article history:
Received 14 August 2014
Received in revised form
22 January 2015
Accepted 9 March 2015
Available online 11 March 2015

Keywords:
Multi-agent system
Provenance
Semantic web
Workflows
Web services

ABSTRACT

Provenance is becoming increasingly important as web services and computational workflows enable new methods by which work is conducted. Yet, there exist sets of questions that cannot be addressed by current provenance capture systems. We address these challenges by leveraging a service provenance ontology that captures execution details of workflow constituent web services. The ontology is used in conjunction with a multi-agent system to automate provenance aggregation and collation. The use of a multi-agent system eliminates the need to modify service interfaces, as was done in previous research. Simulation experiments are used to evaluate multiple agent topologies and identify an efficient and scalable system that scales to large numbers of workflows.

1. Introduction

Recent advances in data publication (e.g. Linked Open Data) have made data discovery significantly easier. Yet, increasing data volumes within many domains make it impractical to download data for local large-scale analysis. Within the geosciences, web services remain a viable solution. Hypotheses are tested through these online tools that combine and mine pools of data (Goble and De Roure, 2009). This evolution in the way that research is conducted is referred to as e-Science (Gray, 2009) with service oriented architectures serving as the common distributed technology (Miles et al., 2005). Within e-Science, the underlying web services allow an increasing volume of analysis to take place and have transformed how scientific research is performed (Miles et al., 2005).

The procedure of how services are combined to perform an experiment can be encoded as a workflow (Miles et al., 2005). A workflow is formally defined as

> "a precise description of a scientific procedure – a multi-step process to coordinate multiple tasks" where "each task represents the execution of a computational process, such as running a program, submitting a query to a database, submitting a job to a compute cloud or grid, or invoking a service over the Web to use a remote resource. Data output from one task is consumed by subsequent tasks according to a predefined graph topology that 'orchestrates' the flow of data." (Romano, 2008)

Several computational tools, such as Taverna Workflow Workbench (Missier et al., 2010) and Kepler (McPhillips et al., 2009), assist in the creation and capturing of workflows so that an experiment can be reviewed, validated, and adapted to reproduce its results. Workflow frameworks are geared toward easy composition of scientific experiments (Barga et al., 2010), which includes allocating and scheduling resources, orchestrating and monitoring the execution, and collecting provenance.

Unfortunately, online analysis is placing more and more "distance" between scientist, data, and analysis tools (Fox, 2012). As depicted in Fig. 1, a workflow may consist of web services spanning geographical and organization boundaries limiting access to the internal execution details of the constituent services. This physical distance leads to a "conceptual distance" that makes it difficult for a scientist to properly interpret the output of a web service. Thus, today's challenges are ensuring that every member of the scientific community can accurately interpret scientific experiments (Zhao et al., 2011).

Currently, recorded provenance is insufficient to facilitate user understanding of a workflow (Chapman and Jagadish, 2010). Web services are treated as "black-boxes," and provenance capturing systems only store information to describe which datasets were used and which web services were run (Chapman and Jagadish, 2010). This coarse-grained information describes what happened but makes it impossible for humans to understand *how* and *why*

* Corresponding author.
E-mail address: tnarock@marymount.edu (T. Narock).

http://dx.doi.org/10.1016/j.cageo.2015.03.004
0098-3004/

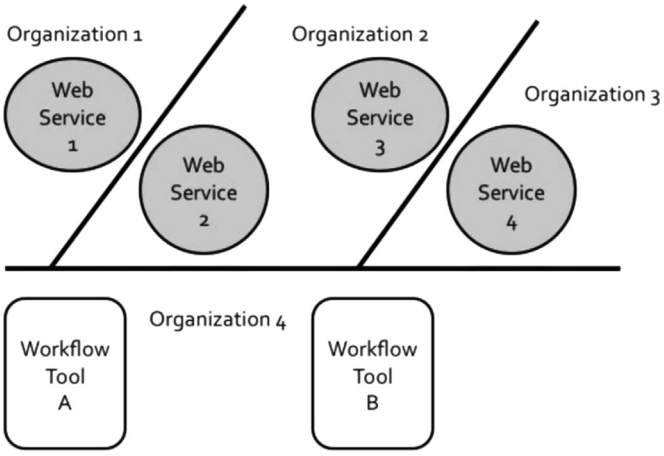

Fig. 1. The web services needed to construct a workflow are often distrubted across organizational boundaries.

the data was manipulated throughout the workflow (Chapman and Jagadish, 2010). Further, such coarse-grained approximations are rarely true and often misleading in understanding a workflow's dependencies and execution (Amsterdamer et al., 2012). Misunderstandings and incorrect conclusions could be drawn from not knowing how an input led to an output and what assumptions and algorithms were applied along the way, which could potentially have profound negative impacts on scientific decision-making (Fox, 2012).

Two parallel but complimentary approaches have been taken to address the lack of provenance in scientific workflows. Chapman and Jagadish (2010) have used relational databases to capture provenance from workflow tools and constituent services. This approach has the advantage of capturing service execution details, thus enabling better understanding of workflows. On the other hand, Zhao et al. (2011) have exposed workflow provenance as Linked Open Data (Bizer et al., 2009). Their approach uses semantics to link workflow provenance to domain specific information as many provenance questions must be answered using both domain-independent and domain specific data (Zednik et al., 2010; Ding et al., 2011). Both approaches have enabled us to better understand workflow results, but they have limitations in their generality. The approach of Zhao et al. (2011) lacks any provenance from the constituent workflow services. The Linked Data applies only to the workflow tool (e.g. the aforementioned Taverna and Kepler tools) and captures the services executed, the inputs and outputs of those services, the time of execution, and any errors that occurred. The internal execution details of the services themselves are excluded. On the other hand, Chapman and Jagadish's (2010) approach lacks the semantics of Zhao et al.'s (2011) work. Further, Chapman and Jagadish's method addresses simultaneous users and concurrently running workflows through the use of unique identifiers. Service interfaces must accommodate a unique identifier as input from the workflow tool. The services and workflow tool then use this identifier when writing to a provenance database. By doing so, web service and workflow tool provenance are linked, allowing subsequent users easy access to complete provenance. The complete provenance trail that this approach offers is vital. Yet, there are questions about the scalability of having to augment each service's interface to be compatible with this approach. We believe, like Chapman and Jagadish (2010) that complete provenance requires information from both the workflow tool and the executing services. Like Zhao et al. (2011), we also believe that such provenance should have a semantic representation. Thus, we propose a framework that merges the complete provenance capture of Chapman and Jagadish (2010) with the semantic approach of Zhao et al. (2011). We do so in a way that requires no changes to existing services. Yet, once the unique identifier is removed there exists no straightforward way to link web service provenance to workflow tool provenance. We address this challenge through the integration of a provenance ontology and a multi-agent system.

2. Related work

2.1. Provenance and workflows

Many provenance questions must be answered using both domain-independent and domain specific provenance data (Zednik et al., 2010; Ding et al., 2011). The Inference Web project was one of the first attempts to combine both types of data and expose them as Linked Open Data (Ding et al., 2011). Inference Web provides explanation of tasks such that users (human or software) can see what was done and how it was done (McGuinness and Pinheiro da Silva, 2004).

The realization that domain-independent and domain specific information were required has since led to three strands of research in regard to provenance and workflows: (1) generating provenance from scientific workflows, (2) annotating provenance metadata with domain specific ontologies, and (3) exposing provenance information in the form of Linked Open Data. These three strands have largely been pursued independently (Zhao et al., 2011). Only recently (Zhao et al., 2011) has work begun on integrating these strands and providing a comprehensive framework. Yet, the current integration of domain semantics, provenance semantics, and Linked Open Data does not include the execution provenance of the constituent web services. This means that questions regarding replacement web services (e.g. same algorithms and assumptions), limitations, and processing steps, cannot be answered by most provenance capture systems. Moreover, the systems that are addressing this issue are syntactic or not at the level of granularity required.

Previous research has led to specific approaches that allow for the submission of execution provenance from the constituent services of a workflow. The Provenance Aware Service Oriented Architecture (PASOA, (Miles et al., 2005)) is a provenance infrastructure based on SOAP web services. This approach uses the SOAP message exchange to capture process details; yet, PASOA is limited to SOAP-based web services and excludes step-by-step process recording (Simmhan et al., 2005). Existing workflow systems, such as Kepler, are extending their systems to accept provenance input from web services via a simple syntactic API (Altintas et al., 2006). Yet, satisfying the diverse needs of the multifarious scientific community in one API is a formidable task (Simmhan et al., 2005). While this is seen as a necessary first step, long-term success is likely to depend on submitting semantic representations of service execution details (Simmhan et al., 2005). The use of ontologies can allow automated provenance verification and enable richer queries to be answered (Simmhan et al., 2005; Ding et al., 2011). Such is the approach taken in this work.

Most recent is the work of Chapman and Jagadish (2010) and Zhao et al. (2011). Despite addressing provenance of service execution and Linked Open Data, respectively, these works suffer from the aforementioned limitations.

2.2. Software agents

Russelland Norvig (1995, p. 31) define an agent as “anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors.” Software agents are then pieces of software designed to interact with an