Conservation of Scientific Workflow Infrastructures Using Semantics

March 2014

Idafen Santana-Pérez

isantana@fi.upm.es
Ontology Engineering Group
Universidad Politécnica de Madrid
Madrid, Spain
• PhD Student AI
  • UPM
  • Since 2011
• MSc Complex Systems
  • UPM
  • 2009-2010
• Computer Science degree
  • ULPGC
  • 2003-2008
Ontology Engineering Group

Directors: A. Gómez-Pérez, O. Corcho

Research Group (~30 people)
- 2 Full Professors
- 5 Associate Professors
- 3 Assistant Professors
- 6 Postdocs
- 12 PhD Students
- 2 MSc Students

Technical support (3 people)
- 2 software engineers
- 1 system administrator

Management (3 people)
- 1 Project Manager
- 2 administrative assistants

Several visitors

50+ Past Collaborators
20+ Past Visitors

http://www.oeg-upm.net/
https://github.com/oeg-upm
@oeg-upm
Data Integration

Semantic e-Science

Ontological Engineering

(Social) Semantic Web and Linked Data

Natural Language Processing

1995

2000

2004

2008

About me
• Ontology Engineering Group

» Asunción Gómez-Pérez

» Oscar Corcho

» María S. Pérez-Hernández

» Daniel Garijo
Introduction

• Experiments in empirical science
  • Primary component of the scientific method
  • Main method for validating a hypothesis
  • **Repeatable procedure**

• Scientific publications
  • Announce a result
  • **Convince readers that the result is correct**

• Computational science
  • In silico science
  • Computational scientific workflow: “a precise, **executable description** of a scientific procedure” [De Roure, 2011]

  • “**Reproducibility in principle underpins the scientific method**” [Goble, 2012]

  • “**Its about capturing, preserving, reusing and curating**” [Goble 2012]
• The replication standard holds that sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party can replicate the results without any additional information from the author [Gary King, 1995]
Goals

• Models to describe an infrastructure in order to reproduce it in the future
  • Models for representing:
    • Execution requirements of a workflow
    • Infrastructure dependencies

• Framework to provide:
  • Means for populating these models
    • Using collaborative annotation (e.g. scientists and IT staff)
    • Automated or semi automated analysis of traces and profiles to extract the requirements

• Algorithms that are able to generate an equivalent infrastructure specification
  • Based on the annotations
  • Infrastructure providers resource availability
  • User policies
Goals

- WICUS Framework overview
Recent Work

- WICUS ontology network
- Workflow annotations
- Catalogues annotations
- Infrastructure Specification Algorithm
Recent Progress

- WICUS ontology network
  - [http://purl.org/net/wicus](http://purl.org/net/wicus)
  - 5 ontologies
  - WICUS Software Stack ontology
  - WICUS Hardware Specs ontology
  - WICUS Scientific Virtual Appliance ontology
  - WICUS Workflow Execution Requirements ontology
  - WICUS Ontology: links the previous ontologies

- Semantics technologies are a mature, standard and flexible way for defining information.
Recent Progress

- WICUS Software Stack ontology
  - [http://purl.org/net/wicus-stack](http://purl.org/net/wicus-stack)
Recent Progress

- WICUS Hardware Specs ontology
  - http://purl.org/net/wicus-hwspecs
Recent Progress

- WICUS Scientific Virtual Appliance ontology
  - http://purl.org/net/wicus-sva
Recent Progress

- **WICUS Workflow Execution Requirements**
  - [http://purl.org/net/wicus-reqs](http://purl.org/net/wicus-reqs)
Recent Progress

• WICUS ontology network
  • http://purl.org/net/wicus
Recent Progress

- Workflow Infrastructure Annotations
  - Annotated 4 Workflow Templates from WINGS
    - 56 possible implementations
  - Text Analytics
  - Computer forensics
    - Script analysis
  - Taverna Workflows (partially)
Recent Progress

- Workflow Infrastructure Annotations
- Workflow Annotation Tool

**WFUS**

*workflow* | Model

**MODEL_WF**

- http://purl.org/net/wcus-reqlWorkflow
- CORRELATIONSCORE_ABS_WF
- CHISQUARED_CONC_WF
- MUTUALINFORMATION_CONC_WF
- INFORMATION_GAIN_CONC_WF
- FEATURES_SELECTION_CONC_WF

**MUTUALINFORMATION_CONC_WF**

- http://purl.org/net/wcus-reqlConcreteWorkflow
- MUTUALINFORMATION_SOFT_REQ
- MUTUALINFORMATION_SOFT_STACK
- MUTUALINFORMATION_SOFT_COMP
- MUTUALINFORMATION_CONF_INFO
- MUTUALINFORMATION_CONF_PAR
- MUTUALINFORMATION_CONF_BIN

**Steps**

- MUTUALINFORMATION_WF_STEP
  - http://www.isi.edu/ac/TextAnalytics/library.owl#MutualInformationClass
Recent Progress

• Catalogue Annotations
  • Software Components Catalogue
    • Catalogue including the Software Components involved in the workflow execution
    • 27 Software Components

• SVA Catalogue
  • Catalogue describing the available Virtual Appliances
  • Includes the Software Components already deployed in the appliance
  • 2 Virtual Appliances from AWS
Recent Progress

- Infrastructure Specification Algorithm
  - Goal: obtain a specification defining what VMs need to be created, what software components must be deployed and their configuration.

```xml
<deployment>
  <node name="Ubuntu_12_04_SVA">
    <provider name="amazon">
      <image>ami-967edc6f</image>
      <instance-type>t1.micro</instance-type>
    </provider>
    <plugin script="JAVA_script.sh">
      <param name="JAVA_HOME">/usr/lib/jvm</param>
    </plugin>
    <plugin script="deploy.jar.sh">
      <param name="DEST_PATH">DIR:</param>
      <param name="JAR_NAME">StopWords.jar</param>
    </plugin>
    <plugin script="deploy.jar.sh">
      <param name="DEST_PATH">DIR:</param>
      <param name="JAR_NAME">RemoveStopWords.jar</param>
    </plugin>
  </node>
</deployment>
```
Recent Progress

- Infrastructure Specification Algorithm
- Jena and SPARQL Queries

GET WF REQUIREMENTS

GET <REQ, STACKS>

GET <REQ, D-GRAPH>

GET AVAILABLE SVA

GET <SVA, STACKS>

CALCULATE REQ-SVA COMPATIBILITY

GET MAX COMPATIBLE REQ-SVA

CLEAN REQ D-GRAPH
Recent Progress

WORKFLOW

- REQ1
- REQ2
- REQ3
Recent Progress

Conservation of Scientific Workflow Infrastructures by Using Semantics
Recent Progress

Conservation of Scientific Workflow Infrastructures by Using Semantics
Conservation of Scientific Workflow Infrastructures by Using Semantics
Recent Progress

Conservation of Scientific Workflow Infrastructures by Using Semantics
Recent Progress

Conservation of Scientific Workflow Infrastructures by Using Semantics
Recent Progress

Conservation of Scientific Workflow Infrastructures by Using Semantics
Recent Progress

Conservation of Scientific Workflow Infrastructures by Using Semantics
Recent Progress

Conservation of Scientific Workflow Infrastructures by Using Semantics
Conservation of Scientific Workflow Infrastructures by Using Semantics

Recent Progress

S13

S14

SVA1

REQ1

S1

S2

S3

S4

S5

S6

S7

S8

S9

S10

S11

S12

S13

SVA3

REQ2

SVA3

REQ3
Recent Progress

Conservation of Scientific Workflow Infrastructures by Using Semantics

(xml code)

<deployment>
    <node name="Ubuntu_12_04_SVA">
        <provider name="amazon">
            <image>ami-967edcfr</image>
            <instance-type>t1.micro</instance-type>
        </provider>
        <plugin script="JAVA_script.sh">
            <param name="JAVA_HOME">/usr/lib/jvm/</param>
        </plugin>
        <plugin script="deploy_jar.sh">
            <param name="DEST_PATH">DIR:</param>
            <param name="JAR_NAME">StopWords.jar</param>
        </plugin>
        <plugin script="deploy_jar.sh">
            <param name="DEST_PATH">DIR:</param>
            <param name="JAR_NAME">RemoveStopWords.jar</param>
        </plugin>
    </node>
</deployment>
Conservation of Scientific Workflow Infrastructures by Using Semantics
Recent Progress

Conservation of Scientific Workflow Infrastructures by Using Semantics
Next Steps

- Implement the enactment system
- New release of the WICUS ontology network
- Implement new version of algorithm
  - Implement versioning
  - Considers graph structure for compatibility, rather than just set intersection.
- Annotate workflows from other platforms, such as Taverna, Kepler, etc.
- Evaluate

Conservation of Scientific Workflow Infrastructures by Using Semantics
Stay at ISI

- Information Sciences Institute
- Feb 28 – May 31
- Pegasus
- Wrangler and PRECEIP deployment tools
- Evaluation
- Study platform requirements
- Capturing runtime, memory usage, and disk usage
- Paper
Questions

Conservation of Scientific Workflow Infrastructures by Using Semantics

@idafensp