Impact of Software Environment on Replicability of Biomedical Workflows

Tomasz Miksa

Vienna University of Technology, Austria
miksa@ifs.tuwien.ac.at
Replicability

- Can I reuse her workflow?
- Do the results match?
- Have I done it correctly?
Replicability

- Current studies show very low reproducibility in:
  - medicine
  - economy
  - computer science
- Reproducibility requires:
  - well documented research workflows
  - precise information on the experiment's environment

3/17  http://dx.doi.org/10.1016/j.jbi.2016.10.011
VFramework

Framework

Original environment

Run static analysis

Run dynamic analysis

Define validation metrics

Redeployment environment

Verify environment

Validate workflow

Context Model

http://dx.doi.org/10.1016/j.jbi.2016.10.011
Context Model

- OWL ontology
- Modular architecture
  - Domain Independent Ontology
  - Domain Specific Ontologies
- Process preservation
Context Model

Validation metrics

Workflow definition

File formats

Software dependencies

Provenance

6/17 http://dx.doi.org/10.1016/j.jbi.2016.10.011
Validation report for the WeatherExample

Evaluation result: There are 2 not fulfilled metrics. Please see tables below for details.

Comparison performed using following workflow execution traces

Original Workflow
ID: 37b4d2f6-e71c-4b67-b7b3-1788ee82977
Timestamp: 2015-10-14 18:58:06.475

Compared Workflow
ID: ede04b87-5f58-4a89-b0c3-e179957cbad0
Timestamp: 2015-11-13 13:59:56.443

Table 1: Overview of requirements

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Description</th>
<th>Is Fulfilled</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>The inputs to the workflow are the same</td>
<td>true</td>
</tr>
<tr>
<td>R2</td>
<td>The outputs of the workflow are the same</td>
<td>false</td>
</tr>
<tr>
<td>R3</td>
<td>The workflow step ExtractTemperature must have identical outputs</td>
<td>true</td>
</tr>
<tr>
<td>R4</td>
<td>The workflow step GetWeatherData must have identical outputs</td>
<td>true</td>
</tr>
<tr>
<td>R5</td>
<td>The workflow step MakeDecision must have identical outputs</td>
<td>true</td>
</tr>
<tr>
<td>R6</td>
<td>The workflow step ExtractWeatherType must have identical outputs</td>
<td>true</td>
</tr>
<tr>
<td>R7</td>
<td>The workflow step VisualiseTemperature must have identical outputs</td>
<td>false</td>
</tr>
<tr>
<td>R8</td>
<td>Execution duration of each of the workflow steps shall be similar</td>
<td>true</td>
</tr>
</tbody>
</table>

Table 2: List of requirements and metrics that failed.

<table>
<thead>
<tr>
<th>Req</th>
<th>Sub-req</th>
<th>Sub-requirement description</th>
<th>Measurement point</th>
<th>Metric</th>
<th>Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>R2</td>
<td>R7.1</td>
<td>The output plot of workflow step VisualiseTemperature must be identical</td>
<td>plot</td>
<td>ImageFingerPrintEquality</td>
<td>false</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ImageResolutionEquality</td>
<td>true</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AbsoluteErrorCount</td>
<td>false</td>
</tr>
<tr>
<td>R7</td>
<td>R7.1</td>
<td>The output plot of workflow step VisualiseTemperature must be identical</td>
<td>plot</td>
<td>ImageFingerPrintEquality</td>
<td>false</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ImageResolutionEquality</td>
<td>true</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AbsoluteErrorCount</td>
<td>false</td>
</tr>
</tbody>
</table>
Use Cases

- Characteristics of Taverna workflows on myExperiment
  - 50% use Beanshells
  - 30% use WSDL web services
  - 15% local tool invocations

- Access to the original environment was necessary

- 5 Taverna workflows from 3 domains
  - 3 biomedical workflows (Leiden University, NL)
  - 1 music classification workflow (TU Wien)
  - 1 sensor data analysis workflow (LNEC, Portugal)
Use Cases

- Workflows required
  - specific dependencies in the environment
  - external services
- Re-executions on different
  - version of operating system
  - distribution
  - system architecture
Results - dependencies detection

- All dependencies were identified
  - R dependencies
  - Ruby dependencies
  - Java libraries
  - local tools (e.g. pdflatex)
Results – external services

- All external communications detected (web services)
- Limited validation for external services (Rserve scripts)
  - global variables
  - steps with no outputs
  - only final result of workflow computation is validated
    • black box testing
Results – instance data

- Dynamic analysis detected
  - data created through shell calls
    - such files are not a part of provenance traces in Taverna
  - ‘real’ workflow outputs
    - Taverna traces can contain paths only, but not the content
    - Taverna workflows can create files not linked to any output

============= running time total: 0.710367914 =============
No of statements in a file 29590 I,
[2015-11-30T15:44:15.159824 #8787] INFO -- : 0.708272803
Results—validation metrics

- detected changes on different stages of workflow processing
  - metadata included in data
    - generation timestamp
  - file format comparison improved the results
    - ZIP archives

\begin{table}
\centering
\begin{tabular}{rrrrrrr}
\hline
& Min & 1Q & Mediana & Media & 3Q & Max \\ 
\hline Residuos (mm) & -2.42 & -1.23 & -0.52 & -0.00 & 0.45 & 5.52 \\ 
\hline
\end{tabular}
\caption{Residuos do modelo}
\end{table}
Recommendations

- Analyse dependencies and evade shell calls
  - e.g. use scripting mechanisms provided by the workflow engine
- Write code that runs on all platforms
  - e.g. do not encode specific paths
- Publish experiment setup and context
  - e.g. exact versions of tools used
- Publish validation data
  - e.g. provenance but also other files created during execution
- Test the replicability on your own
  - e.g. try rerunning your experiment in a clean virtual machine

http://dx.doi.org/10.1016/j.jbi.2016.10.011
Conclusions

I have repeated her experiment in the same way!
I got the same results!
I can reuse any part of it!

VFramework

Original experiment

Re-executed experiment

17/17  http://dx.doi.org/10.1016/j.jbi.2016.10.011