PubFlow
a scientific data publication framework for marine science

Peer Brauer, Prof. Dr. Wilhelm Hasselbring
Software Engineering Group
Kiel University
München, 23.10.2013

funded by DFG
www.pubflow.de
Motivation
Motivation
Motivation
Agenda

PubFlow Framework
Evaluation Scenario
Conclusion
What is PubFlow about?

Creating a scientific workflow environment for data publication
Introducing role based working models to the domain of data management
Increasing the degree of automation in data management
Features

Build upon proven workflow technology

Build in support for BPMN, BPEL
Extensible by other workflow engines
Designed for high throughput
Features
Provenance Awareness

Automatically capturing of provenance data
Integrated W3C Prov-O compliant provenance archive
Workflow based provenance browser
# Features

Provenance Awareness

<table>
<thead>
<tr>
<th></th>
<th>Feature Name</th>
<th>Type</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>Informed Rural Passenger Information Infrastructure</td>
<td>Application</td>
<td>PROV-O</td>
</tr>
<tr>
<td>33</td>
<td>PubFlow Provenance Archive</td>
<td>Application / Framework / API</td>
<td>PROV-O, PROV-XML</td>
</tr>
<tr>
<td>34</td>
<td>PROV Python library</td>
<td>Framework / API</td>
<td>PROV-N, PROV-JSON</td>
</tr>
<tr>
<td>35</td>
<td>csv2rdf4lod-automation</td>
<td>Application</td>
<td>PROV-O</td>
</tr>
<tr>
<td>36</td>
<td>recoprov</td>
<td>Application</td>
<td>PROV-O, PROV-N</td>
</tr>
<tr>
<td>37</td>
<td>DataFAQs</td>
<td>Application</td>
<td>PROV-O</td>
</tr>
<tr>
<td>38</td>
<td>provx2o</td>
<td>Application</td>
<td>PROV-O, PROV-XML</td>
</tr>
<tr>
<td>39</td>
<td>Hedgehog</td>
<td>Application</td>
<td>PROV-XML</td>
</tr>
<tr>
<td>40</td>
<td>QuerioCity research prototype</td>
<td>Application / Framework / API Service</td>
<td>PROV-O</td>
</tr>
<tr>
<td>41</td>
<td>Tinga Provenance Service</td>
<td>Service</td>
<td>PROV-O, PROV-JSON</td>
</tr>
<tr>
<td>42</td>
<td>Human Computation ontology</td>
<td>Vocabulary Extension</td>
<td>PROV-O</td>
</tr>
<tr>
<td>43</td>
<td>tavernaprov</td>
<td>Vocabulary Extension</td>
<td>PROV-O</td>
</tr>
<tr>
<td>44</td>
<td>The Open Provenance Model for Workflows (OPMW)</td>
<td>Vocabulary Extension</td>
<td>PROV-O</td>
</tr>
<tr>
<td>45</td>
<td></td>
<td></td>
<td>PROV-O</td>
</tr>
</tbody>
</table>
Features
Provenance Awareness

Apache ODE

Kieker.workflow.Analyzer

Kieker.workflow.Visualizer

Kieker.monitoring.Writer

Kieker.workflow.monitoring.ode.Probe

Kieker.WorkflowMonitor Controller

Kieker.workflow.Processor

Kieker.workflow.Space

KiekerMonitoring-Log / Stream

Kieker.workflow.Visualizer

Kieker.workflow.Analyzer
Features

Graphical Workflow Editor

Supports graphical DSLs
Data managers can easily define own workflows
Workflows can be transformed to selected target execution environment
Custom User Interfaces

- Build upon proven technology
- Modular, flexible Architecture
- Various Workflow Engines
- Variable Datastores

Integrations:
- Workflow Engines
  - BPMN 2.0 Engine (JBPM)
  - BPEL 2.0 Engine (ODE /Jetty)
- WF Computational Plugins
- Presentation
  - Ticketsystem (Jira)
  - Portal (Liferay)
- User-Frontend
  - Frontend Endpoints
- Communication
  - User-Frontend
- Administration-frontend
  - Custom User Interfaces
  - Modular, flexible Architecture
  - BPEL 2.0 Engine (ODE /Jetty)

Engines:
- BPMN 2.0 Engine (JBPM)
- BPEL 2.0 Engine (ODE /Jetty)

Plugins:
- BPMN 2.0 Workflow Broker
- BPEL 2.0 Engine (JBPM)

Custom User Interfaces:
- Custom User Interfaces
- Modular, flexible Architecture
- Various Workflow Engines
- Variable Datastores

Integrations:
- Workflow Engines
  - BPMN 2.0 Engine (JBPM)
  - BPEL 2.0 Engine (ODE /Jetty)
- WF Computational Plugins
- Presentation
  - Ticketsystem (Jira)
  - Portal (Liferay)
- User-Frontend
  - Frontend Endpoints
- Communication
  - User-Frontend
- Administration-frontend
  - Custom User Interfaces
  - Modular, flexible Architecture
  - BPEL 2.0 Engine (ODE /Jetty)
Section 2

The Evaluation Scenario
Evaluation Scenario
Evaluation Scenario

- Observation
- Simulation & Analysis
- Visualization
- Research Paper
- Review
- Institutional Repository
- Data Curation
- Data Center
- Publisher
Evaluation Scenario

Diagram showing the evaluation scenario involving data curation, data center, and publisher, with the institutional repository as the central point of interaction.
Evaluation Scenario

PubFlow – OCN_To_Pangaea

Scheduler

PubFlow

Load from DB

Map to Pangaea

Open Tasks?

To Pangaea

Anwender

Edit Dataset

Exception-handling
Evaluation Scenario
Conclusion

PubFlow increases the degree of automation in the data publication process. It is build upon proven workflow technology. Brings the division of work to data management. Collects provenance information.