CAPS

Capturing and Managing Provenance Information in Scientific Workflows
AGENDA

‣ Motivation

‣ Research Questions

‣ CAPS

‣ Evaluation
Motivation
MOTIVATION

Provenance

How was the scientific data I use for my research created?
<table>
<thead>
<tr>
<th>Coordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position (RA):</td>
</tr>
<tr>
<td>Position (Dec):</td>
</tr>
<tr>
<td>Field of view:</td>
</tr>
<tr>
<td>Orientation:</td>
</tr>
</tbody>
</table>
The oddly shaped Pencil Nebula (NGC 2736) is pictured in this image from ESO's La Silla Observatory in Chile. This nebula is a small part of a huge remnant left over after a supernova explosion that took place about 11,000 years ago. The image was produced by the Wide Field Imager on the MPG/ESO 2.2-metre telescope at ESO's La Silla Observatory in Chile.

Credit: ESO
License: Creative Commons Attribution 3.0 Unported license
The oddly shaped Pencil Nebula (NGC 2736) is pictured in this image from ESO's La Silla Observatory in Chile. This nebula is a small part of a huge remnant left over after a supernova explosion that took place about 11,000 years ago. The image was produced by the Wide Field Imager on the MPG/ESO 2.2-metre telescope at ESO's La Silla Observatory in Chile.

**Credit:** ESO

**License:** Creative Commons Attribution 3.0 Unported license
MOTIVATION

Definition Provenance

Provenance information describes how, when, where and by whom data was created.
MOTIVATION

W3C Prov-DM

Source: http://www.w3.org/TR/2013/NOTE-prov-primer-20130430/
MOTIVATION

Scientific Workflows

... automate recurrent tasks

... are widely used in data intensive domains

... orchestrate complex analysis or computations

Source: www.myexperiment.org/workflows/197.html
MOTIVATION

Scientific Workflows

- ... can have many different forms
- ... consist mainly of Activities, Gateways and Events

Quelle: http://docs.jboss.org/jbpm/v6.0.1/userguide/images/Chapter-1-Overview/Overview.png
MOTIVATION

Problemstatement

- Scientific workflow engines are often domain specific and developed by small teams with less resources > only really needed functionality is implemented

- also business products are widely used

❓ How to assure the capturing of provenance information in these systems?
Research Question
RESEARCH QUESTIONS

Aspect-Oriented Provenance Capturing

Data provenance as a cross-cutting concern

- Is it possible to use aspect-oriented programming techniques to acquire provenance information from a scientific software system?

- How to determine, which methods and modules have to be monitored?
RESEARCH QUESTIONS

Capturing Provenance Information in scientific workflows

‣ What are the demands for capturing provenance information in scientific workflows?

‣ Which informations have to be gathered from the systems?
RESEARCH QUESTIONS

Provenance Information Reconstruction

Integrating provenance information from different systems and reconstruction the provenance graph

› What are the demands for constructing a provenance graph out of monitoring informations received from different software systems?

› How can the provenance information be filtered out of the normal monitoring records?
RESEARCH QUESTIONS

Provenance Information Archival

Archiving the provenance information

› How to store the provenance information, so it can be accessed by scientist and data managers to validate their data?

› How to link the provenance information back to the processes and the data it evolved of?
CAPS

Integration in the Guest System

Application

GUI

Business Logic

3. Party Libraries

Workflow

Persistence

Filesystem / Database …

Network

CAPS

Weaver

Runtime Observer

Provenance Filter

Provenance Archive
CAPS

Integration in the Guest System

Analyze Application & Integrate Provenance Awareness
CAPS

Integration in the Guest System

Collect Provenance Information
Basic Workflow

- Deploy a Application
- Execute Application
- Integrate CAPS in Application
- Analyze Monitoring Output
- Archive Monitoring Output
- Create Provenance Graph
- Archive Provenance Graph
- Display Provenance Graph

Time

Peer Brauer
Software Engineering Group - Kiel University
24.03.2014
CAPS

Basic Workflow

Deploy a Application

Execute Application

Integrate CAPS in Application

Analyze Monitoring Output

Archive Monitoring Output

Create Provenance Graph

Archive Provenance Graph

Display Provenance Graph

User

CAPS

Time
Deploying an Application

- Upload an application to CAPS
- Choose an application profile
- Create a runtime configuration
- Create an application wrapper
Basic Workflow

- Deploy a Application
- Execute Application
- Integrate CAPS in Application
- Analyze Monitoring Output
- Archive Monitoring Output
- Create Provenance Graph
- Archive Provenance Graph
- Display Provenance Graph
CAPS

Analyzing and Filtering the Results

- Observation data analysis is build upon Kieker Webgui

- Data is processed by an integrated pipe and filter analysis framework

- Dedicated provenance filter reconstruct the provenance graph, workflow

- Additional information can be processed (Memory usage, CPU time, …)

- Own filters and analysis plugins can be added

Source: http://kieker-monitoring.net/features/webgui/attachment/kieker-webgui-02/
Basic Workflow

1. User Deploy a Application
2. CAPS Integrate CAPS in Application
3. User Execute Application
4. Analyze Monitoring Output
5. Archive Monitoring Output
6. Create Provenance Graph
7. Archive Provenance Graph
8. Display Provenance Graph

Peer Brauer
Software Engineering Group - Kiel University
24.03.2014
CAPS

Archiving the Provenance Data

- Provenance archive build upon Neo4j, EMF and Prov-DM
- API, Webservice and WebGUI for accessing the provenance information
- Export of the provenance data to data centers (i.e. Zenodo)
Evaluation
EVALUATION

PubFlow

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

Peer Brauer
Software Engineering Group - Kiel University
24.03.2014
EVALUATION

System Landscape

- OCN Editor
- „Start Smart“
- OCN Datenbank
- local Data Sources
- Oceanrep / Kielprints
- Sckripts, Excelsheets, …
- OSIS
- WDC Mare
- Oceanrep / Kielprints
EVALUATION

System Landscape

- OCN Editor
- OCN Datenbank
- "Start Smart"
- local Data Sources
- Oceanrep / Kielprints
- PubFlow
- OSIS
- WDC Mare
- Oceanrep / Kielprints
EVALUATION

PubFlow : Goals

- Development of a workflow based tool for data publication
- Integration of PubFlow in the existing systems
EVALUATION

PubFlow

Presentation
- Config.-Interface
- Workflow Browser*
- Website*
- Portal* (Liferay)
- Ticketsystem (Jira)

Integration
- Mail (JEE Mail API)
- JMS (Active MQ)
- Webservice (JaxWS)
- Frontend Endpoints
- Workflow Broker

User-Frontend
- Admin-Frontend

Communication
- Mail (JEE Mail API)
- JMS (Active MQ)
- Frontend Endpoints
- Workflow Broker

User-Frontend
- Config.-Interface
- Workflow Browser*
- Website*
- Portal* (Liferay)
- Ticketsystem (Jira)

Integration
- Mail (JEE Mail API)
- JMS (Active MQ)
- Frontend Endpoints
- Workflow Broker

Workflow
- BPMN 2.0 Engine (JBPM)
- BPEL 2.0 Engine (ODE/Jetty)
- WF Computational Plugins

Core Modules
- Security (Shiro)
- Scheduling (Quartz)
- Repositories
- Config.-Files

Security
- (Shiro)
- Scheduling
- (Quartz)
- Repositories
- (Common)

Repositories
- Common
- Workflow Definitions
- Config.-Files

Data
- HsqlDB
- Portal DB* (DB2)
- OceanRep* (Postgresql)
- OCN (Postgresql)

Databases
- HsqlDB
- Portal DB* (DB2)
- OceanRep* (Postgresql)
- OCN (Postgresql)
EVALUATION

PubFlow: The Workflow

- Observation
- Simulation & Analysis
- Visualization
- Research Paper
- Review
- Institutional Repository
- Data Curation
- Institutional Archive
- Data Center
- Publisher
EVALUATION

PubFlow: The Workflow

Observation -> Institutional Repository -> Data Curation

Simulation & Analysis -> Visualization

Research Paper -> Institutional Archive

Review

Data Center

Publisher

Institutional Archive

Data Center
EVALUATION

PubFlow: The Workflow

- Loading the Data
- Transformation
- Export the Data to the WDC
EVALUATION

PubFlow

Presentation

Config.-Interface
Workflow Browser* Website* Portal* (Liferay)

Integration

JMS (Active MQ) Webservice (JaxWS)

Frontend Endpoints

HTTP (Jetty)

Workflow Broker

BPMN 2.0 Engine (JBPM)

BPEL 2.0 Engine (ODE/Jetty)

WF Computational Plugins

Workflow

User-Frontend

User-Communication

User-Endpoints

Workflow

Msg.-Bus (Active MQ)

Msg.-Broker

BPMN 2.0 Engine (JBPM)

BPEL 2.0 Engine (ODE/Jetty)

Workflow Computational Plugins

User Forming

PubFlow Core Modules

Security (Shiro)

Scheduling (Quartz)

Repositories

Common

Data

HsqlDB

Portal DB* (DB2)

OceanRep* (Postgresql)

OCN (Postgresql)

Workflow Definitions

Config.-Files

PubFlow Core

Modules

Admin-Frontend

Peer Brauer

Software Engineering Group - Kiel University

24.03.2014
EVALUATION

PubFlow

Presentation
- Config.-Interface
- Workflow Browser
- Website
- Portal (Liferay)
- Ticketsystem (Jira)

Integration
- HsqlDB
- Portal DB (DB2)
- OceanRep (Postgresql)
- OCN (Postgresql)
- Workflow Definitions
- Config.-Files

- HTTP (Jetty)
- JMS (Active MQ)
- Webservice (JaxWS)

Workflow
- Workflow Broker
- BPEL 2.0 Engine (ODE/ Jetty)

- BPMN 2.0 Engine (JBPM)

Computational
- WF Computational Plugins

- Mail (JEE Mail API)
- Mail (Camel)

- Security (Shiro)
- Scheduling (Quartz)
- Repositories
- Common

Data

- User-Frontend
- Portal (Capelli)
- Website (Capelli)

Admin-Frontend
- Security
- Scheduling
- Repositories
- Config.-Files

Repositories

Common

WEB SERVICES

BPEL 2.0

BPMN 2.0

WF Computational Plugins

Portal DB

OCN

Peer Brauer
Software Engineering Group - Kiel University
24.03.2014
EVALUATION

PubFlow

Presentation
- Configuration Interface
- Workflow Browser
- Website
- Portal (Liferay)
- Ticket System (Jira)
- Workflow Broker
- JMS (Active MQ)
- Webservice (JaxWS)
- Frontend Endpoints

Integration
- Mail (JEE Mail API)
- Message Router (Camel)
- Security (Shiro)
- Scheduling (Quartz)
- Repositories
- Common

Data
- HSQLDB
- Portal DB (DB2)
- OceanRep (Postgresql)
- OCN (Postgresql)
- Workflow Definitions
- Configuration Files

Workflow
- BPMN 2.0 Engine (JBPM)
- BPEL Engine (ODM/Neteum)
- WF Computational Plugins

Repositories
- Common
- Configuration Files
- Workflow Definitions

Website
- Portal (Liferay)
- Portal DB (DB2)
- OceanRep (Postgresql)
- OCN (Postgresql)

CAPS
EVALUATION

PubFlow

Presentation
- Config.-Interface
- Workflow Browser* (Jetty)
- Website* (Liferay)
- Portal* (Liferay)
- Ticketsystem (Jira)
- Workflow Broker
- BPMN 2.0 Engine (JBPM)
- BPEL 2.0 Engine (ODE / Jetty)
- WF Computational Plugins

Integration
- Mail (JEE Mail API)
- JMS (Active MQ)
- Webservice (JaxWS)
- Frontend Endpoints
- BPMN 2.0 Engine (JBPM)
- BPEL 2.0 Engine (ODE / Jetty)
- WF Computational Plugins

User-Service
- User-Frontend
- Portal* (Liferay)
- Portal DB* (DB2)
- OCN (Postgresql)
- OCN-Database (Postgresql)
- OCN-Workflow <Activity>
- OCN-Dataset <Entity>
- OCN-Workflow <Activity>
- OCN-Dataset <Entity>
- PubFlow-System <Entity>
- DataManager <Entity [Person]>
- ActedOnBehavior
- Used
- WasGeneratedBy
- WasDerivedFrom
- WasAssociatedWith

Data
- HsqlDB
- OceanRep* (Postgresql)
- OCN (Postgresql)
- Portal DB* (DB2)
- Filesystem
- Config.-Files
- Workflow Definitions
- Repositories
- Common
- Scheduling (Quartz)
- Security (Shiro)

CAPS

Peer Brauer
Software Engineering Group - Kiel University
24.03.2014
CONCLUSION

- CAPS captures provenance information using AspectJ and workflow monitoring
- Uses proven technology like Kieker, Neo4J, Emf, GWT, …
- Archives provenance information in an integrated archive
- Makes provenance information accessible through different interfaces
Thanks for your attention!