FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Diplomarbeitspräsentation

Capturing and Visualizing Provenance Information

Masterstudium:

Software Engineering and Internet Computing

Fenghong Zhang

Technische Universität Wien
Institute of Information Systems Engineering (ISE)
Vienna University of Technology
BetreuerIn: Ao.Univ.Prof. Dipl.-Ing. Dr.techn.
Andreas Rauber

Introduction

In the perspective of modern science, a proper methodology for managing and monitoring scientific workflows plays an important role. Currently, we are missing such a tool, which can automatically catch the provenance data with little user effort and represent the provenance information into a human-readable and intractable structure.

Provenance captures all computational steps, its output, input, and environment. The goals of this thesis is to provide a tool for scientific workflow provenance documentation. The tool aims to supply a prototype to solve the following questions:

- How to automatically track, monitor and capture system provenance information at a wide range?
- How to transform the captured provenance data into provenance ontology?
- How to present the collected provenance data into a human-readable format with interaction possibilities?

Theoretical Basis

"Reproducibility has the potential to serve as a minimum standard for judging scientific claims when full independent replication of a study is not possible." [1]. The term scientific workflow(SWF) is the description of a process for accomplishing a scientific objective. The software products–Workflow Management System(SWfMS) is designed to help users with creating and executing the workflows. But they require a high learning curve and not implemented with general-purpose languages. The tool Noworkflow overcomes the problem but it can only apply to Python.

The provenance concept and the version control system are used in most of the SWfMS. The tracking of the provenance data is beginning from the creation and end up with the final results[2]. PROV-O[3] is the forthcoming data model and is provided by W3C with specifications as the first official standard, we introduce the PROV-O to describe our SWF provenance. Moreover, the version control system records the changes to the file over time. We will still using those concepts in our thesis.

Project Structure

The tool contains three components, the application console records scientific computational tasks and initializes the workspace for storing the results and the logfiles. The File Monitor is the monitoring tool for detecting the log file entrance. The Graph visualizer integrates the visualization of ontology and user interactions. Its provenance ontology creator is used for translating and converting the scientific workflows into the Extensible Markup Language (XML) format and generate the visualized graph accordingly. The structure of the project and its library usages can be shown as the Figure 1 below:

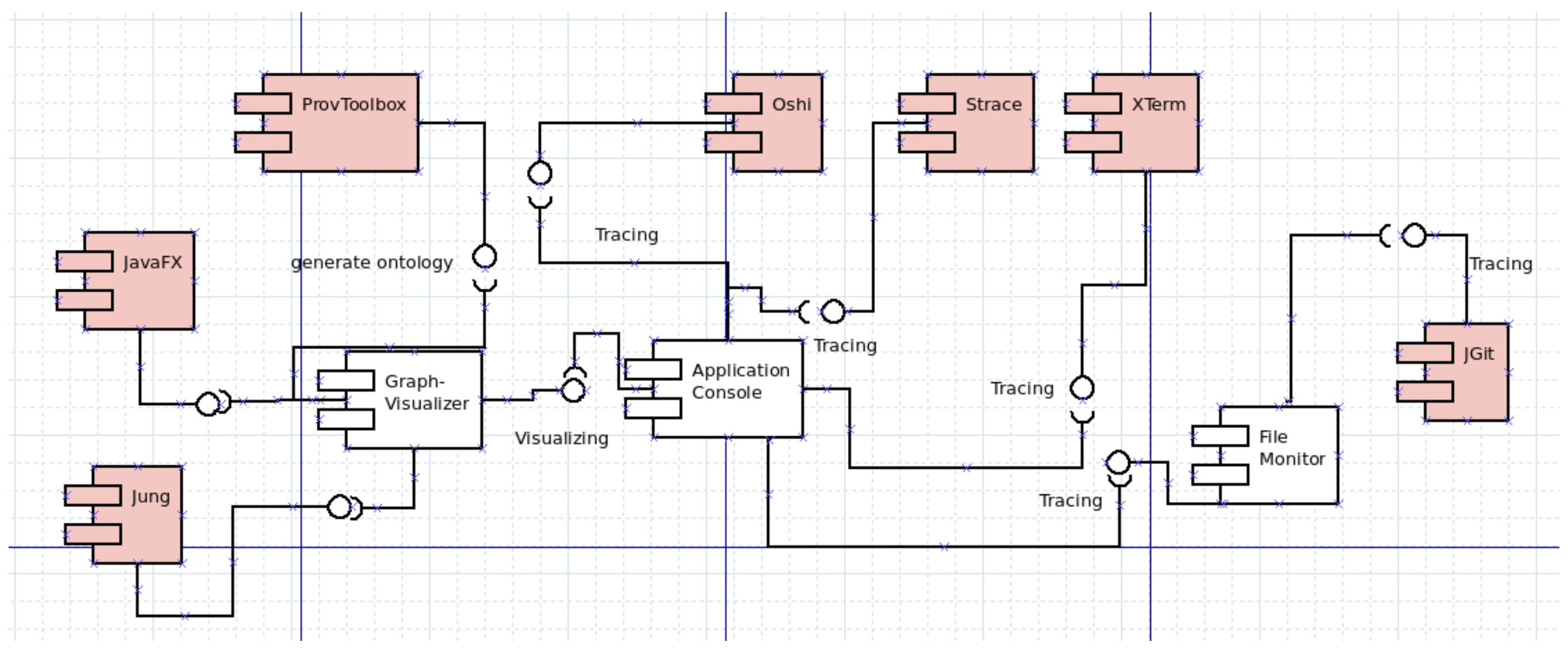

Figure 1: Main Components of the Prosci

Result

We construct the sample project with the following steps: 1.download dataset from external web service: (`http://spatialkeydocs.s3.amazonaws.com/FL`); 2. unzip the dataset; 3. using weka with version 3.8 and version 3.9 for computational processing. In this step, a new document "out.txt" is generated. 4. extract a subset from the original dataset and rerun the weka processing with two different versions. The result of the ontology is demonstrated in Figure 2. The version of the Weka doesn't impact the result, but by using the subset of the dataset we generate a new out.txt with "v2". Furthermore, we can verify that all activities and files are recorded correctly by Prosci tool.

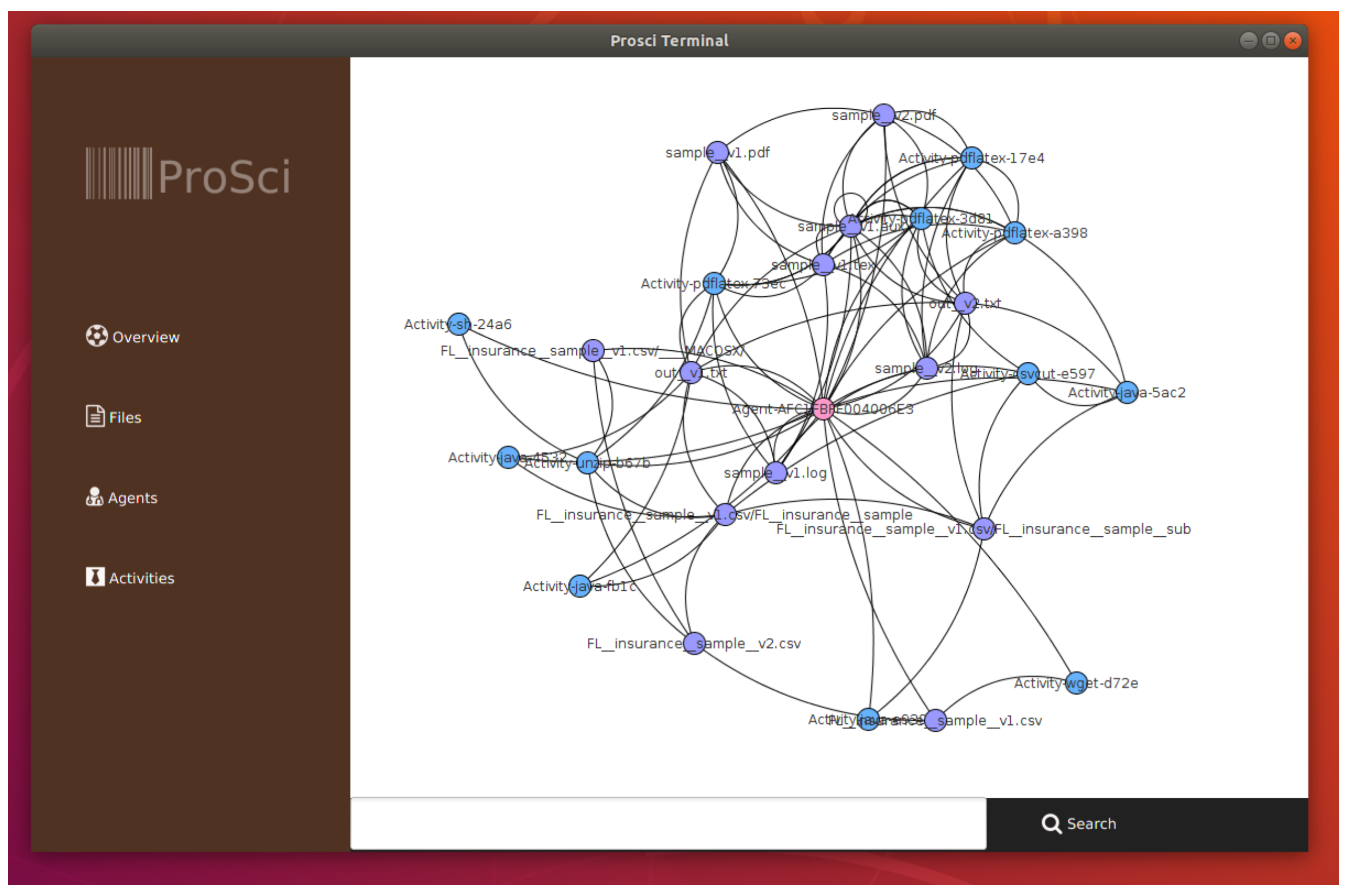

Figure 2: Ontology Graph of Sample Experiment

Conclusion

In our ProSci project, we have investigated a new tool, which integrates several APIs and tools. The whole project is built upon Java 8, using the Maven build tool, based on the provenance ontology theory. ProSci enables the tracking of the provenance data during the scientific workflow execution, visualizes the provenance metadata with the user-friendly GUI. From the result of the sample project we notice that the functionalities such as monitoring, tracking and capturing of the provenance data of a scientific experiment are approved. And the transforming mechanism of the provenance data into ontology graph and the user interactive ability are available. Therefore, we construct a meaningful Graph-Visualizer for representing the provenance information, which in return benifits the reproduction of the scientific experiment.

References

- Roger D. Peng.
  Reproducible research in computational science.
  Science,334(6060):1226–1227, 2011..
- Satya S Sahoo et. al. Minning, and Amit P Sheth.
  A unified framework for managing provenance information in translational research. BMC bioinformatics, 12(1):461, 2011.
- Timothy Lebo et. al.
  PROV-O: The PROV Ontology. W3C Recommendation. World Wide Web Consortium, 4 2013.

Kontakt: e1425097@student.tuwien.ac.at