Combining P-Plan and the REPRODUCE-ME Ontology to Achieve Semantic Enrichment of Scientific Experiments using Interactive Notebooks

Sheeba Samuel, Birgitta König-Ries

Heinz-Nixdorf Chair for Distributed Information Systems
Friedrich-Schiller University, Jena, Germany
sheeba.samuel@uni-jena.de, birgitta.koenig-ries@uni-jena.de

Abstract. End-to-end reproducibility of scientific experiments requires scientists to share their experimental data along with the computational environment. Interactive notebooks have recently gained widespread popularity among scientists because they allow users to document their experiments along with the code, visualize the results inline and selectively execute the code. In a multi-user environment where users can run and modify the shared notebooks, it becomes essential to capture the provenance of notebooks along with the experiments which used them. In this paper, we propose a way to capture provenance of these interactive notebooks and convert them into semantic descriptions so that a user can query the difference between the results, steps, errors and the execution environment of the code. We use the REPRODUCE-ME ontology extended from PROV-O and P-Plan to describe the provenance of notebook execution. We evaluate our prototype in a multi-user environment provided by JupyterHub.

Keywords: Notebooks, Provenance, Reproducibility, Experiments, Ontology

1 Introduction

Scientific experiments are a complex set of processes which involve multiple agents, activities, computational environment, input, and output. Additional challenges emerge with the collaborative and distributed experiment environment in terms of code sharing and execution. The inspiration for our work arises from the Collaborative Research Center (CRC) ReceptorLight[1], where scientists from multiple research institutes collaborate to develop high-performance microscopy techniques to understand the function of membrane receptors. In such a distributed and collaborative environment, it is necessary to understand the provenance of results generated by the researchers. In our previous work [9], we developed a prototype to capture the non-computational parts of an experiment

[1] http://www.receptorlight.uni-jena.de/

which includes the descriptions, agents, execution environment, devices, materials and methods used. To capture the provenance of the computational part of an experiment, we use Jupyterhub[2], which is a multi-user version of Jupyter Notebooks. It is an open source initiative which supports centralized deployment, centralized user-authentication and advances collaboration among scientists to document and run the code in any programming languages.
Several tools have been introduced to capture provenance of computational experiments. Scientific Workflow Management Systems capture provenance by running experiments as workflows. But, because of their steep learning curve, scientists still prefer writing scripts [6]. This has motivated approaches that aim to capture provenance from scripts and notebooks [[5], [2], [7]]. The YesWorkflow tool [5] is a language-independent tool where provenance data from a script is rendered as a workflow with the help of special annotations added by the user. Another recent approach is to convert notebooks into workflows where notebook developers need to follow a set of guidelines in writing code [2]. Carvalho et al. [1] present a methodology to convert scripts into workflow research objects with the help of tools like YesWorkflow, Research Objects, and PROV. All of these approaches have the limitation that they require changes to scripts by the user. Pimentel et al. [7] collect provenance data from IPython Notebooks by integrating noWorkflow [6] to the notebooks. However, this approach is limited to Python scripts. In our work, we capture and semantically describe the provenance of notebook execution in a multi-user environment using the REPRODUCE-ME ontology [8] by extending PROV-O [4] and P-Plan [3] to give a complete picture of a scientific experiment.

2 Development

We use the REPRODUCE-ME ontology [8] to describe the provenance of a notebook and its execution. In order to do this, we extend P-Plan [3] to represent the steps, plans, input and output variables and their relationship with each other. The prefixes *prov*:, *p-plan*: and *repr*: are used to indicate the namespace of all terms of PROV-O, P-Plan, and REPRODUCE-ME respectively. A *p-plan*:*Plan* consists of smaller steps *p-plan*:*Step* which consumes and produces *p-plan*:$Variable$. A *repr*:*Experiment* and *repr*:*Notebook* are the subclasses of *p-plan*:*Plan* and the *repr*:*Notebook* is related to *repr*:*Experiment* using the object property *p-plan*:$isSubPlanOfPlan$. A cell of the notebook, a *repr*:*Cell*, is a *p-plan*:*Step* which generates an output which is described as a *p-plan*:$Variable$. The source of the cell is described as an input variable. The creation of the notebook is described using *prov*:$generatedAtTime$ and the modification time using *repr*:$modifiedAtTime$. The order of the execution of cells is described using *p-plan*:$isPrecededBy$. The *repr*:*Session*, a subclass of *p-plan*:*Activity*, describes the session of a notebook user who is described using *prov*:*Agent*. The execution environment of a notebook is described using *repr*:*Setting* which includes *repr*:*ProgrammingLanguage*, *repr*:$Version$ and *repr*:$Kernel$.

[2] `https://jupyter.org/`

JupyterHub is installed and connected to our prototype so that users can create new notebooks, run and share them. The notebooks are stored in a centralized place so that they can be shared and run by scientists that belong to a group. Our prototype fetches the metadata of the notebooks from the Jupyter Notebook and JupyterHub REST APIs which provide details of the notebooks, the kernel, and the programming language used, the sessions and the users. The metadata that is useful for scientists is stored along with the other experimental data. The captured provenance data is then mapped to the ontology using ontology-based data access technique. The prototype provides a dashboard which runs SPARQL queries and visualizes the experimental data including the people who are involved in the experiment, the devices and their settings, publications used in the experiment and the notebook data. Figure. 1 shows the project dashboard in our prototype. In this way, the prototype provides a complete picture of an experiment. Listing 1.1 shows an example SPARQL query to find all the notebooks used in an experiment and their metadata.

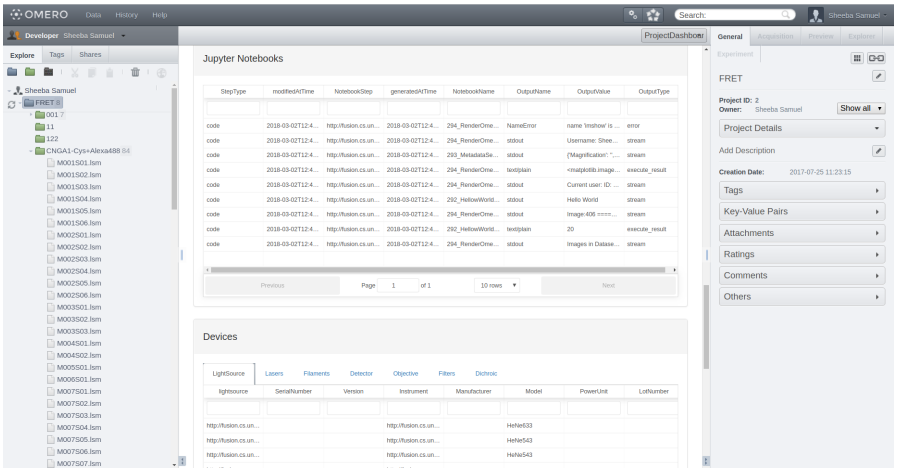

Fig. 1: The Project Dashboard in the prototype [9]

```
Select DISTINCT * WHERE {
   ?notebook a :Notebook ; :name ?NotebookName ;
   prov:generatedAtTime ?generatedAtTime ;
   :modifiedAtTime ?modifiedAtTime ;
   p-plan:isSubPlanOfPlan ?experiment .
   ?experiment a :Experiment .
   ?NotebookStep p-plan:isStepOfPlan ?notebook ;
   p-plan:hasOutputVar ?output ; :type ?StepType .
   ?output prov:value ?OutputValue ;
   :type ?OutputType ; :name ?OutputName
}
```

Listing 1.1: SPARQL query for Notebooks used in an experiment

The REPRODUCE-ME ontology, the mappings and the SPARQL queries used for evaluation are publicly available[3].

[3] `http://fusion.cs.uni-jena.de/fusion/repr/`

3 Conclusion and Future Work

The REPRODUCE-ME ontology was initially developed for microscopy-based experiments. Since scientists use scripts to perform data analysis, we decided to expand the ontology by extending W3C vocabularies to describe the widely used notebooks. In this paper, we semantically enrich the scientific experimental data using notebooks in a multi-user environment provided by JupyterHub. The prototype provides a dashboard which visualizes the experimental data along with the notebooks used to generate the final results. This allows the user to visualize the complete path taken for an experiment from its input to its output along with the execution environment. As future work, we aim to evaluate the prototype based on scalability measures.

Acknowledgements

This research is supported by the Deutsche Forschungsgemeinschaft (DFG) in Project Z2 of the CRC/TRR 166 High-end light microscopy elucidates membrane receptor function - ReceptorLight. We thank Christoph Biskup, Kathrin Groeneveld and Tom Kache from University Hospital Jena, Germany, for providing the requirements to develop the proposed approach and evaluating the system.

References

1. Carvalho, L.A.M.C., Belhajjame, K., Medeiros, C.B.: Converting scripts into reproducible workflow research objects. In: 2016 IEEE 12th International Conference on e-Science (e-Science). pp. 71–80 (Oct 2016)
2. Carvalho, L.A.M.C., Wang, R., Gil, Y., Garijo, D.: Niw: Converting notebooks into workflows to capture dataflow and provenance (2017)
3. Garijo, D., Gil, Y.: Augmenting PROV with plans in P-Plan: scientific processes as linked data. CEUR Workshop Proceedings (2012)
4. Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., et al.: PROV-O: The PROV Ontology. W3C Recommendation 30 (2013)
5. McPhillips, T.M., Song, T., Kolisnik, T., Aulenbach, S., Belhajjame, K., et al.: Yesworkflow: A user-oriented, language-independent tool for recovering workflow information from scripts. CoRR abs/1502.02403 (2015)
6. Pimentel, J.a.F., Murta, L., Braganholo, V., Freire, J.: noworkflow: A tool for collecting, analyzing, and managing provenance from python scripts. Proc. VLDB Endow. 10(12), 1841–1844 (Aug 2017)
7. Pimentel, J.F.N., Braganholo, V., Murta, L., Freire, J.: Collecting and analyzing provenance on interactive notebooks: When ipython meets noworkflow. In: 7th USENIX Workshop on the Theory and Practice of Provenance (TaPP 15). USENIX Association, Edinburgh, Scotland (2015)
8. Samuel, S., König-Ries, B.: REPRODUCE-ME: ontology-based data access for reproducibility of microscopy experiments. In: The Semantic Web: ESWC 2017 Satellite Events, Portorož, Slovenia. pp. 17–20 (2017)
9. Samuel, S., Taubert, F., Walther, D., König-Ries, B., Bücker, H.M.: Towards reproducibility of microscopy experiments. D-Lib Magazine 23(1/2) (2017)