PubFlow: provenance-aware workflows for research data publication

Peer C. Brauer
Software Engineering Group
Department of Computer Science
Kiel University

Wilhelm Hasselbring
Software Engineering Group
Department of Computer Science
Kiel University

Abstract

In this paper we present a workflow oriented data publication framework called PubFlow. PubFlow is an ongoing research project with the goal to create a framework, which alleviates the process of data publication. A main feature of PubFlow is its provenance capturing mechanism.

We also present an approach for collecting provenance information in a scientific workflow environment like PubFlow and give an outlook on a data archive for storing provenance information. This archive will be based on a NoSQL graph database and the W3C provenance ontology PROVO[1].

1 Introduction

Today the management of research data is one of the major challenges of sciences [1]. Data is no longer collected only for one experiment or research project, but is reused in many projects. At the same time the amount of data, that is captured and used as base for scientific work, is growing rapidly. Jim Gray and others described this change of scientific paradigms in their book the Fourth Paradigm [6]. They gave an overview over the development of the scientific paradigms from the beginning of science until now. Their conclusion is, that today we are facing a new scientific paradigm, that science is becoming more and more data centric. They postulate, that this new paradigm demands new tools and approaches for data management. From their point of view, one of the main components of each scientific data infrastructure is a system for capturing provenance information for the research data throughout the whole scientific process.

Not only have the authors of the fourth paradigm postulated this. Also Fleischer and Jannaschk conclude this [3]. They analyze the current situation of data managers and give some advice, how the whole scientific data managing process can be improved. Beside the need for a semi-automated system for research data publication, they identified a need for automated provenance capturing to record the provenance information of the research data during the whole scientific process. This leads to a fundamental question. What is the provenance information of research data? At different situations it showed up, that there is no simple answer to this question. Every data manager and every scientist, even those who work in this domain, will give a slightly different answer. In the context of our work we define provenance information as the history of data. The provenance information of a piece of research data contains all the information needed to revert all operations, which were performed on this data. Everything else belongs to the meta-data of this specific dataset. In summary one can say, provenance data of a specific set of processed research data is exactly that data, that describes the inverse map, which is needed to get back to the original raw data. In this paper we first introduce a semi-automated publication system for research data - PubFlow which we are currently working on. We will give an introduction to the PubFlow system and its areas of application. Further we will explain PubFlows provenance collection mechanisms, which assure the collection of the provenance information of the research data during the whole publication process. The provenance capturing system is based on Kieker[2] [7, 8], an application monitoring framework, which is also a recommended tool of the SPEC Research Group[3].

2 PubFlow

PubFlow[4] is a publication workflow system. This means, PubFlow provides the tools and the infrastructure, which is needed by scientists and data managers to get the research data out of the institutional data repositories into the publicly available data centers. The current area of application is the domain of marine science. Here, the system is used to get the data out of an institutional

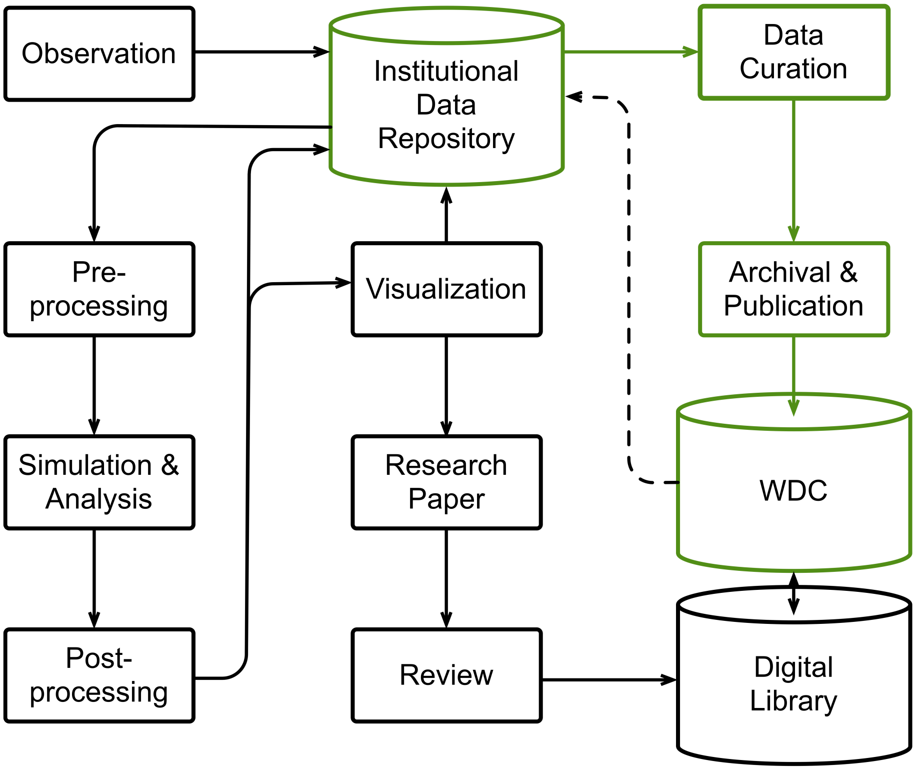

Figure 1: data flow in research

data repository located at the GEOMAR[5] in Kiel into the WDC-Mare Pangaea[6]. Figure 1 shows a simplified data flow diagram for research data. It starts with the observation and ends with the archival in the world data centers. PubFlow provides the infrastructure to transfer the research data from the local repositories to the WDCs[7]. One of the main ideas of PubFlow is to bring the role-based division of work into the domain of scientific workflow environments. Currently most scientific workflow environments are based on the idea, that scientist design their own workflows in an iterative way to fit their specific needs. There are no different roles like the designers, who create the workflows, and the scientists, who only use them. This is different in PubFlow. It allows data managers to define their own, customized workflows for data publication, which then can be used by researchers to upload their data to the archives. PubFlow provides a different user interface for each role. Data managers use the integrated workflow editing component to create their workflows in a domain specific, but BPMN 2.0 compliant, graphical language. They can also access the provenance information record for each workflow execution.

The scientist on the other hand accesses PubFlow through a task management interface like i.e. JIRA [8]. They do not have direct access to the workflows, but can initiate new workflow runs by creating new tickets or tasks. When created, these tickets automatically start the correct publication workflows. If human interaction is needed during the execution of a workflow, the system creates a new task in the ticket system and assigns it to the scientist, who started the workflow. PubFlow is built upon different proven technologies. The core of the system is the Apache ODE workflow engine [9]. Apache ODE is an open source project developed by the Apache Foundation. ODE executes workflows following the WS-BPEL standard [10]. The publication workflows, which are defined in a BPMN 2.0[11] compliant workflow language, are transformed to BPEL by the use of a model based transformation framework [4]. The provenance information is collected by the use of the Kieker workflow monitoring framework.

3 PubFlow and provenance

One of the main features of PubFlow is its provenance information collection. This feature consists of two different parts. First the collection of the provenance data. In PubFlow the collection of provenance data starts, when a dataset is imported into the system and ends, when the data is publicized into the public archives. During this whole process, the provenance information is collected and aggregated, such that a complex provenance graph (Figure 2) evolves, which contains the complete history of the data modifications and conversions done by PubFlow. The provenance data is collected by the help of the Kieker monitoring framework, which has been extended in PubFlow to enable workflow monitoring. The monitoring with Kieker and how the provenance information is derived from this data is described in the next subsection. Another important part of PubFlow's provenance information system is the provenance data store. This store is build upon the widely know graph database Neo4j[12] and the Prov-DM of the W3C provenance group[13].

3.1 Provenance information collection

As mentioned in the last section, PubFlow collects provenance information for the processed data sets by the use of a workflow monitoring extension to the Kieker framework [8, 7]. In this paper, we will only give a short overview over the functionality of the provenance capturing mechanism. More information about Kieker's workflow monitoring extension can be found in [2]. Kieker was chosen as basis for the provenance capturing mechanism, because it is a modular, open source monitoring framework, which can easily be extended by user defined components. The development of the framework started in 2006. Since then, it has evolved from a small tool for monitoring the response times of Java applications to a powerful monitoring and analysis tool. Currently, the standard implementation provides modules for monitoring and analyzing not only the runtime behavior of software systems, but also its inner structure. Therefore it includes different sorts of monitoring probes and analysis components.

To collect provenance information, we added specially designed Kieker monitoring probes into all the different components of PubFlow. The main focus was laid on the Apache ODE workflow engine. A new extension for Kieker was created, to enable Kieker to monitor the execution of BPEL workflows within the workflow engine. The workflow probes keep track of every change, which is done to the original scientific data by the workflow and passes the event logs and descriptions to the kieker analysis component. Here the monitoring data which was collected by all the different probes integrated in each component of the PubFlow framework is aggregated to a provenance graph and stored in Kiekers provenance archive.

3.2 Provenance information representation

PubFlow stores the collected provenance information in an integrated provenance archive. This archive, like the proost provenance archive of the DLR[14] [5], is build upon the widely known Neo4J graph database to store the collected provenance data. But unlike proost our provenance archive uses a java implementation of the PROV specification for provenance on the Web to represent the provenance graph. The provenance information archive is an independent module, which can not only used by PubFlow, but can also be integrated in other applications or used as a stand alone application. The provenance archive provides two different ways to insert data in the archive. If it is embedded in another application, developers can use a simple Java API to interact with the archive module. In standalone mode, the archive can be accessed through a webservice interface.

In the near future the provenance archive will be extended with a provenance browser. The provenance browser is a web interface for accessing the stored provenance information. The browser will provide two different views of the provenance information, a textual view and a graphical view of the provenance graph. The provenance graph shown in figure 2 was created with a first version of the provenance browser.

4 Our case study

The PubFlow framework is developed, tested and evaluated in the context of the Kiel marine science. In cooperation with our project partners from GEOMAR, ZBW[15] and the KDMT[16], we worked out an evaluation scenario for the PubFlow framework containing different typical tasks, which are needed to transfer research data to the archives. The first use case we carried out, was the automatic transfer of oceanographic research data, more precisely measuring data collected by a CTD[17] probe, from the institutional data repository of the GEOMAR[18] to the WDC mare PANGAEA. To transfer the data we implemented a BPMN 2.0 workflow containing all the steps needed to upload the data. This BPMN 2.0 workflow describes how the different modules and filters for data handling and conversion provided by the PubFlow framework should be orchestrated, so that the original research data set can be converted to an export data format which the is uploaded to the WDC mare.

5 Future work

Currently we have several ideas on how to add additional functionality to the PubFlow system, once the system is completed. One idea is to add an additional view to the

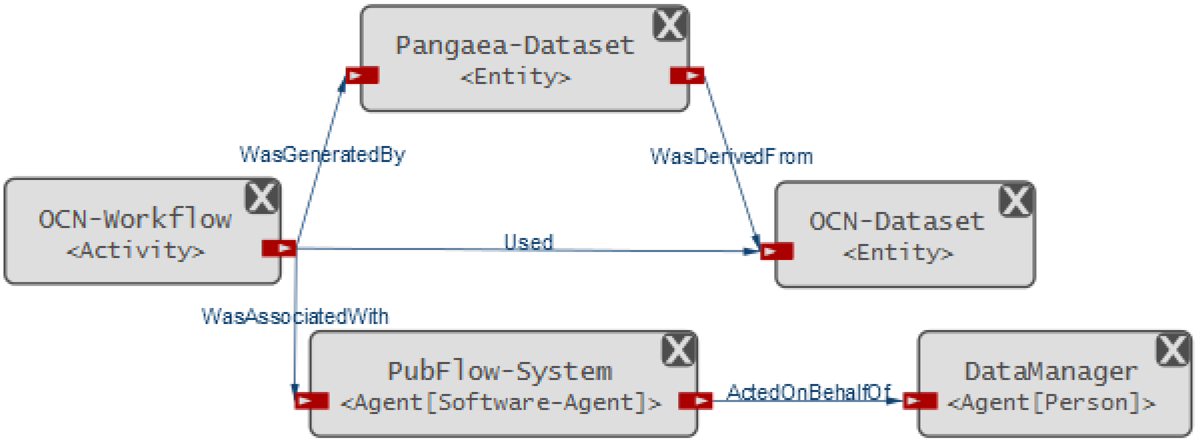

Figure 2: sample provenance graph

system, which allows scientists to create their own explorative workflows such that they can use PubFlow not only as a publication system, but also as a classic scientific workflow environment. A different task will be the integration of autonomous sensors. We think it would alleviate the daily work of data managers and scientist, if autonomous research equipment can directly upload its data to the data archives or institutional repositories. Also some improvements to the provenance archive would be helpful. We think of an enhanced provenance archive, which can be used as a public available provenance data archive just like the way world data archives are used for normal research data. It should be possible for provenance aware applications like, but not limited to, PubFlow to store their provenance information in this system, so it can be used by scientist from all over the world to validate the research data from the WDCs.

6 Availability

PubFlow is currently in the implementation phase. You can find supplementary information about PubFlow and Kieker under:

```
http://www.pubflow.de
http://kieker-monitoring.net
```

References

[1] Bell, G., Hey, T., and Szalay, A. Computer science. beyond the data deluge. *Science 323*, 5919 (2009), 1297–8.

[2] Brauer, P. C., and Hasselbring, W. Capturing provenance information with a workflow monitoring extension for the kieker framework. In *Proceedings of the 3rd International Workshop on Semantic Web in Provenance Management* (Mai 2012), vol. 856 of *CEUR Workshop Proceedings*, CEUR-WS.

[3] Fleischer, D., and Jannaschk, K. A path to filled archives. *Nature Geoscience 4* (2011), 575–576.

[4] Scherp, G., and Hasselbring, W. Towards a model-driven transformation framework for scientific workflows. *Procedia Computer Science 1*, 1 (2010), 1513 – 1520. ICCS 2010.

[5] Schreiber, A., Ney, M., and Wendel, H. The provenance store proost for the open provenance model. In *Provenance and Annotation of Data and Processes*, P. Groth and J. Frew, Eds., vol. 7525 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2012, pp. 240–242.

[6] Tolle, K. M., Tansley, D. S. W., and Hey, A. J. G. The fourth paradigm: Data-intensive scientific discovery, 2011.

[7] van Hoorn, A., Rohr, M., Hasselbring, W., Waller, J., Ehlers, J., Frey, S., and Kieselhorst, D. Continuous monitoring of software services: Design and application of the kieker framework. Forschungsbericht, Kiel University, November 2009.

[8] van Hoorn, A., Waller, J., and Hasselbring, W. Kieker: A framework for application performance monitoring and dynamic software analysis. In *Proceedings of the 3rd joint ACM/SPEC International Conference on Performance Engineering (ICPE 2012)* (April 2012), ACM, pp. 247–248.

Notes

[1] http://www.w3.org/2011/prov/

[2] http://kieker-monitoring.net/

[3] http://research.spec.org/projects/tools.html

[4] http://www.pubflow.de

[5] http://www.geomar.de/

[6] http://www.pangaea.de/

[7] WDC stand for World Data Center

[8] http://www.atlassian.com/de/software/jira

[9] http://ode.apache.org/

[10] http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html

[11] http://www.omg.org/spec/BPMN/2.0/

[12] http://www.neo4j.org/

[13] http://www.w3.org/TR/2012/CR-prov-dm-20121211/

[14] Proost is an open source provenance archive build upon the open provenance model (http://software.dlr.de/p/proost/home/)

[15] http://zbw.eu

[16] KDMT stands for Kiel Data Management Team (http://www.geomar.de/zentrum/einrichtungen/rz/daten/)

[17] a CTD probe is a instrument for analyzing the conductivity, temperature and pressure of ocean water in the complete head of water

[18] http://portal.geomar.de