PubFlow: a scientific data publication framework for marine science

Peer C. Brauer, *Kiel University,* peer.brauer@informatik.uni-kiel.de *(Germany)*
Wilhelm Hasselbring, Kiel University, wilhelm.hasselbring@informatik.uni-kiel.de

We are facing a deluge of data. Improvements in technology and research methods have considerable increased the amount of data collected by scientific experiments. An example for this evolution of scientific instruments and methods is the development in the domain of depth measurement. For years this was a data poor science. The depth of the sea was calculated by using a hand lead or later a simple echo sounder. Today things have changed. Modern multi-beam echo sounders produce a detailed and also data rich height profile of the sea floor.

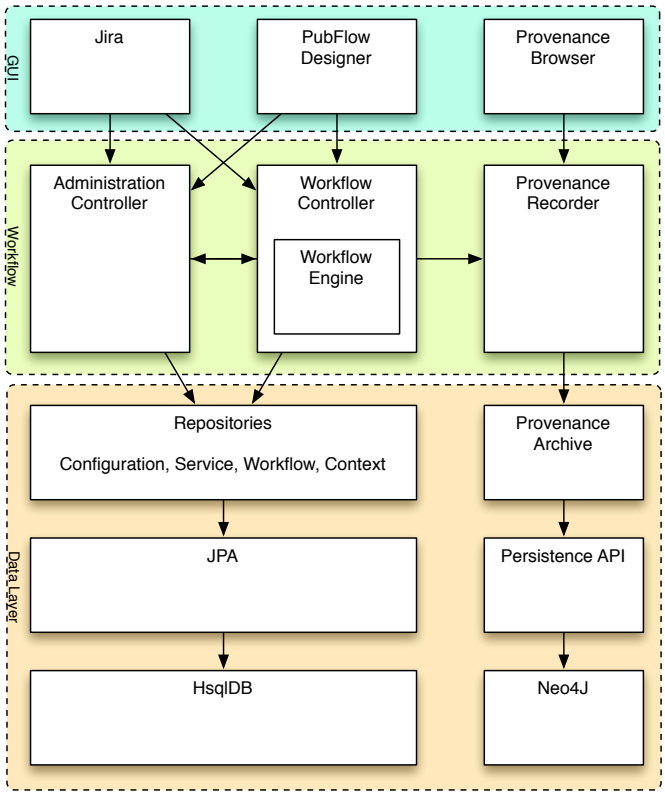

Figure 1 : The architecture of the PubFlow system

But these developments rise a new question, how to deal with the bulk of data created by all these experiments? Scientific data is too valuable to just be used once. On the other hand, new research methods, which developed during the last years and which Jim Gray [1] referred to as the fourth paradigm of science, base on the reuse of huge amounts of data from other experiments. So new ways and methods have to be discovered to alleviate the reuse of data and to get it out of the local repositories into public available archives [3]. PubFlow [4] is our suggestion for such a tool.

PubFlow

PubFlow is a data publication framework for scientific data, build on top of proven business workflow technologies like BPMN 2.0[1], Apache ODE [2] and JBoss JBPM[3]. It brings automation and the division of work to the domain of scientific data management. Pubflow is based on the assumption, that data managers know best about the processes and guidelines, which have to be followed to publish research data to a public available archive. Unfortunately the amount of scientific data is so overwhelming, that data managers alone can not curate each dataset and upload it to the archives. Researchers, institutes and funding agencies on the other hand want their research data to be published. This factor was considered, when the PubFlow system was planed. In PubFlow the role of the data managers is to define the publication workflows and to take care for complex tasks. The publication process for a specific datset on the other hand is started by a scientist. He chooses a predefined workflow meeting his requirements from a list and starts it through a normal ticket system

[1] omg.org/spec/BPMN/2.0/

[2] ode.apache.org

[3] jboss.org/jbpm

like i.e. Jira[4] providing his datset as input. After this the PubFlow system runs the selected workflow, which was predefined by the data managers, on the dataset the scientist uploaded. Every time a problem occurs and the workflow can not be continued, PubFlow creates a new ticket in the ticket system and assigns it to a datamanager or to the researcher, who uploaded the dataset to the PubFlow system. If the problem described by the ticket is marked to be solved, PubFlow continues the workflow execution.

Figure 1 depicts the architecture of the PubFlow system. On the GUI-Layer at the top one can find three user interface components. This reflects the division of work in the PubFlow system. Jira is the interface for the researchers to start the workflows, the PubFlow designer is the interface for the datamanagers to create and maintain the workflows. The third interface provides access to pubflows provenance system [2,4]. During the exection of each workflow PubFlow collects provenance information for the data currently processed by the workflow. This information is available through the PubFlow Provenance Browser.

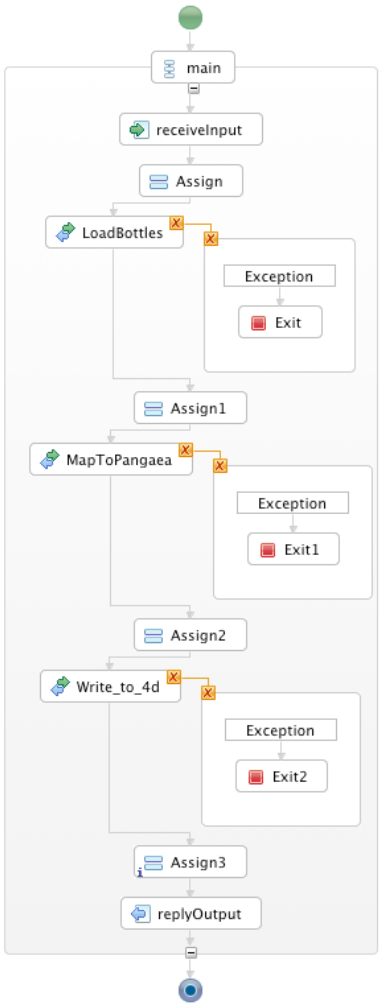

Figure 2 : PubFlow workflow

The evaluation scenario

The PubFlow system is evaluated in cooperation with the Kiel Datamanagement Team[5] located at the Geomar Helmholtz Centre for Ocean Research Kiel. The first publication workflow we implemented for PubFlow is a workflow for publishing data collected by a CTD probe. This data is stored in an institutional repository and has to be transfered to the world data center mare - Pangaea[6]. Figure 2 shows this workflow. It mainly consists of five steps. In a first phase the information, the scientist provided, when he started the task, is loaded from the ticketsystem into the workflow engine. The next step is to fetch the data from the institutional repository. Now PubFlow performs predefined mapping and conversion operations on the dataset, so the output data format is compatible to the one used by the world data center. At last the data is written to the specified output format and exported. Although it is still under active development, PubFlow has already proven to be a very helpful tool for data managers by automating simple or periodic data management tasks.

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, 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] BRAUER, P. C., AND HASSELBRING, W. PubFlow: provenance-aware workflows for research data publication In: 5th USENIX Workshop on the Theory and Practice of Provenance (TaPP '13), April 2-3, 2013, Lombard .

---

[4] atlassian.com/jira

[5] geomar.de/en/service/data-management/

[6] pangaea.de