Start smart and finish wise:

The Kiel Marine Science Provenance-Aware Data Management Approach

Peer Brauer
Software Engineering Group
Kiel University
pcb@informatik.uni-kiel.de

Andreas Czerniak
GEOMAR Helmholtz Centre for Ocean Research Kiel
aczerniak@geomar.de

Wilhelm Hasselbring
Software Engineering Group
Kiel University
wha@informatik.uni-kiel.de

Abstract

While creating or processing scientific data, it is very important to capture and to archive the corresponding provenance data. “Start smart and finish wise” is our approach for a provenance aware tooling, which helps data managers and scientists not only to manage their data, but also to capture their scientific data in the field, to record the provenance data, to store it for further analysis and finally to publish the scientific data to the data centres. The tool chain consists of four major components, (1) the digital Pen for capturing the (meta) data and the corresponding provenance information in the field, (2) the OCN database for data-acquisition workflows and the data repository, (3) the PubFlow framework for scientific data publication and (4) CAPS for capturing provenance data in Java based scientific software. During each processing step in “Start smart and finish wise” the provenance data for the scientific data is captured and archived.

Keywords Data provenance, Data Publication, Scientific Workflows, Data Management

1. Introduction

Today, the way science and specially marine science is done is changing rapidly. New scientific methods, techniques and new tools have turned computer-based marine science during the last decades into a data-intensive scientific field. These new developments enable scientists to create a much more detailed model of the oceans and to gain a deeper understanding of the processes inside ocean, than it was possible before. But the constantly increasing complexity and quantity of scientific data brings new data management requirements. The quantity and complexity of scientific data has reached a point, where it is nearly impossible to curate, publish or to archive these data sets manually [BHS09, HTT09]. At Kiel, the Kiel Data Management Team (KDMT)[1] at the GEOMAR is addressing this problem during the last years. The team developed a process to automize the capturing of scientific data as well as its transfer to an institutional archive, which is a component of the OCN system (our ocean database). Here, the PubFlow framework[2] [BH13] takes over. The PubFlow framework is responsible for publishing the data from OCN to the world data center (Pangaea[3] in our case). During this whole process the provenance data, for the scientific data processed, is captured and saved:

1. During the observation the scientists capture their data using a smart pen and the digital ink software with predefined paper forms.
2. The data entered in these forms is automatically extracted and stored in the institutional repository.
3. PubFlow loads this data from the institutional repository and performs some basic checks on the data.
4. PubFlow publishes the data to the world data centres.

This approach is not meant to replace the work of the data managers, but to assist them and to free them from recurrent tasks. Due to the nature of scientific data, data-management workflows cannot be fully automated.

The rest of our paper is divided in five sections. In the first four sections we explain the tool chain and the tools used in the data management approach described in this paper. These tools were created by the KDMT or by the PubFlow project. Finally we summarize our work and give an outlook of that what might come in the near future.

2. Capturing data with a smart pen and digital ink software

Keeping the quality of scientific data high requires to start with data management in an very early stage in the process of the creation of scientific data. The Kiel Data Management Infrastructure (KDMI) places data capturing straight in the data creation process. To take advantage of a scientist’s habit, we chose to use familiar but specifically prepared paper forms, which can be related to their digital representation using the capabilities of digital pens and the digital ink software. Figure 1 shows an overview of the Kiel Data Management Infrastructure and the flow of data from research vessels into OCN database component.

In our approach the archival process of the scientific data already starts, when the scientist takes his notes in the field. We created field notebook forms, which are printed on a special dot-paper. To fill these forms, scientist can either use a normal pen or an intelligent smart pen [CFS$^+$12]. This intelligent digital pen captures the information entered in the form while the scientist writes it down. The information and data collected is stored inside the pen until the

[1] Kiel DataManagement Team - www.geomar.de

[2] http://www.pubflow.uni-kiel.de

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

TAPP ’14, 06 12–13, 2014, Cologne, Germany.
Copyright © 2014 ACM 978-1-nnnn-nnnn-n/yy/mm. . . $15.00.
http://dx.doi.org/10.1145/nnnnnnn.nnnnnnn

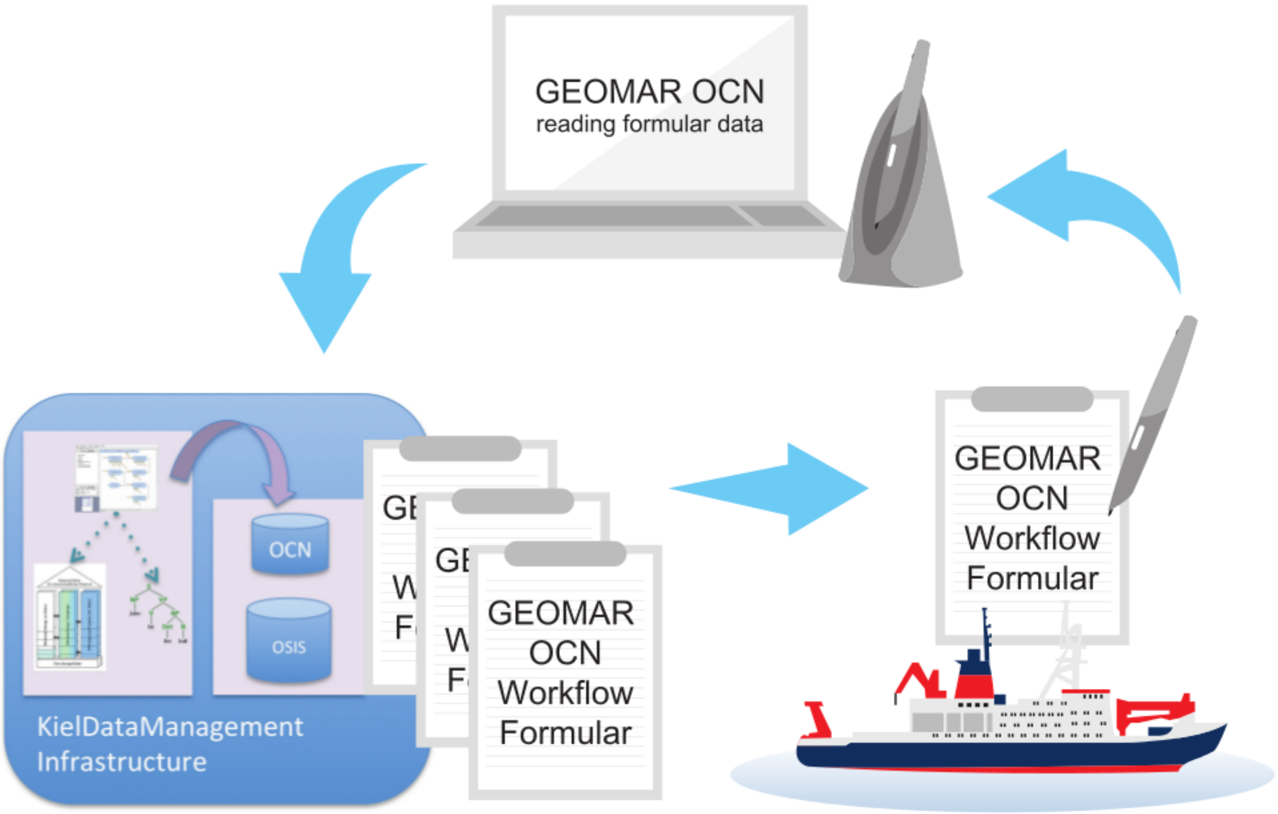

Figure 1. The flow of special paper-based forms and loss-less transfer into OCN

scientist uploads it to the KDMI infrastructure. Here the handwriting is recognized by the use of the digital ink software and a electronic representation of the form from the field notebook is created. The digital copy of his data is then made accessible to the scientist. This way, scientists can simultaneously access their written documents, while having a digital representation of their documents and data at hand. Smart pens and integrated software components allow loss-less transfer of data and information from paper into digital content that can be serialized into the OCN database component.

Another benefit of the use of smart pens is that the scientists don't have to bother with the archival process of the data. This way the scientific data is delivered to the data infrastructure without the need of extra work by the scientists. The data archival process is not longer a annoying duty, the contrary is the case. Having a digital copy of their data is an incentive for scientists to use our tools.

In the data flow of scientific data in Kiel marine science, the smart pen covers the capturing of the scientific data during the observation and its transfer to the institutional repository the OCN database.

3. The OCN components

At the GEOMAR, the data captured by the digital pen is preserved in the OCN database. An institutional repository, not only for scientific data, but also for the provenance information associated with this data. Additionally to the classic provenance information, like where, when and by whom the data was collected, the OCN database component contains further information about the scientific workflow of marine scientist's in the form of a description not only in the workflow itself, but also of every single step of the workflow. These detailed workflow definitions are already defined by data managers and scientist in advance of an expedition or an experiment. To enter the information about the workflow in the OCN database, the OCN Workflow Designer is used. The designer allows it to describe the workflows by the use of a graphical domain-specific language. Already in this early stage the scientist defines, which data should be captured when. But later in the field he is not limited to his selections, the OCN infrastructure allows it always to add further observations or experiments to an existing workflow. During the expedition all the data collected is associated to this workflow description.

The provenance information stored in the OCN database and some additional information from the Ocean Science Information System (OSIS) is also used for constantly refining the dot paper forms used in the field.

4. Data Publication with PubFlow

PubFlow is a data publication framework for scientific data, build on top of proven business workflow technologies like BPMN 2.0, Apache ODE and JBoss JBPM. 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 scientific data to a publicly 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. Scientists, institutes and funding agencies on the other hand want their scientific 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 dataset

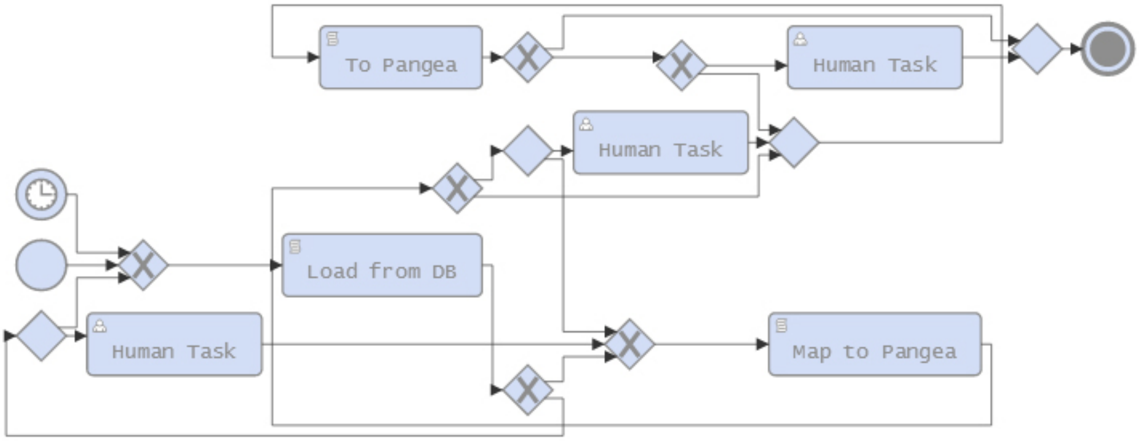

Figure 2. The OCN workflow executed in PubFlow (simplified)

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 like i.e. Jira providing his dataset 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 data manager or to the scientist, 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.

In our approach PubFlow the research data is transferred from the institutional archive, the OCN database, into the World Data Center for oceanographic data Pangaea.[4] As described above, exporting the data is as easy for the scientist as filling out a simple web form. To simplify the process of uploading the data, we choose Jira,[5] a commercial issue management framework developed by Atlassian, for interacting with the scientist. Jira is already widely used in many domains and well know by many scientist and data managers. We created a specially designed Jira Plugin, which allows PubFlow, to connect to any Jira instance, which has the plugin installed. Once the plugin has been added to the Jira instance, PubFlow starts creating new issue types, workflows, and dialogues within this instance. Immediately, scientists can access these new functions and start uploading their data by creating a new issue ticket.

Once the issue is created, it is transferred to the PubFlow system by the use of a SSL-encrypted and certified web-service connection. PubFlow performs basic security and plausibility checks on transmitted data. If any error occurs within these tests, PubFlow informs the user, who created the issue, and allows him to correct the information he provided. This whole communication process is processed by the use of Jira and Jira issues. In the case that all needed information was transmitted to the PubFlow system, PubFlow loads the corresponding data publication workflow from its integrated workflow repository, determines its type and deploys it in the execution environment. Figure 2 depicts a simplified version of the workflow used to transfer research data out of the OCN database into the world data center Pangaea. PubFlow supports publication workflows defined in BPMN 2.0 and BPEL. To integrate other types of workflows, PubFlow provides a simple way of adding new workflow engines to the PubFlow workflow environment. The publication workflow is executed in a dedicated workflow container and communicates with the rest of the PubFlow system just by the use of a ActiveMQ message bus. In the future, we will add a new deployment scenario for the workflow engine container to the PubFlow system. The plan is to add a cloud deployment option to the system, such that the new workflows can be started on dedicated cloud nodes. This option will allow us to process much larger data sets or long-living processes, that would other wise decrease the performance of the whole Pubflow system.

Once the workflow is finished, it informs the user by creating a new Jira ticket and assign it to him. The ticket contains all results of the workflow run, like:

- The information where the data was published
- DOIs assigned to the data set
- The link to the published dataset
- Warnings produced by the workflow engine
- A pointer to the provenance information collected during the execution of the data set

As mentioned above, during the execution of the publication workflow the provenance information for the data set is recorded and processed by the CAPS framework.

5. The CAPS Framework

CAPS[6] is responsible for capturing, analyzing and archiving the provenance information for the data published by PubFlow. In CAPS data provenance is seen as a cross-cutting concern. To capture the provenance information for a research data set, CAPS integrates in the software,[7] which is used to process the data, by the use of AspectJ[8] and different mechanisms and programming interfaces provided by the Java JVM. The integration process is described in [Bra14]. Once integrated in the software, CAPS collects the provenance information and processes it by the use of Kieker [RvHM$^+$08, vHWH12]. Kieker is a monitoring framework

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

[5] http://www.jira.atlassian.com

[6] CAPS stands for Capturing and Archiving Provenance information in Scientifc workflows

[7] CAPS is limited to software based on Java

[8] http://www.eclipse.org/aspectj/

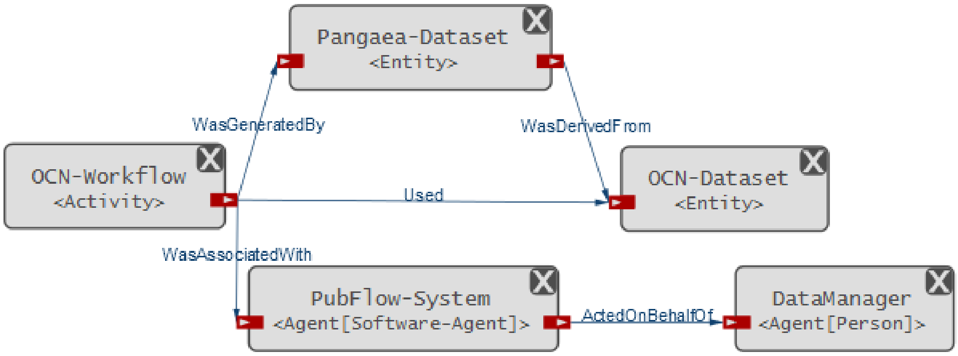

Figure 3. Provenance Graph for OCN Dataset processed by PubFlow (simplified)

and provides a powerful data analysis toolkit, which is employed by CAPS to reconstruct the provenance graph for the research data processed from the observation records delivered by the observation pointcuts CAPS integrates into the host software via AspectJ.

The provenance graph (Figure 3) is stored in an integrated provenance archive build upon the W3C Prov-O provenance model,[9] the Eclipse Modelling Framework[10] and Neo4j.[11] We choose a graph database as a persistence layer for our provenance model, because it allows us to perform sophisticated search and insert operations on the provenance graph.

Currently we are working on a web-based graphical user interface for the provenance archive, that can be used to visualize and browse the provenance graph. Figure 3 shows a screenshot of a simplified provenance graph of the OCN workflow visualized by our provenance browser. As one can see, the graph contains a selection of information of a data sets history. Which information are seen to be relevant to describe the history of a dataset, depend on the data manager, who configures the provenance capturing mechanism in CAPS. In this scenario, the provenance graph contains information about the raw data, that was used to create the Pangaea data set, the workflow, that was run by the PubFlow system, and the user who started the system. More information about each component in the provenance graph can be accessed by clicking on the component. When clicked, the component will expand and display more detailed information.

6. Conclusion and Future Work

”Start smart and finish wise” has some benefits compared to other existing data management approaches. Since all components of our approach are integrated, data provenance is assured during the whole data management process. Digitizing the data directly in the field reduces the loss of information and of data quality due to transcription errors and using paper-based forms increases the acceptance of our approach among the scientists.

PubFlow reduces the workload of the data managers, as it uses semi-automatic publication workflows based on established business workflow technologies. The provenance information for the research data processed in ”Start smart and finish wise” is in a first step assured by the smart pen and the OCN-Designer and, during the data publication process, by the CAPS framework.

As a next step, we plan to enhance the PubFlow framework, such that it can be easily adapted by different domains, and is not limited to marine science. Furthermore we plan to open the provenance archive integrated in CAPS, such that the provenance data stored in this archive can be accessed by other applications via standardized programming interfaces. Also we are developing a strategy for exporting the provenance graph to external data centers, thus provenance data and its research data are available in the same resource. More information about our approach is available through the website of the Kiel Data Management Team [12] and the PubFlow project website.[13]

References

[BH13] Peer C. Brauer and Wilhelm Hasselbring. PubFlow: a scientific data publication framework for marine science. In *Proceedings of the International Conference on Marine Data and Information Systems (IMDIS 2013)*, volume 54, pages 29–31, Lucca, Italia, September 2013.

[BHS09] Gordon Bell, Tony Hey, and Alex Szalay. Beyond the data deluge. *Science*, 323(5919):1297–1298, 2009.

[Bra14] Peer Christoph Brauer. Caps : Capturing and managing provenance information in scientific workflows. In *ZBW PHD Springschool 2014*, März 2014.

[CFS+12] Andreas Czerniak, Dirk Fleischer, Carsten Schirnick, Pina Springer, and Hela Mehrtens. The next generation of data capturing - digital ink for the data stewards of the future. In *AGU Fall Meeting 2012*, 2012.

[HTT09] Tony Hey, Stewart Tansley, and Kristin Tolle. Jim Gray on eScience: A transformed scientific method. *Fourth Paradigm, Microsoft Research Redmond, WA*, 2009.

[RvHM+08] Matthias Rohr, André van Hoorn, Jasminka Matevska, Nils Sommer, Lena Stoever, Simon Giesecke, and Wilhelm Hasselbring. Kieker: Continuous monitoring and on demand visualization of Java software behavior. In Claus Pahl, editor, *Proceedings of the IASTED International Conference on Software Engineering 2008 (SE'08)*, pages 80–85, February 2008.

[vHWH12] André van Hoorn, Jan Waller, and Wilhelm Hasselbring. 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)*, pages 247–248. ACM, April 2012.

[9] http://www.w3.org/TR/prov-o/

[10] https://www.eclipse.org/modeling/emf/

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

[12] https://portal.geomar.de/de/about-us

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