Einsatz kommerzieller und Open-Source Software für wissenschaftliche Workflows zur Datenpublikation in PubFlow

Marc Adolf & Wilhelm Hasselbring
Christian-Albrechts-Universität zu Kiel
Institut für Informatik, Arbeitsgruppe Software Engineering
{mad|wha}@informatik.uni-kiel.de

In wissenschaftlichen Arbeiten gehört es immer stärker zur Anforderung, dass Ergebnisse zusammen mit den ihnen zugrunde liegenden Daten und Datenprodukten publiziert werden. Weiterhin fördert die Veröffentlichung von Daten größere Projekte, die durch die umfangreiche Datenbasis Zusammenhänge in größeren Skalen erforschen können [1]. Um die erfassten und verarbeiteten Daten öffentlich zur Verfügung zu stellen, müssen sie in eine allgemeine Form gebracht werden, die je nach Publikationsplattform variieren kann. Auch die Datenquellen können unterschiedliche Formate für die gleiche Messaktivität bereitstellen. Der Weg von gemessenen, lokalen hin zu publizierten Daten mit einheitlichen Formaten kann stark schwanken und sehr aufwändig sein [2]. Um Wissenschaftler und Datenkuratoren in diesem Vorgang zu unterstützen, bietet PubFlow die Möglichkeit für verschieden Publikationsvorgänge feste Arbeitsabläufe, oder Workflows, zu definieren und umzusetzen [3]. Alle Vorgänge werden durch ein Ticketsystem verwaltet. Für jedes Datenpaket, das veröffentlicht werden soll, legt der Wissenschaftler, der diese Daten veröffentlichen will, ein neues Ticket an. PubFlow nutzt für die Verwaltung unter anderem JIRA (`www.atlassian.com/software/jira`) und jBPM (`http://www.jbpm.org/`. Dabei werden auch verschiedene Rollen unterstützt: Wissenschaftler, Datenkuratoren/-manager und PubFlow als technisches System. Das System bietet dabei die Möglichkeit Teile der Workflows, soweit möglich, automatisiert abzuarbeiten. Dadurch können Datenmanagern wiederkehrende und repetitive Aufgaben abgenommen werden. Idealerweise beginnen die Workflows schon mit der Datenerhebung, beispielsweise auf einem Forschungsschiff [4]. Insgesamt werden durch die automatisierten Workflows klare Abläufe geschaffen, die die Datenpublikation vereinfachen und beschleunigen können. Insbesondere da alle Verarbeitungsschritte fest angegeben sind, ist es möglich über den gesamten Vorgang auch Provenienzdaten zu sammeln [5, 6, 7]. Aktuell arbeiten wir daran, das PubFlow-System für eine bessere Skalierbarkeit und Wartbarkeit in Microservices aufzuteilen [8], insbesondere auch vor dem Hintergrund der Weiterentwicklung der zugrunde liegenden kommerziellen und Open-Source Software.

Literatur

[1] Christy, T., "Earth science: Big geochemistry", Nature 523, 293–294, 2015, doi:10.1038/523293a

[2] Fleischer, D., Janaschk, K., "A path to filled archives", Nature Geoscience 4, 575–576, 2011, doi:10.1038/ngeo1248

[3] Brauer, P. C. und Hasselbring, W., "PubFlow: a scientific data publication framework for marine science", In: International Conference on Marine Data and Information Systems (IMDIS 2013), 2013, Lucca, Italy

[4] Brauer, P. C., Czerniak, A. und Hasselbring, W., "Start Smart and Finish Wise: The Kiel Marine Science Provenance-Aware Data Management Approach", In: 6th USENIX Workshop on the Theory and Practice of Provenance (TaPP 2014), June 2014, Köln

[5] Brauer, P. C. und Hasselbring, W., "Capturing provenance information with a workflow monitoring extension for the Kieker framework", In: The Third International Workshop on the role of Semantic Web in Provenance Management , 2012, Heraklion, Kreta

[6] Brauer, P. C. und Hasselbring, W., "PubFlow: provenance-aware workflows for research data publication", In: 5th USENIX Workshop on the Theory and Practice of Provenance (TaPP '13), April 2013, Lombard

[7] Brauer, P. C., Fittkau, F. und Hasselbring, W., "The Aspect-Oriented Architecture of the CAPS Framework for Capturing, Analyzing and Archiving Provenance Data", In: 5th International Provenance and Annotation Workshop (IPAW 2014), June 2014, Köln

[8] Hasselbring, W., "Microservices for Scalability", In: 7th ACM/SPEC International Conference on Performance Engineering (ACM/SPEC ICPE 2016), March 2016 , Delft, NL