PubFlow: provenance-aware workflows for research data publication

What is PubFlow about?

PubFlow is a data publication framework for research data. This means PubFlow can be used by scientists or data managers to transfer research data from a local or institutional data repository into a World Data Center (WDC).

One of the goals of PubFlow is to create an infrastructure that alleviates the process of data publication to get research data out of the institutional repositories into public available data centers so it can be reused.

Data archival workflow

PubFlow can be integrated in different existing software systems, which are already used by many scientist and data managers. Usually PubFlow is started by creating a ticket in a ticket system like Jira. When the ticket is created, PubFlow starts the corresponding workflow and - in the case of a human task - interacts with the user through the ticket system. Finally the research data is uploaded to an archive.

Provenance information collection

PubFlow collects provenance information for the processed research data during runtime. This information is collected by Kieker, a monitoring framework developed by our working group. The provenance information collection is independent from the PubFlow system. This assures that it is not affected by errors, which might occur in PubFlow. The collected data is stored in a repository build upon W3Cs Prov-O notation and neo4j. It will be available as Google Code project soon.

The consortium

PubFlow is funded by

You can find more information about PubFlow in our paper and on our homepage.