Geophysical Research Abstracts
Vol. 16, EGU2014-6208-1, 2014
EGU General Assembly 2014

DataSync - sharing data via filesystem

Damian Ulbricht and Jens Klump
GFZ German Research Centre for Geosciences, Telegrafenberg, 14473 Potsdam, Section CeGIT

Usually research work is a cycle of to hypothesize, to collect data, to corroborate the hypothesis, and finally to publish the results. In this sequence there are possibilities to base the own work on the work of others. Maybe there are candidates of physical samples listed in the IGSN-Registry and there is no need to go on excursion to acquire physical samples. Hopefully the DataCite catalogue lists already metadata of datasets that meet the constraints of the hypothesis and that are now open for reappraisal. After all, working with the measured data to corroborate the hypothesis involves new methods, and proven methods as well as different software tools. A cohort of intermediate data is created that can be shared with colleagues to discuss the research progress and receive a first evaluation. In consequence, the intermediate data should be versioned to easily get back to valid intermediate data, when you notice you get on the wrong track. Things are different for project managers. They want to know what is currently done, what has been done, and what is the last valid data, if somebody has to continue the work.

To make life of members of small science projects easier we developed Datasync [1] as a software for sharing and versioning data. Datasync is designed to synchronize directory trees between different computers of a research team over the internet. The software is developed as JAVA application and watches a local directory tree for changes that are replicated as eSciDoc-objects into an eSciDoc-infrastructure [2] using the eSciDoc REST API. Modifications to the local filesystem automatically create a new version of an eSciDoc-object inside the eSciDoc-infrastructure. This way individual folders can be shared between team members while project managers can get a general idea of current status by synchronizing whole project inventories. Additionally XML metadata from separate files can be managed together with data files inside the eSciDoc-objects.

While Datasync's major task is to distribute directory trees, we complement its functionality with the PHP-based application panMetaDocs [3]. panMetaDocs is the successor to panMetaWorks [4] and inherits most of its functionality. Through an internet browser PanMetaDocs provides a web-based overview of the datasets inside the eSciDoc-infrastructure. The software allows to upload further data, to add and edit metadata using the metadata editor, and it disseminates metadata through various channels. In addition, previous versions of a file can be downloaded and access rights can be defined on files and folders to control visibility of files for users of both panMetaDocs and Datasync. panMetaDocs serves as a publication agent for datasets and it serves as a registration agent for dataset DOIs.

The application stack presented here allows sharing, versioning, and central storage of data from the very beginning of project activities by using the file synchronization service Datasync. The web-application panMeta-Docs complements the functionality of DataSync by providing a dataset publication agent and other tools to handle administrative tasks on the data.

[1] http://github.com/ulbricht/datasync
[2] http://github.com/escidoc
[3] http://panmetadocs.sf.net
[4] http://metaworks.pangaea.de