1. FDM-Werkstatt | 14.-16.0.2023 | KWI Essen

The 1. FDM-Werkstatt was organized by fdm.nrw.

Program

Mittwoch, 14.06.2023, ab 18:00 Uhr

- Auftaktveranstaltung

Donnerstag, 15.06.2023

- 09:00-13:00 Uhr:
  - Session 1: Using DataLad to interface RDM infrastructures
  - Session 2: Data quality assurance and detection using GitLab CI/CD
- 14:00-18:00 Uhr:
  - Session 1: Connect Datalad to Coscine Resources
  - Session 2: The coffee bean project — Best practice example for combined use of GitLab and Coscine

Freitag, 16.06.2023

- 09:00-15:00 Uhr:
  - Session 1: With metadata to better data! (Meta)Data transfer from and to Coscine
  - Session 2: One tool to rule them all: research (data) management with Emacs

Abstracts

The coffee bean project — Best practice example for combined use of GitLab and Coscine

The goal of this workshop is to create an example project that connects GitLab and Coscine as a demonstration for good research data management. The starting point is the coffee bean project in GitLab (https://gitlab.com/fdm-nrw-gitlabexamples/pythondatavisualization), which will be extended by a connection to Coscine and a comprehensive documentation. Besides the technical connection of Coscine and GitLab, files will be described with an appropriate application profile and example files will be created, described, and deposited in Coscine. Previous knowledge of the participants can span from none to very good knowledge of Coscine, GitLab and data management in general.

Person/s in charge

- Sophia Leimer (UDE)
- Henning Timm (UDE)

Proprietary binary data in research and Marble

Instruments are used in experiments and save the data in proprietary binary formats, generally. Existing export-functionality in the vendor software delivers the data in lower accuracy and without metadata, although the latter allow to better understand for instance the calibration. The goal of the event is to teach about binary data and to help people to decipher their own proprietary data. During the event we will also address legal and technical aspects of proprietary data and will use a software with graphical user interface to decipher example files.

Requirements

One should now about the background of the individual measurement that one wants to decipher (for instance, one should have done the measurement him/herself). For the event it is helpful to

have curiosity about technical things and be open to learn about them. It is not necessary to be able to code. The computer should have a python installation and one should be able to install other python packages.

Person/s in charge

- Steffen Brinkmann (FZ Jülich)

Data quality assurance and detection using GitLab CI/CD

One of the goals of the "FAIR Dataspaces" project (https://www.bildung-forschung.digital/digitalezukunft/de/technologie/daten/fair-data-spaces/fair-data-spaces.html) is to provide the quality assurance of research data. One of the results of the project is an automated workflow to analyze and transform research data and to verify data artifacts. Using csv files, this workshop teaches how to automatically analyze them in your own project using the CI/CD workflow of GitLab and output the results on a reporting page.

Person/s in charge

- Jonathan Hartman (RWTH Aachen University)

With metadata to better data! (Meta)Data transfer from and to Coscine

The Coscine research data platform (www.coscine.de) provides an API interface to transfer metadata annotated data to Coscine in automated processes. In the workshop, we will show in small-scale steps how to move data to Coscine using a JupyterNotebook (Python) and Coscine's personal authentication token, and how to specify the metadata using the application profile provided by the application. Prior knowledge of Python is desirable.

Person/s in charge

- Nicole Parks, RWTH Aachen University

Using DataLad to interface RDM infrastructures

Every field, institution, consortium or lab faces a different particular RDM challenge and therefore selects their solutions accordingly. Independent of the capabilities and maturity of each solution, this heterogeneity represents a challenge for an individual researcher or initiative, because interoperability between these different choices has to be established for efficient collaboration, or even to transition from one solution to another when the requirements or the environment changes.

DataLad is a dezentralized RDM system that is designed for interoperability with a wide range of RDM services. Using a portable, self-contained representation of a dataset that can comprise any number and any size of files, information can be precisely tracked, and shared with collaborators or stored (redundantly) across many institutional and commercial storage infrastructures.

In this workshop we demonstrate how DataLad features like automated recording of re-executable provenance records can be used to facilitate reproducible open-science. Moreover, we show how an individual researcher can iterate and collaborate on a dataset with research outcomes, and eventually deposit it for publication and long-term preservation using resources available to scientists. The workshop ends with an overview on the technical components used to achieve

DataLad's interoperability to illustrate how support for additional systems and services can be added via 3rd-party contributions.

Person/s in charge

- Michael Hanke (FZ Jülich)
- Adina Svenja Wagner (FZ Jülich)
- Stephan Heunis (FZ Jülich)

One tool to rule them all: research (data) management with Emacs

Many different technical tools are used in everyday research. This not only includes open source tools and the greater the number of tools, the worse is collaborating with colleagues. The tool GNU Emacs (www.emacs.org) is not only open source, but flexible and powerful enough to technically represent and directly support the complexity of a research (data) cycle (cf. https://www.zbmed.de/ueber-uns/profil-zbmed/forschungskreislauf/). However, it lacks concrete guidance with reference to applied research. The goal of this workshop could be to write an article explaining the stages of the research cycle using the capabilities of Emacs or to write down best practices.

Person/s in charge

- Lukas C. Bossert (RWTH Aachen Universitx)

Connect Datalad to Coscine Resources

As a group, we try to link Datalad to Coscine resources and work with the data in Coscine. For example, data is downloaded from Coscine using Datalad commands, updated, and uploaded again. These changes and processing steps are tracked and documented using Datalad.

Person/s in charge

- Nicole Parks (RWTH Aachen University) und Michael Hanke (FZ Jülich)

Using Wikidata for charting RDM

Wikidata is a free and open database for structured data on about everything – including RDM. Through a SPARQL endpoint, the data in Wikidata can be queried and displayed. In this workshop, we provide an introduction into editing and querying Wikidata. We draw on a number of examples related to RDM to illustrate possible applications and we try to expand the data on RDM in Wikidata and monitor our progress through our queries.

While tools integrated into Wikidata provide an excellent starting point, their options for data processing and visualization are somewhat limited. For participants with basic knowledge of Jupyter Notebooks we therefore offer to move the data from Wikidata to Juypter and try some more sophisticated analyses. No prior knowledge of Wikidata is required for participation. To allow us to better plan the workshop, please indicate whether you have basic skills using Jupyter.

Person/s in charge

- Matthias Fingerhuth (fdm.nrw)