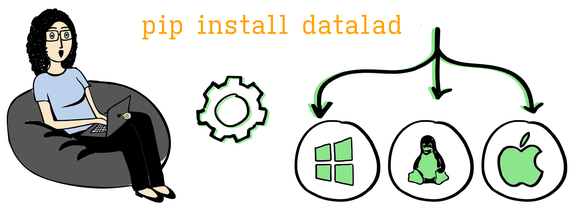
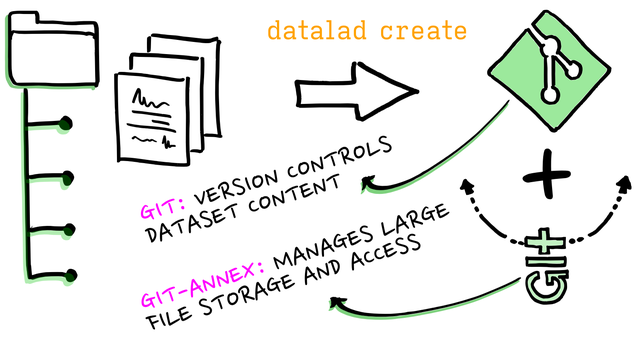
DataLad: distributed system for joint management of code, data, and their relationship

Michał Szczepanik, Christian Mönch, Michael Hanke
Institute of Neuroscience and Medicine, Brain and Behaviour (INM-7), Research Center Jülich, Jülich, Germany

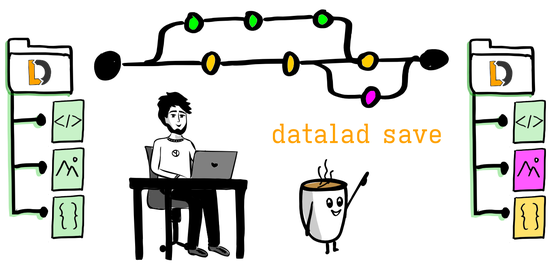
(1) DataLad[1] is a free and open-source software for data version control and provenance tracking. It is available for all major operating systems (via pip and other package managers, e.g. apt, conda & brew).



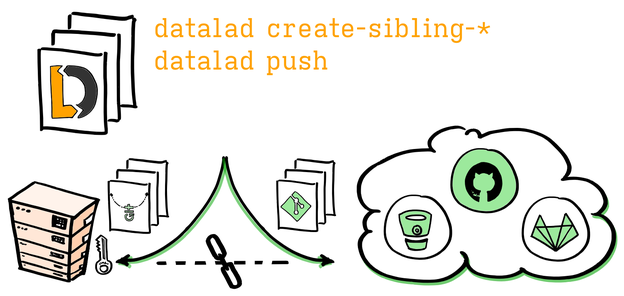
(2) DataLad datasets are modular collections of files in directories which are version-controlled with Git and git-annex. DataLad tracks files, agnostic of data formats. With git-annex, Git becomes a "filename and metadata tracker".



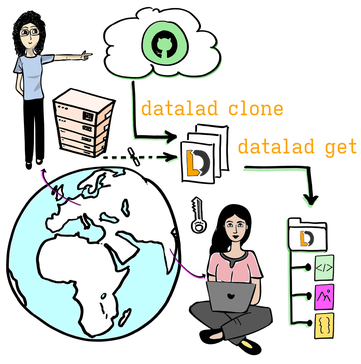
(3) Collaborate, create, branch, add, save, merge, and track: DataLad adds a range of useful concepts and functions, and provides a unified user interface, but basic workflows are based on Git.



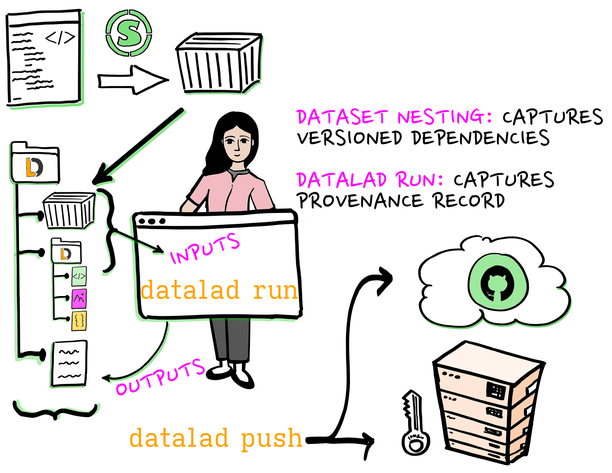
(4) DataLad is built with decentralized collaboration in mind[2]. You can combine multiple storage locations ("dataset siblings"). File content availability is machine-tracked, and retrieval is automated.



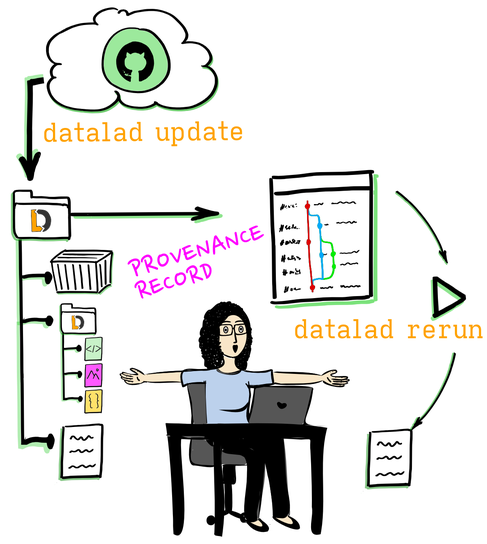
(5) With DataLad, data access starts with a lightweight clone (dataset history, naming & availability metadata). Large data content is accessed through a get operation. Credentials and encryption are supported, if needed.



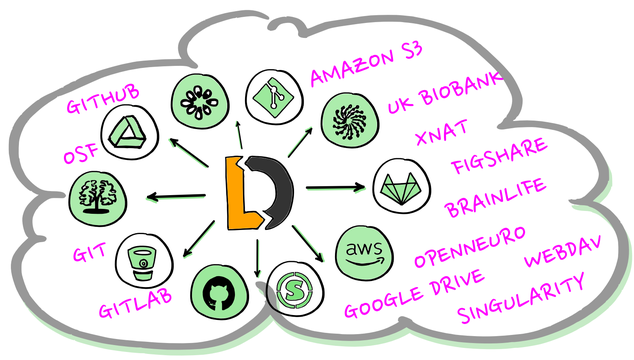
(6) DataLad provides utilities for nesting datasets and associating them with software containers to create reproducible pipelines.



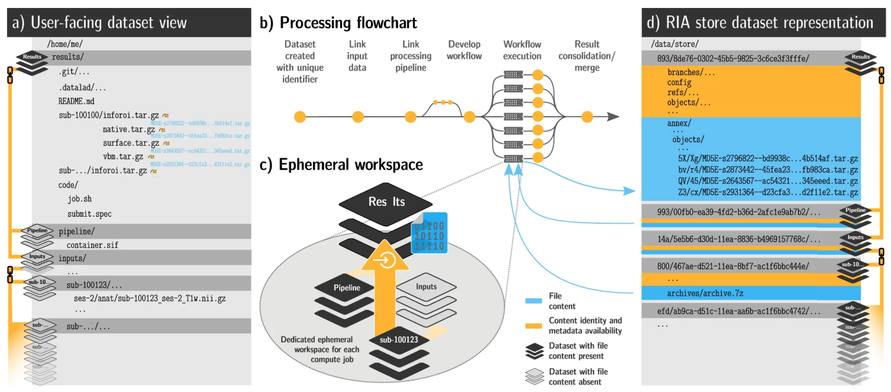
(7) DataLad can capture structured provenance records, allowing to easily rerun and reproduce the results.



(8) DataLad does not store your data, but it can integrate with various storage services. Additional integrations, as well as other functionality, can be provided with DataLad extensions (Python packages) created based on DataLad extension template.



(9) FAIRly big[3]: A template for decentralized, reproducible processing demonstrates how compute jobs can be efficiently partitioned - from clusters to PCs.



(10) Extension highlights

MetaLad extension for semantic metadata handling
docs.datalad.org/projects/metalad

DataLad Catalog: web views of dataset metadata
docs.datalad.org/projects/catalog

DataLad Gooey: GUI for basic dataset operations
docs.datalad.org/projects/gooey

datalad.org
handbook.datalad.org
docs.datalad.org
youtube.com/datalad
github.com/datalad
@datalad@fosstodon.org

Illustrations were created by Stephan Heunis.

Federal Ministry of Education and Research BMBF 01GQ1905 BMBF 01GQ1411
NIH National Institute of Biomedical Imaging and Bioengineering 1P41EB019936 -01A1
DFG SFB1451-INF
NSF NSF 1912266 NSF 1429999
Human Brain Project #H2020-945539
VirtualBrainCloud H2020-826421
European Regional Development Fund

1. Halchenko et al., (2021). DataLad: distributed system for joint management of code, data, and their relationship. Journal of Open Source Software, 6(63), 3262, https://doi.org/10.21105/joss.03262
2. Hanke M, Pestilli F, Wagner A, Markiewicz C, Poline J, Halchenko Y. In defense of decentralized research data management. Neuroforum. 2021;27(1): 17-25. https://doi.org/10.1515/nf-2020-0037
3. Wagner, A.S., Waite, L.K., Wierzba, M. et al. FAIRly big: A framework for computationally reproducible processing of large-scale data. Sci Data 9, 80 (2022). https://doi.org/10.1038/s41597-022-01163-2