# Distributed data management for large collaborative projects: DataLad ecosystem in Collaborative Research Center 1451

Michał Szczepanik, Stephan Heunis, Christian Mönch, Adina Wagner, Alexander Q. Waite, Laura Waite, Michael Hanke
Psychoinformatics Group, Institute for Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Jülich

## Collaborative Research Center

- Consortium-type funding by German Research Foundation (DFG)
- CRC1451: Key Mechanisms of Motor Control in Health and Disease
- 24 research groups: animal, human, clinical & cross-linking
- INF project goal: extend workflows to help foster a rich data collection for cross-species computational modelling
- Challenges: multi-site, no central storage storage, diverse data, heterogeneous workflows [1]
- Approach: distributed data management [2]

## What is DataLad

- Free and open source distributed data management software [3]
- Built on top of Git and git-annex
- Track file identity, availability, provenance
- Unified access to distributed data
- File identity and content are separated
- Encrypted git-annex workflow optional
- Extensible through extension packages (extension-template)
- CRC: Findability, Accessibility, Interoperability, and Reusability of data is achieved with DataLad as an overlay structure; projects retain full control about applicable standards, data sharing
- Extensively documented: handbook.datalad.org

## DataLad Next

- DataLad next: broadly-applicable additional functionality
- UX improvements
- WebDAV interface: integrate cloud storage (e.g. Nextcloud) with DataLad workflows
- Export tree mode and Git remote helper: DataLad or point&click access to archived files
- URL operations for custom & credentialled http(s) and ssh access: more flexible storage and sharing

## DataLad MetaLad

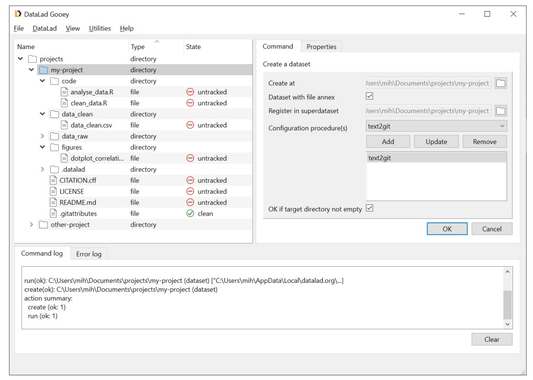
- MetaLad: DataLad extension for semantic metadata handling
- Extraction, aggregation, filtering and reporting
- Extractor-based workflows (extractor: produces metadata)
- JSON (with defined outer structure) as metadata format
- CRC approach: capture immediately, curate perpetually

## DataLad Catalog

- Generate a user-friendly web catalog from structured metadata
- Catalog: data description, not data storage
- Content is generated and displayed as a website (HTML+JS)
- Can be actionable with DataLad
- CRC: data portal, entry point to datasets


## DataLad Gooey

- Graphical User Interface for DataLad
- Make key data management tasks more accessible and more convenient, without requiring command line




## DataLad Tabby

- DataLad Tabby: tooling for tabular metadata, JSON-LD in disguise
- Format specification for describing datasets
- A list of files is equivalent to DataLad dataset without history (identity + location)
- Can be prepared in a spreadsheet
- Translates to JSON-LD, with complexity hidden away

## Useful infrastructure

- G-node GIN: repository store with Git & git-annex support (large file content, DOI service) hosted by LMU Munich [4]
- Sciebo: Nextcloud-based regional academic cloud service (for North Rhine-Westphalia, Germany)

**References**

[1] Mittal et al. (2023), Data management strategy for a collaborative research center https://doi.org/10.1093/gigascience/giad049
[2] Hanke et al. (2021), In defense of decentralized research data management, https://doi.org/10.1515/nf-2020-0037
[3] Halchenko et al. (2021), DataLad: distributed system for joint management of code, data, and their relationship, https://doi.org/10.21105/joss.03262
[4] Kalantari et al (2023), How to establish and maintain a multimodal animal research dataset using DataLad, https://doi.org/10.1038/s41597-023-02242-8

Funded by
Deutsche Forschungsgemeinschaft
German Research Foundation
(431549029-INF)