INF

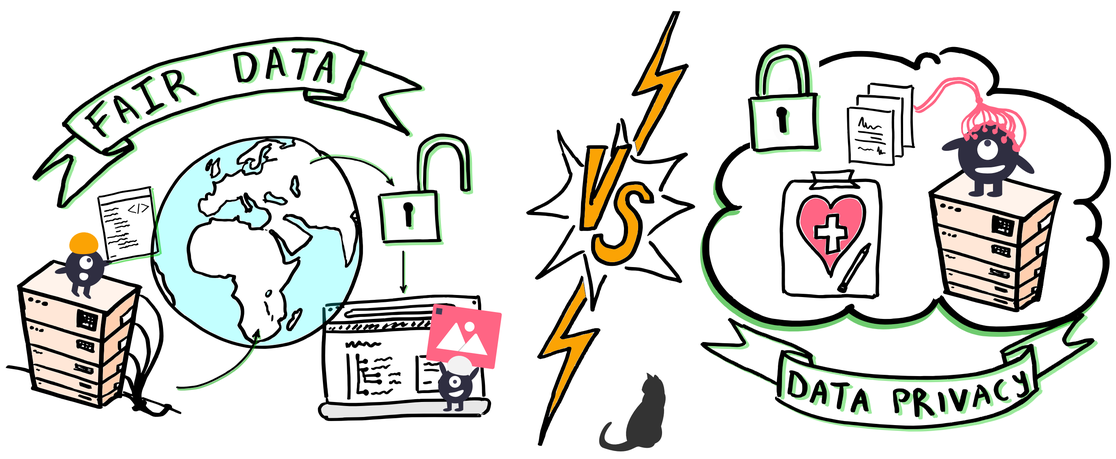
Distributed Research Data Management in CRC 1451

rdm.sfb1451.de

Michał Szczepanik[1], Stephan Heunis[1], Christian Mönch[1], Adina Wagner[1], Michael Hanke[1]
[1]Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Forschungszentrum Jülich

The privacy challenge

The importance and benefits of making research data findable, accessible, interoperable, and reusable (FAIR) are clear[1]. But of equal importance is our ethical and legal obligation to protect the personal data privacy of research participants.

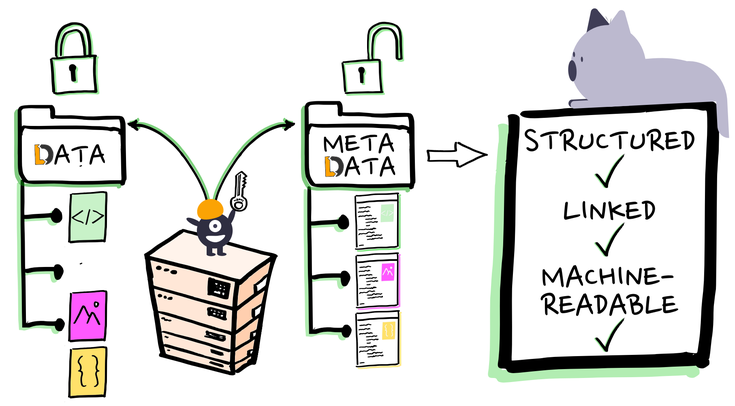


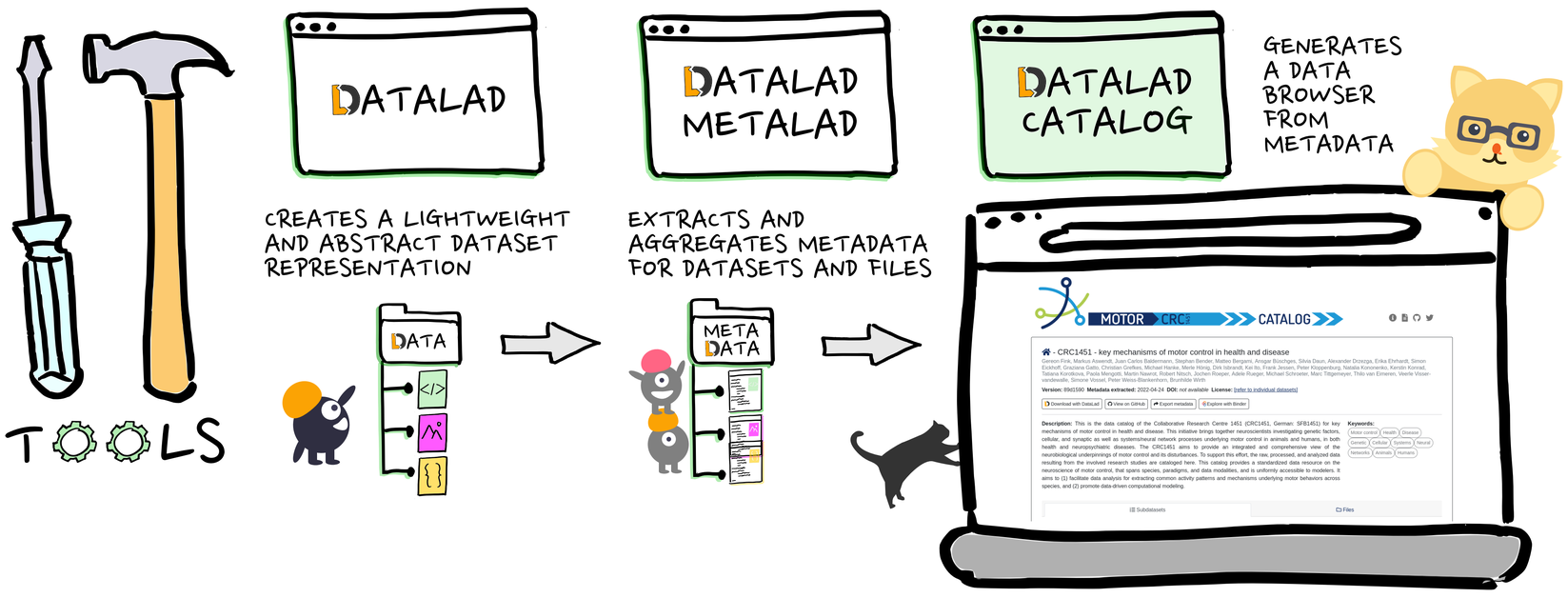
The decentralization challenge

The CRC1451 is comprised of multiple teams, each using their own workflows and storage solutions. The data is collected and processed at different sites. There is no central storage for archival; the data lives in various places, some open and some closed. But the decentralization can actually be advantageous[2], as long as there is a common layer of for exchanging availability information - the more automated, the better.

The opportunity

Structured, linked, and machine-readable metadata presents a powerful opportunity to address these challenges. Metadata provides both high-level information about our research data (such as study design, participant information, and data acquisition parameters) and lower-level descriptive aspects of each file in the dataset (such as filenames, relative paths, sizes, and formats). With this metadata, we can create an abstract representation of a full dataset that is separate from the actual data content. This means that the content can be stored securely to maintain privacy, while the metadata can be shared openly to comply with FAIR principles.

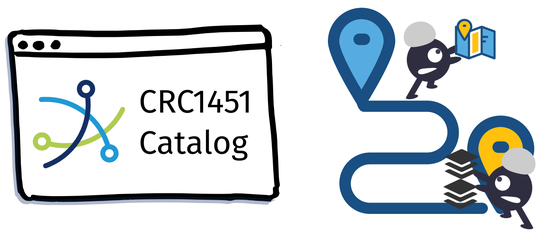




The toolset

These principles are achievable in practice with a free and open source toolset. DataLad[3] software can be used for decentralised management of data as lightweight, portable and extensible representations. Datalad-metalad can extract structured high- and low-level metadata and associate it with datasets or with individual files. And at the end of the pipeline, datalad-catalog can turn the structured metadata into a user-friendly data browser!

Central access to distributed data



The catalog provides a landing point for the consumer. It displays metadata: authors, modality, species, tasks, number of subjects, subdatasets, file names… + an actionable URL for retrieval.

Each dataset may be in a different place, but the location is also encoded with the metadata. DataLad retrieves file content, if access is possible.

Distributed storage

The publisher can choose the most suitable kind of storage, and decide on access rights.

Local storage. Unavailable from the outside. Only the metadata (including file availability information) is exposed. E.g. local NAS, hard drives on a shelf.

Cloud storage, private access. While the metadata is publically available, automated access to file content requires credentials. E.g. **Dracoon**, Sciebo, GIN, Amazon S3, Backblaze B2.

Cloud storage, public access. For anybody, getting file content is as simple as: `datalad clone & datalad get`.

DataLad Support

Virtual office hours every Thursday at 16:00 are a great opportunity to discuss problems and solutions with members of the Psychoinformatics Group at FZJ.

Direct contact (m.szczepanik@fz-juelich.de) to discuss use cases is very welcome.

We held two ECR workshops on DataLad. A write-up of the **workshop materials** is online.

For particular bugs, problems, and feature requests, opening a new issue in **DataLad's GitHub** Issue section is the best way of bringing it to the attention of DataLad developers. A lot of development is user-driven.

General user support and a growing knowledge base on DataLad-based solutions is provided on **Neurostars.org** forum under datalad tag (category).

References
1 Wilkinson et al., (2016). DOI: 10.1038/sdata.2016.18
2 Hanke et al., (2021). DOI: 10.1515/nf-2020-0037
3 Halchenko et al., (2021). DOI: 10.21105/joss.03262

2nd CRC1451 Annual Retreat
Moers, 1-3 June 2022

get poster PDF

Funded by
Deutsche Forschungsgemeinschaft
German Research Foundation