DATA ... LAD

Michael Hanke & Adina Wagner

JÜLICH Forschungszentrum

hhu Heinrich Heine Universität Düsseldorf

USE CASES IN DATA STORAGE AND RETRIEVAL

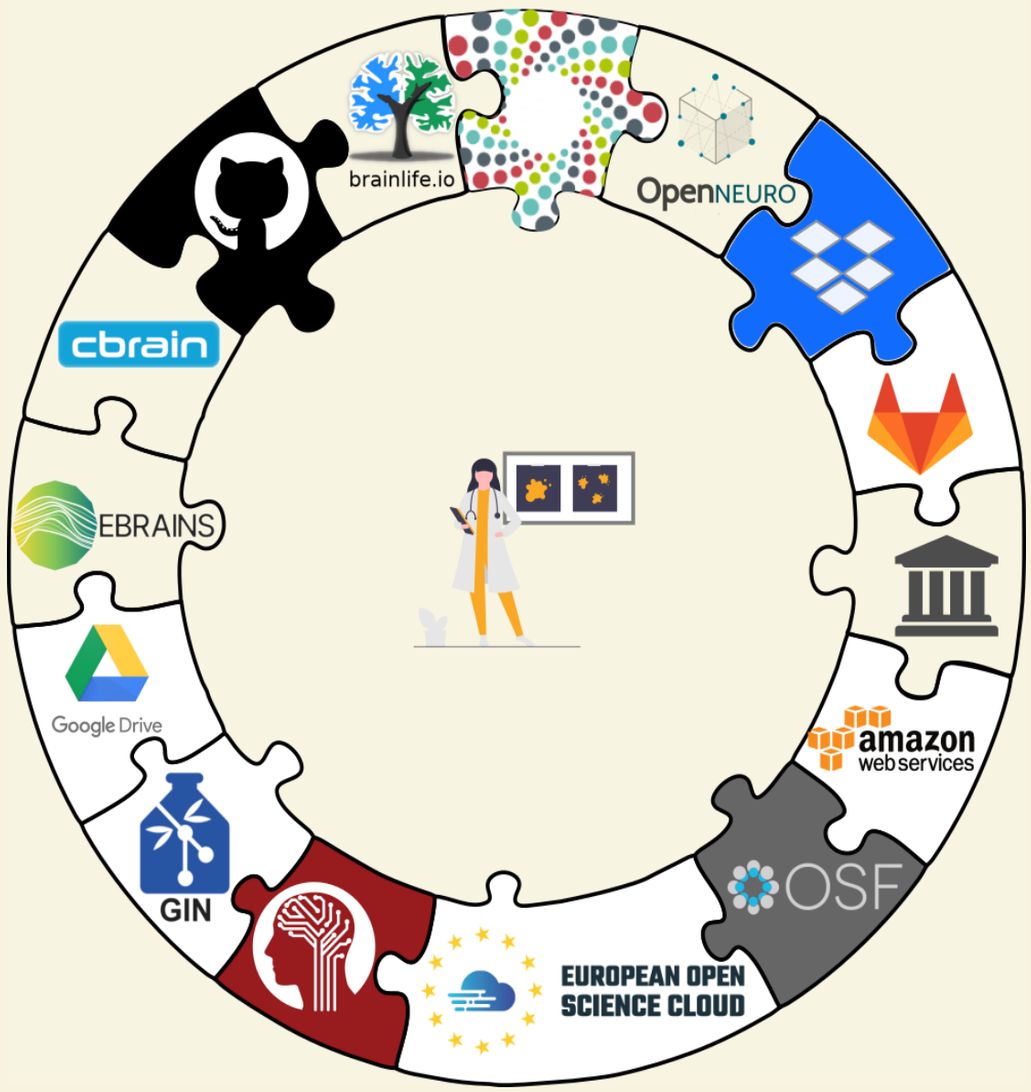
- DataLad: Joint management of digital objects through their entire life cycle

  DataLad's features,

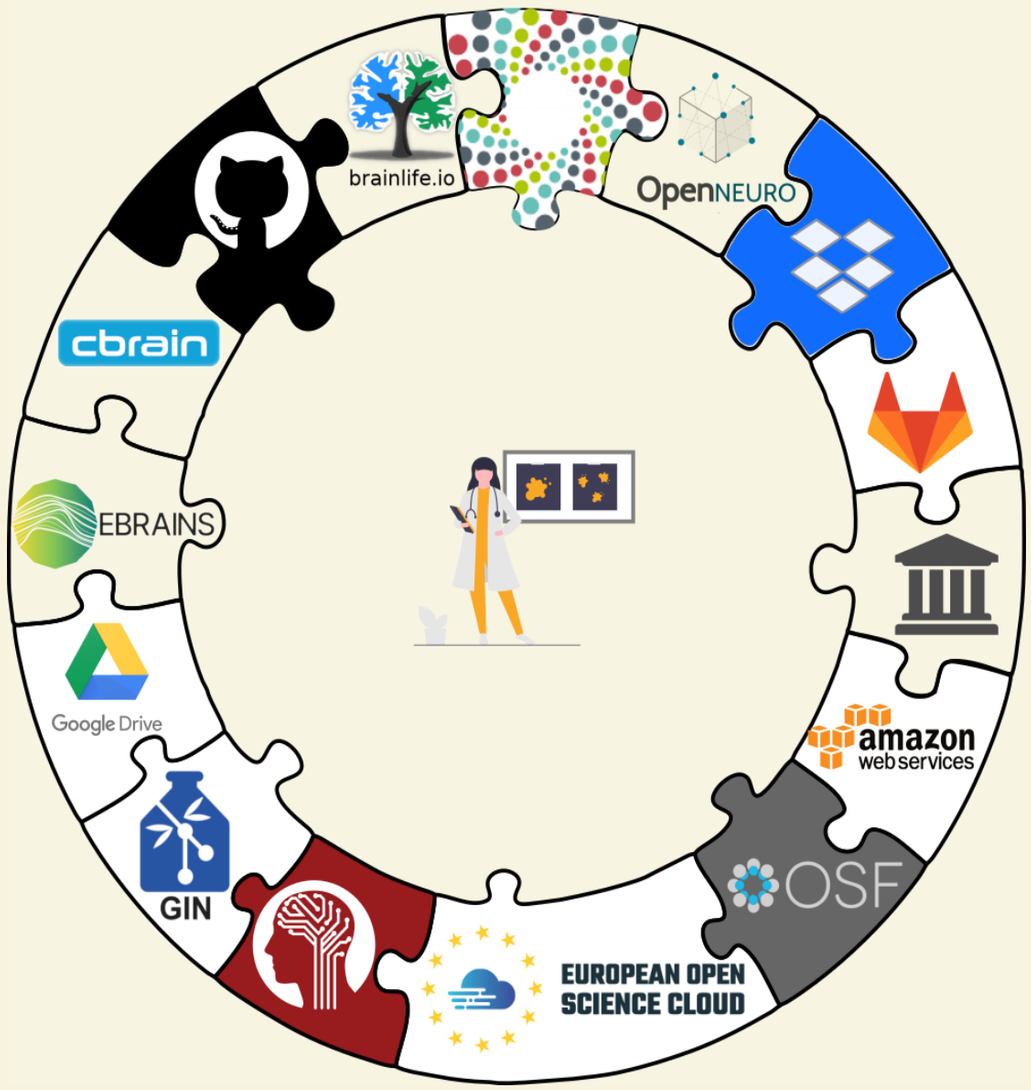
  - Version control
  - Transport logistics
  - Interoperability
  - Provenance capture

  enable a range of common use cases for storing and retrieving data

DATA SHARING & CONSUMPTION

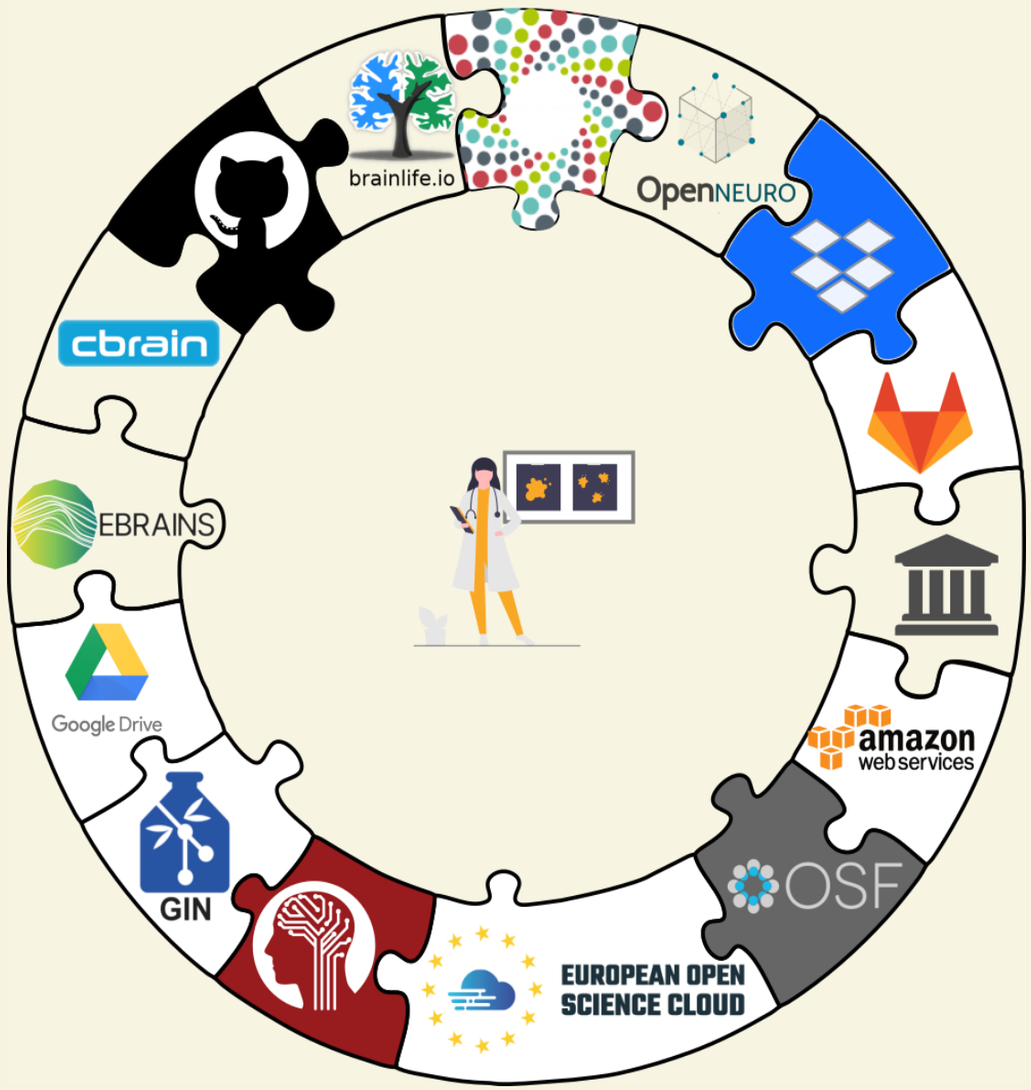


DATA SHARING & CONSUMPTION



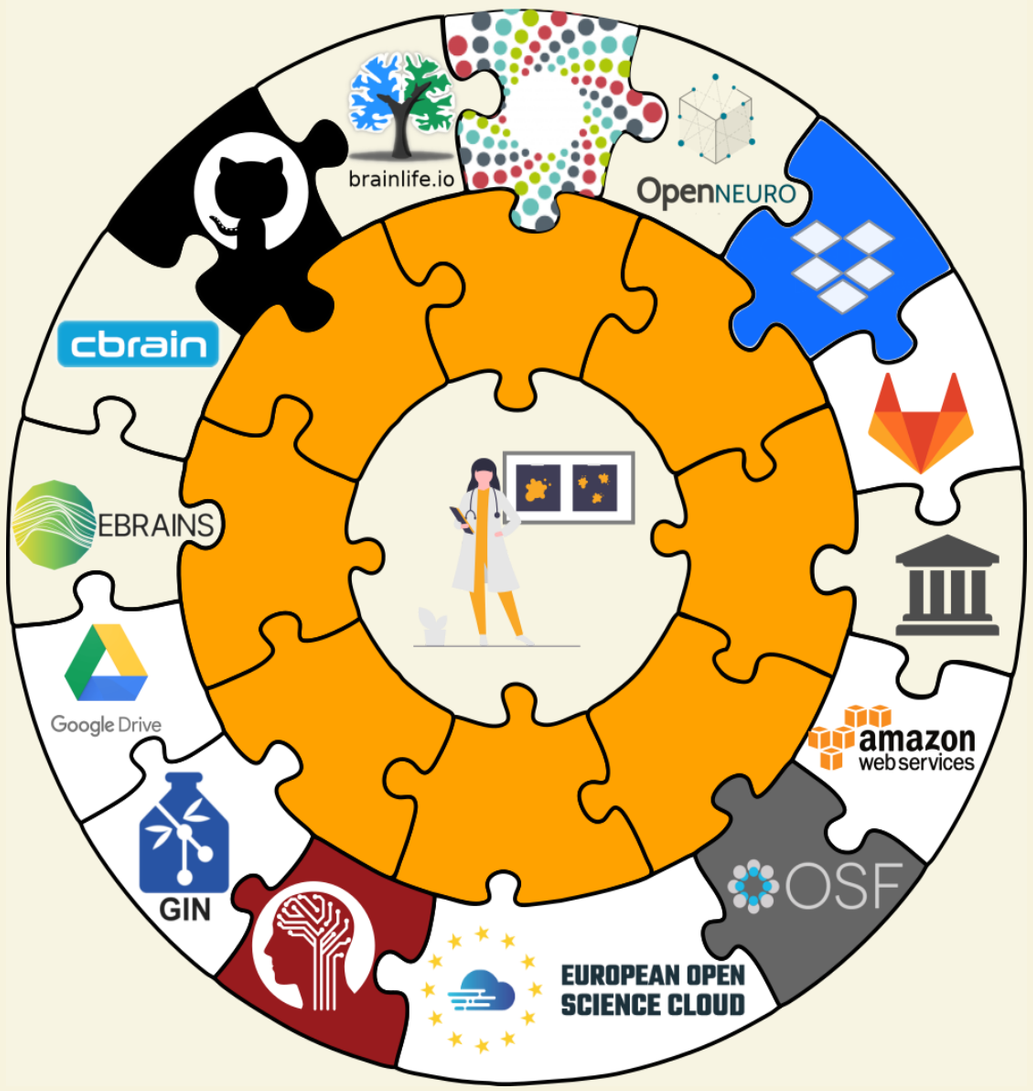
- Many options for sharing and storing data

DATA SHARING & CONSUMPTION



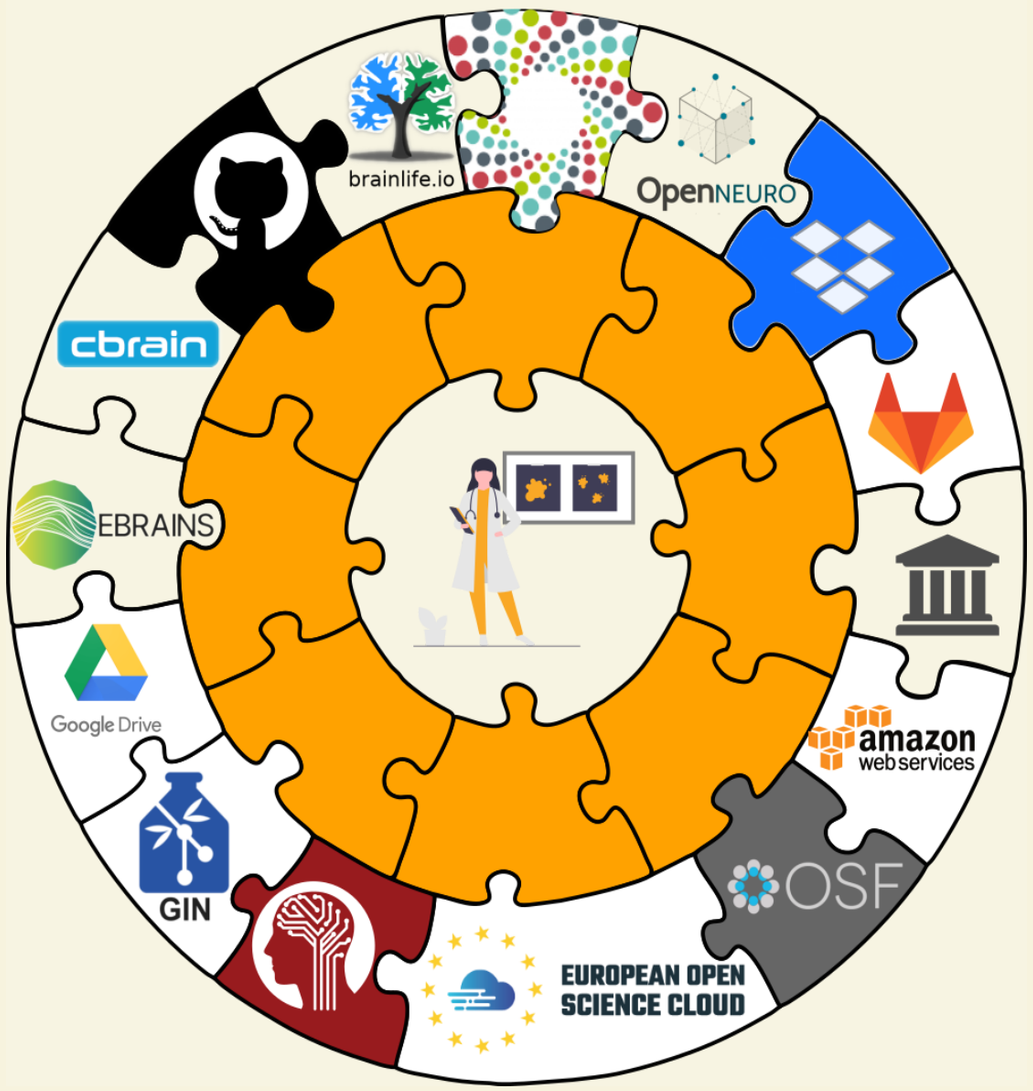
- Many options for sharing and storing data
- Often university-/lab-wide standards, but lack of generic and interoperable workflows

DATA SHARING & CONSUMPTION



- Generic interface to a variety of third party services & storage providers

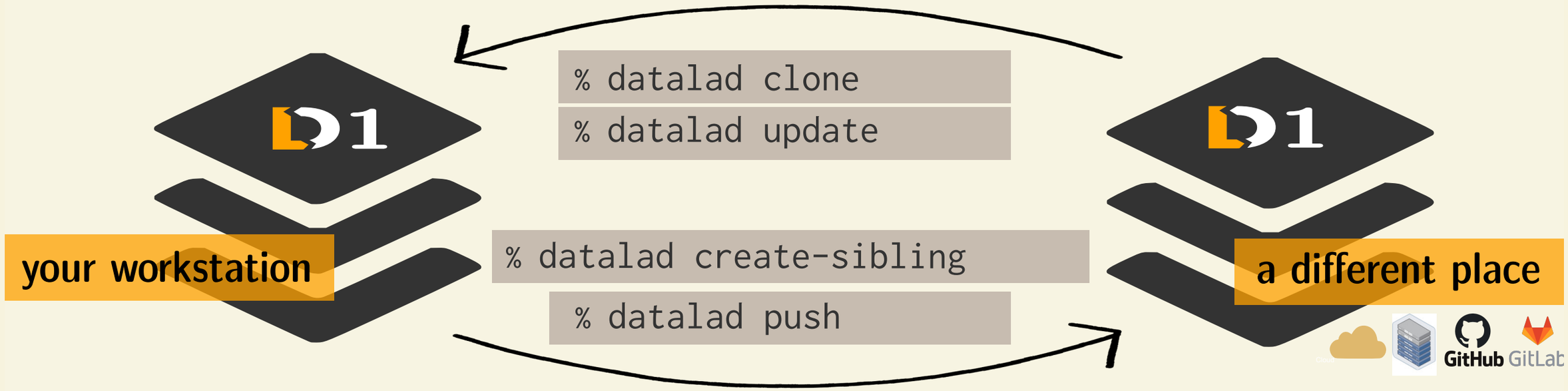
DATA SHARING & CONSUMPTION



- Generic interface to a variety of third party services & storage providers
- Enables interoperable workflows independent of the chosen service

DATA SHARING & CONSUMPTION

Consume existing datasets and stay up-to-date



Create sibling datasets to publish to or update from

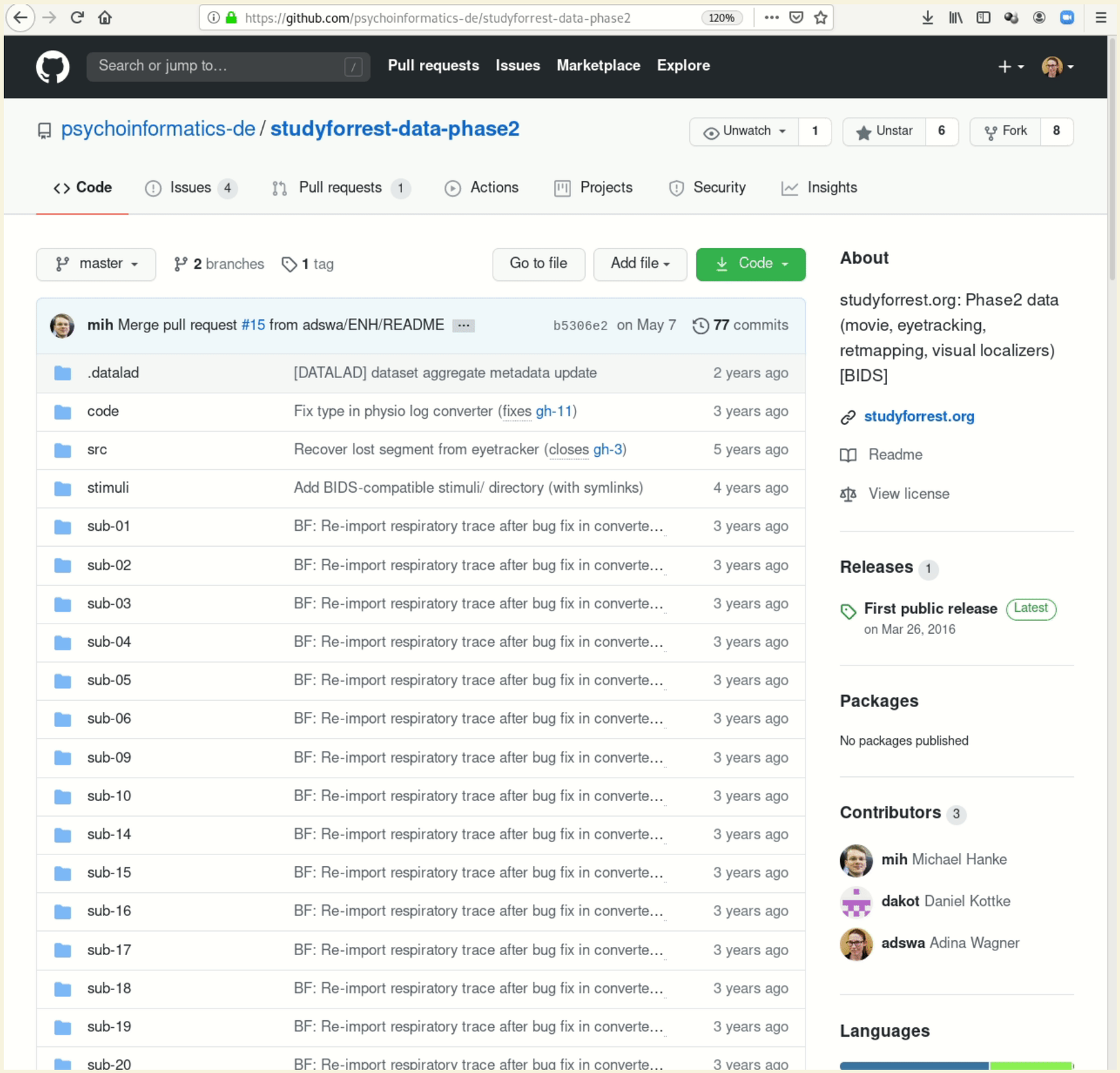
DATA SHARING & CONSUMPTION

DATA SHARING & CONSUMPTION

- **Publish (or consume) datasets** via GitHub, GitLab, Gin, OSF, or similar services

DATA SHARING & CONSUMPTION

- **Publish (or consume) datasets** via GitHub, GitLab, Gin, OSF, or similar services



DATA SHARING & CONSUMPTION

- **Publish (or consume) datasets** via GitHub, GitLab, Gin, OSF, or similar services

dkp / sub-219_bids_datalad

Watch 1 | Star 0 | Fork 0

Files | Issues 0 | Pull Requests 0 | Wiki

This is a BIDS MRI dataset generated from sub-219 (me), so no privacy issues. I used heudiconv with the datalad flag to generate it.

7 Commits | 4 Branches | 0 Releases

Branch: master | sub-219_bids... | GIN HTTPS SSH git@gin.g-node.org:/dkp/sub-

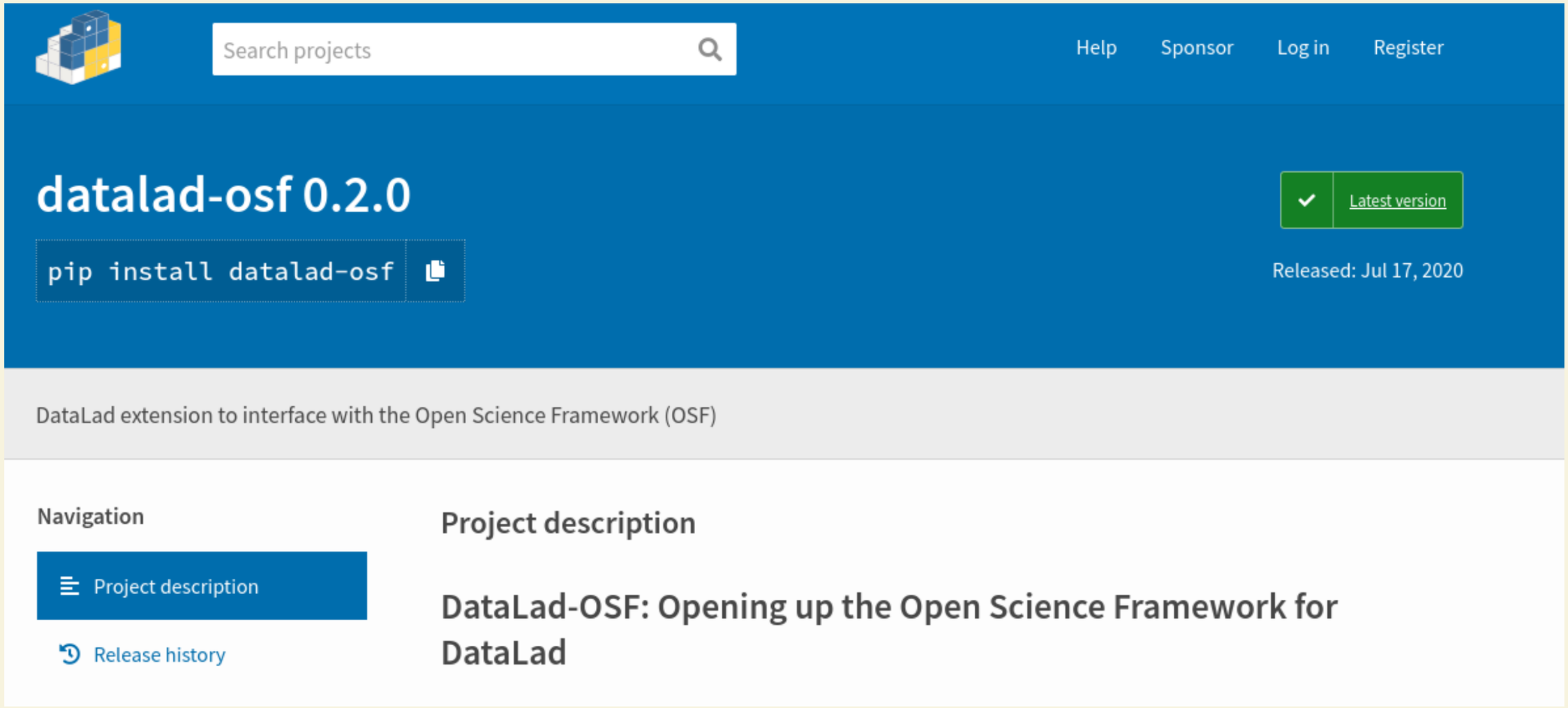
root	0ef0b644b0	removed derivatives	3 months ago
.datalad	eff691fb27	[DATALAD] new dataset	15 years ago
.heudiconv @ 7cbcd6df26	41bf012430	Converted subject 219, session itbs, 1857 dicoms	15 years ago
code	eb0468638b	Preparing for 219, session itbs, 1857 dicoms	15 years ago
sub-219	41bf012430	Converted subject 219, session itbs, 1857 dicoms	15 years ago
.gitattributes	d8d9540de6	Custom .gitattributes	15 years ago
.gitmodules	0ef0b644b0	removed derivatives	3 months ago
CHANGES	eb0468638b	Preparing for 219, session itbs, 1857 dicoms	15 [illegible] ago
README	eb0468638b	Preparing for 219, session itbs, 1857 dicoms	
dataset_description.json	eb0468638b	Preparing for 219, session itbs, 1857 dicoms	
participants.tsv	eb0468638b	Preparing for 219, session itbs, 1857 dicoms	
task-rest_bold.json	41bf012430	Converted subject 219, session itbs, 1857 dicoms	

README

gin.g-node.org

DATA SHARING & CONSUMPTION

- **Publish (or consume) datasets** via GitHub, GitLab, Gin, OSF, or similar services



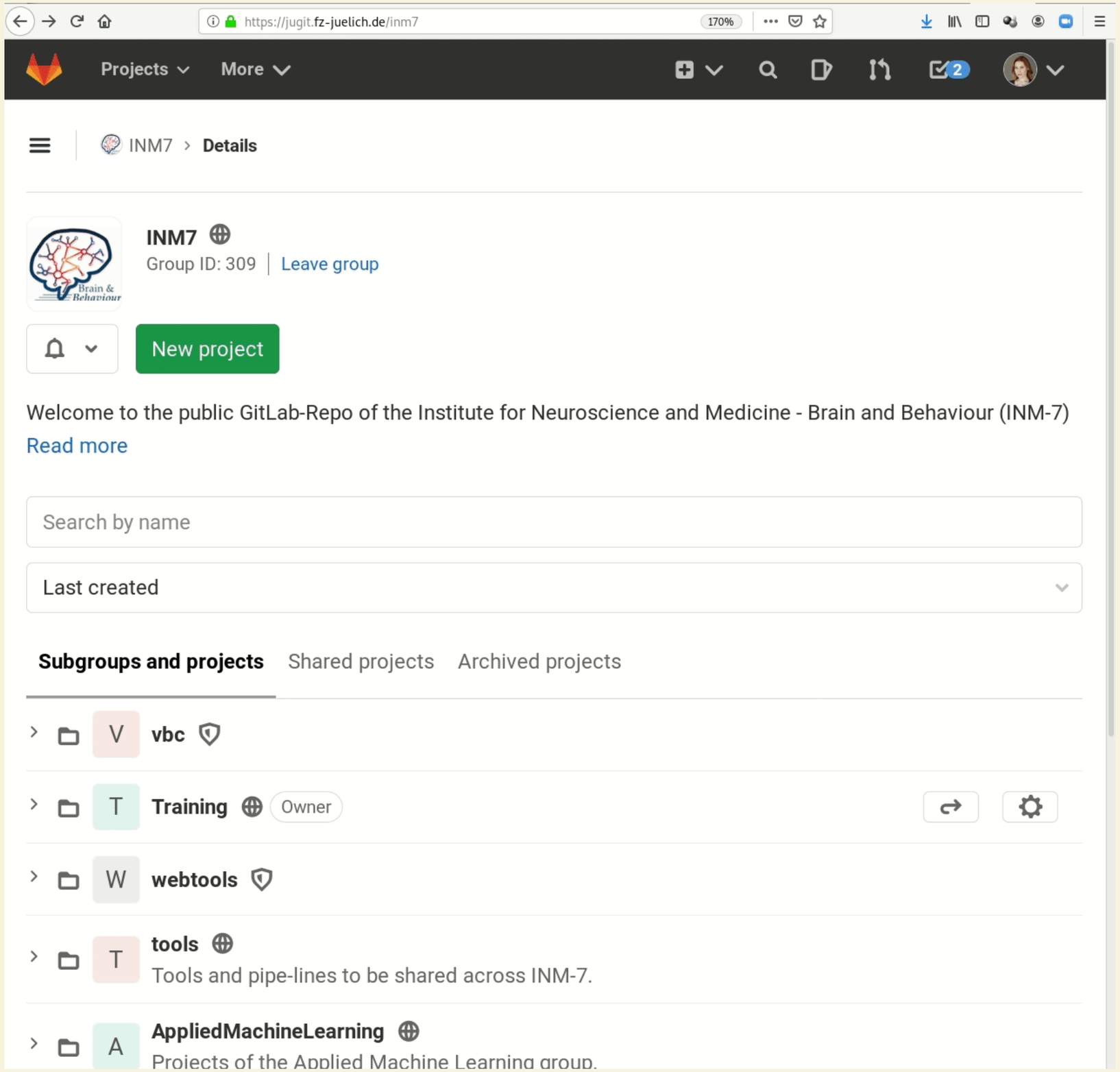
DATA SHARING & CONSUMPTION

DATA SHARING & CONSUMPTION

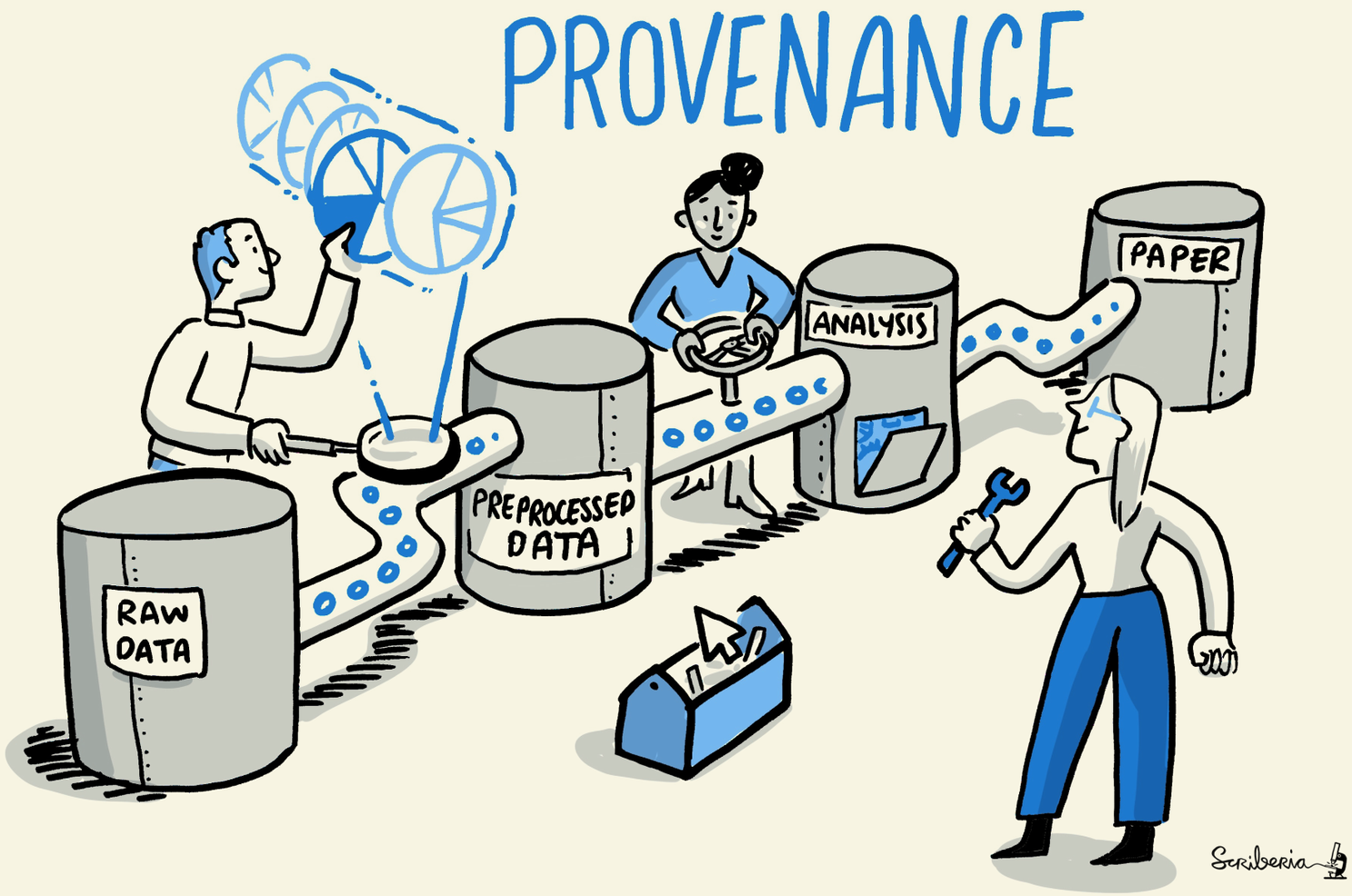
- Special Case: **Central data management** and archival system

DATA SHARING & CONSUMPTION

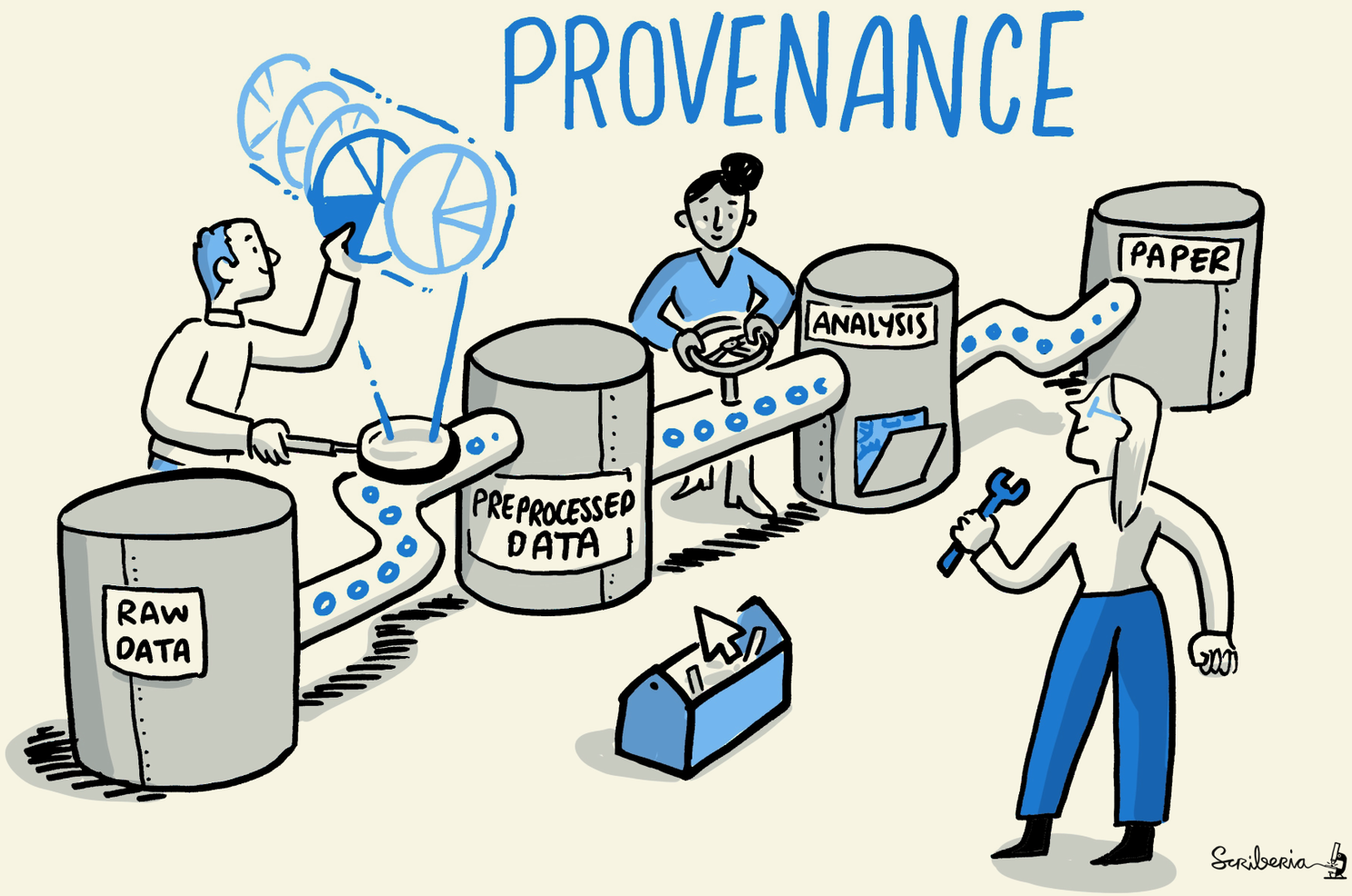
- Special Case: **Central data management** and archival system



BEYOND DATA: PROVENANCE

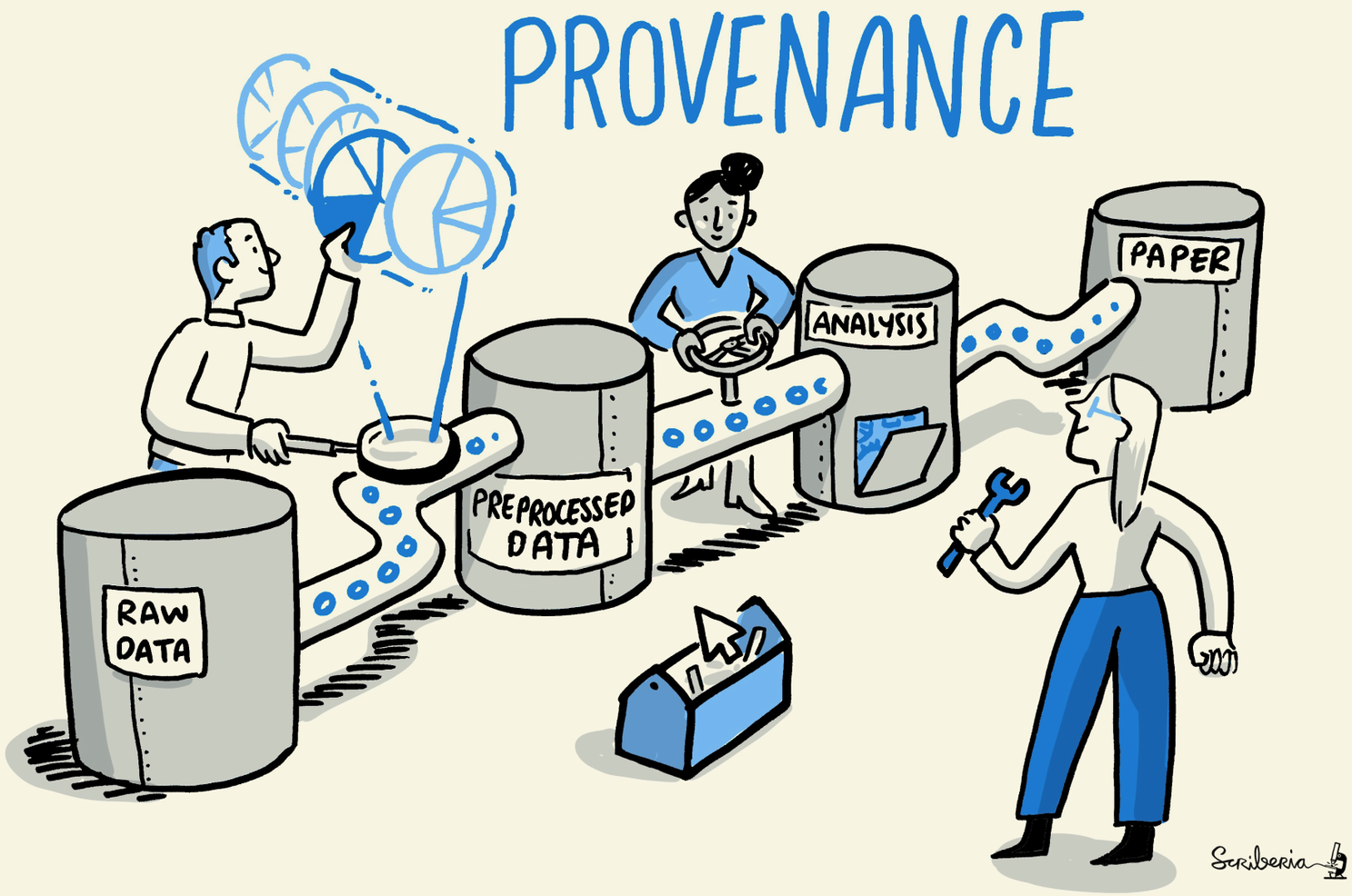


BEYOND DATA: PROVENANCE



- Create, share & use provenance of research objects

BEYOND DATA: PROVENANCE



- Create, share & use provenance of research objects
- Link code, data, software container and re-executable analysis records

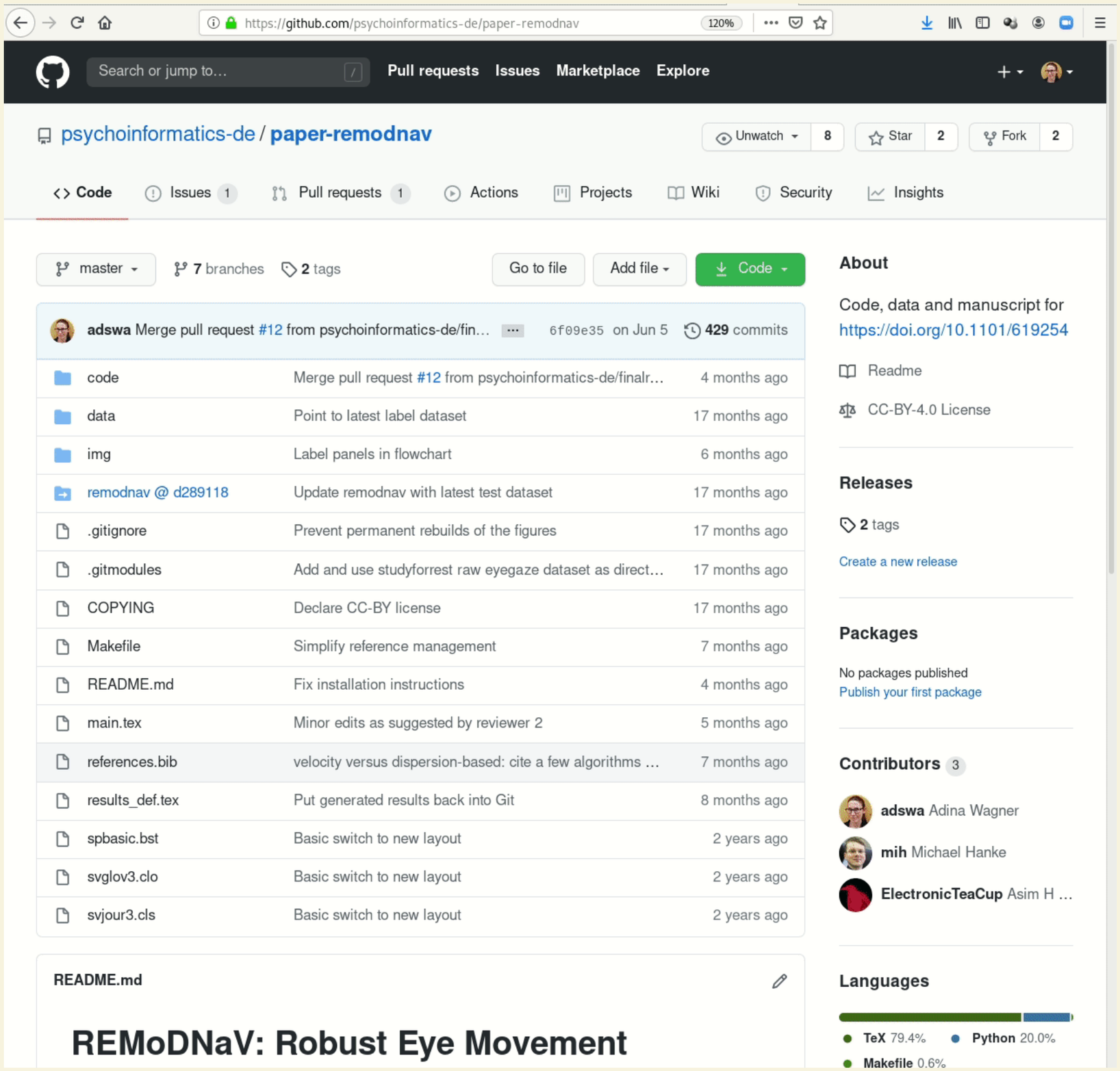
BEYOND DATA: PROVENANCE

BEYOND DATA: PROVENANCE

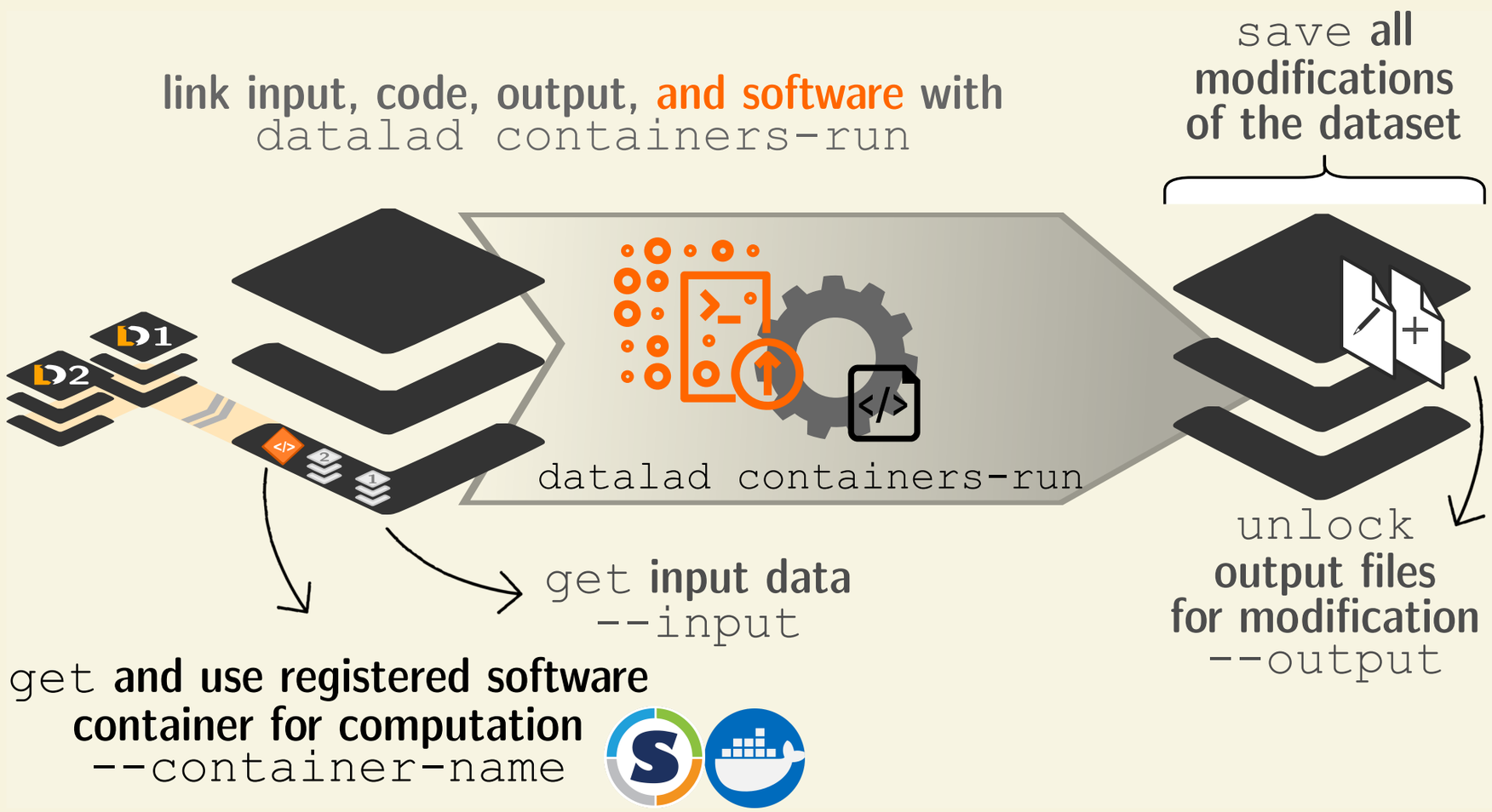
- **Creating and sharing reproducible, open science**: Sharing data, software, code, and provenance

BEYOND DATA: PROVENANCE

- **Creating and sharing reproducible, open science**: Sharing data, software, code, and provenance



CONTAINERIZED WORKFLOWS



PROVENANCE CAPTURE

- **Computational provenance**: Datasets can track **software containers**, and perform and record computations inside it:

```
$ datalad containers-run -n neuroimaging-container \
  --input 'mri/*_bold.nii --output 'sub-*/LC_timeseries_run-*.csv' \
  "bash -c 'for sub in sub-*; do for run in run-1 ... run-8;
     do python3 code/extract_lc_timeseries.py \$sub \$run; done; done'"

-- Git commit -- Michael Hanke < ... @gmail.com>; Fri Jul 6 11:02:28 2019
    [DATALAD RUNCMD] singularity exec --bind {pwd} .datalad/e...
    === Do not change lines below ===
    {
     "cmd": "singularity exec --bind {pwd} .datalad/environments/nilearn.simg bash..",
     "dsid": "92ea1faa-632a-11e8-af29-a0369f7c647e",
     "inputs": [
      "mri/*.bold.nii.gz",
      ".datalad/environments/nilearn.simg"
     ],
     "outputs": ["sub-*/LC_timeseries_run-*.csv"],
     ...
    }
    ^^^ Do not change lines above ^^^
---
 sub-01/LC_timeseries_run-1.csv | 1 +
...
```

PROVENANCE CAPTURE

- All recorded transformations can be re-computed automatically

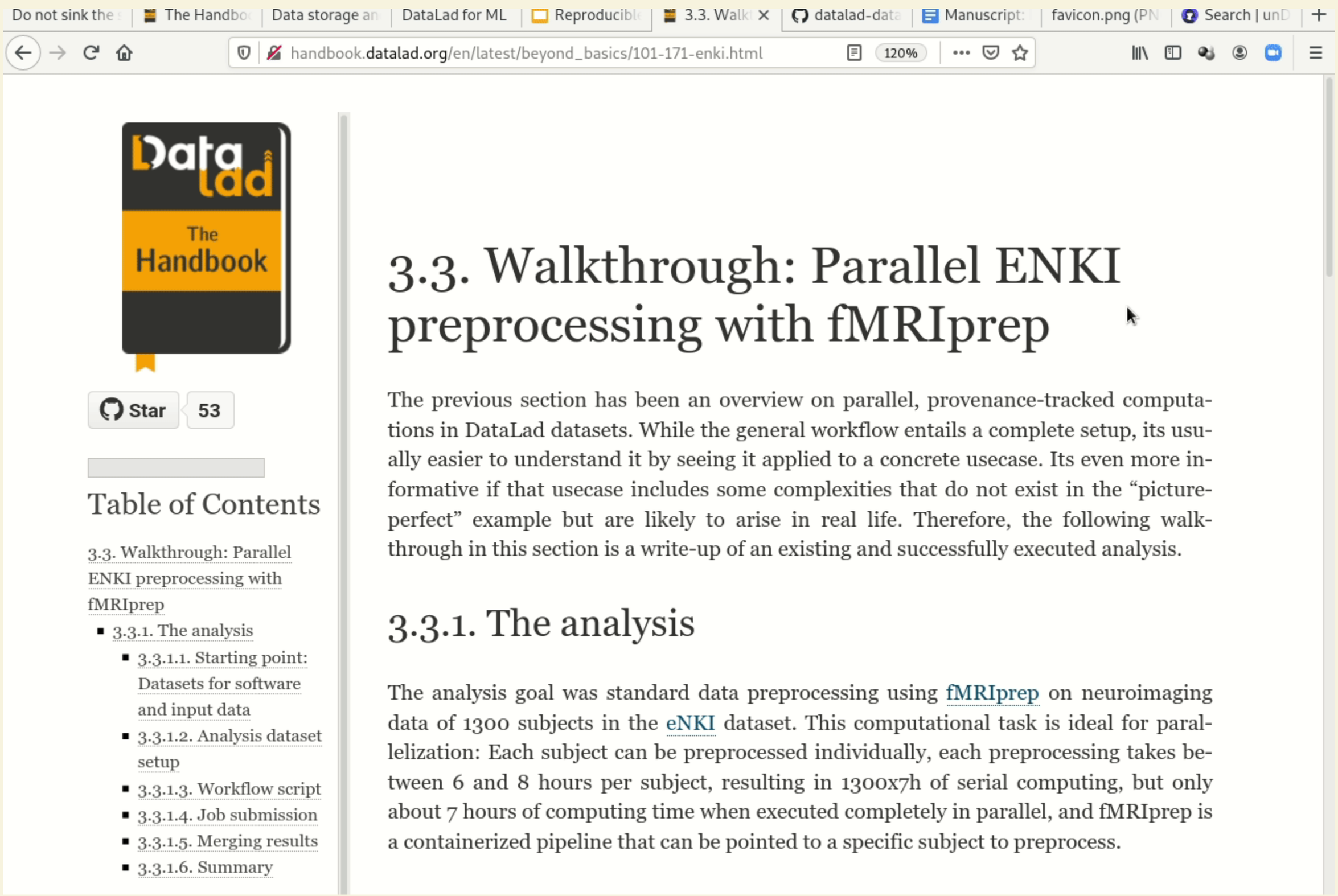
```
$ datalad rerun eee1356bb7e8f921174e404c6df6aadcc1f158f0
[INFO] == Command start (output follows) =====
[INFO] == Command exit (modification check follows) =====
add(ok): sub-01/LC_timeseries_run-1.csv (file)
...
save(ok): . (dataset)
action summary:
  add (ok: 45)
  save (notneeded: 45, ok: 1)
  unlock (notneeded: 45)
...
```

  - Aid with the reproducibility of a result and verify it (via content hash)
  - Use complete capture and automatic re-computation as alternative to storage and transport

CONTAINERIZED WORKFLOWS

```
(master) adina@juseless in /data/group/psyinf/HCP_structural on git:master
❯ datalad containers-run -n .tools/fmriprep \                                  1 !
   --input '.source/sub-170631' \
   --output '.' \
   ".source . participant --participant-label 170631 --fs-license-file .tools/license.txt --skip-bids-validation --anat-only -w /tmp"
```

LARGE-SCALE, CONTAINERIZED WORKFLOWS



CONCEPTS: DATA SHARING WITHOUT PRIVACY BREACH

- Expose (sensitive) file content versus anonymized metadata on a per-file basis

Exposed (anonymized)

VERSION HISTORY

FILE NAMES

SALTED FILE CONTENT HASHES

```
MD5E-s3016749--64364d53b5f41fbe5bccfe5e.jpg   patient_record/photo.jpg
MD5E-s3016749--8ccc1989f2ab223874428092.json  patient_record/insurance.json
MD5E-s932874--ad9e369d8c908feb6d3fe84e.json   patient_record/visit/1.json
MD5E-s762273--1ec751b8cb432d16f6a39e64.json   patient_record/visit/2.json
MD5E-s30983443--30dcd89ecfd5dd546f8ccc19.img  patient_record/imaging/brain/20031.img
MD5E-s73826423--9432d12b1581f220c8b75953.img  patient_record/imaging/thorax/20201.img
```

METADATA

```
{"type": "patient", "id": "122bf-342-422",
 "imaging": ["brain", "thorax"], "sex": "male",
 "diagnosis": ["F51", "G12", "G52"], "age": 52...}
```

ANON IDS

Tracked

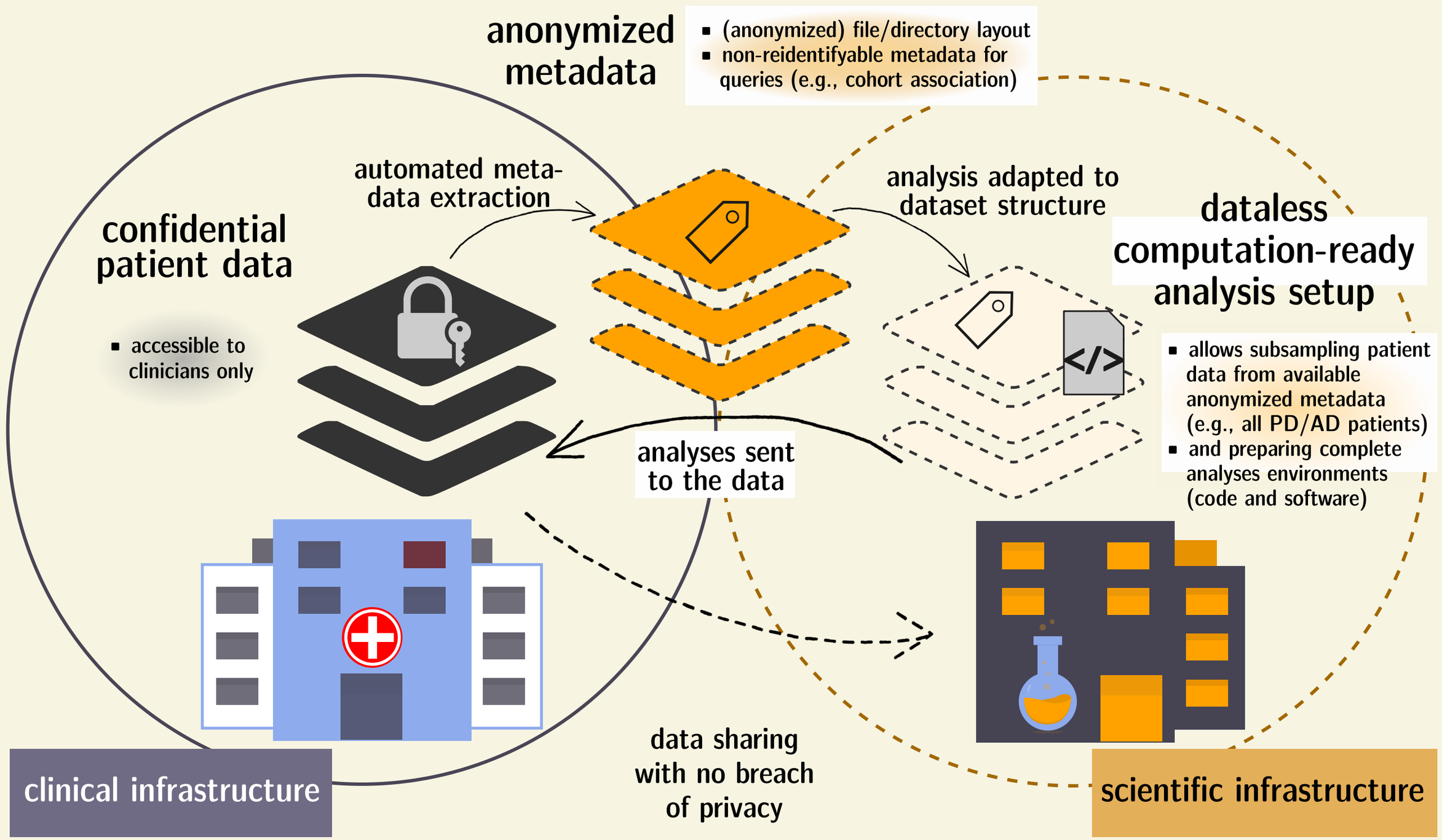
PRISTINE INFORMATION

```
{"account": "293924",
 "id": "a399b30033d"...}
{"date": "2020-04-21",
 "diagnosis": "...",
 "notes": "..." ...}
```

Authorized/local access only

CONCEPTS: COMPUTATION TO DATA

- Expose sufficient anonymized metadata to allow code development, but don't share raw data, only results



FIND OUT MORE

More use cases & comprehensive user documentation in the DataLad Handbook (handbook.datalad.org)

- High-level function/command overviews, Installation, Configuration, Cheatsheet

---

- Narrative-based code-along course
- Independent on background/skill level, suitable for data management novices

---

- Step-by-step solutions to common data management problems