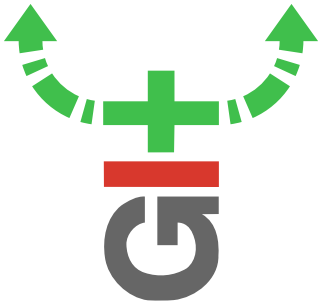


http://git-annex.branchable.com

git-annex

is a distributed version control system for large files developed by Joey Hess. git-annex enriches regular git repositories with meta-information and facilities for retrieving and/or sending data from/to remote locations. Actual data itself is not directly committed to git, making git-annex repositories lightweight and easily manageable, while still providing access to all data upon request.

git-annex features

Data versioning
Every file is identified by its checksum (*e.g.*, SHA256), thus guaranteeing unambiguous versioning

Special remotes
Besides accessing data from other git remotes (*e.g.*, via ssh), data can be offloaded to the cloud (*e.g.*, Amazon S3 and Glacier), or just some remote computer (via ssh or rsync)

Association with web origins
Every file can have a list of URLs from which it can be downloaded

Encryption
Data can be uploaded to remotes encrypted, guaranteeing privacy even when offloading sensitive data

Synchronization
git-annex assistant provides Dropbox-like operation: keep your data files in sync, or automatically offloaded, backed-up, *etc.* across multiple hosts

Semi-compatible with GitHub
As a git-annex repository is just a git repository, it can be pushed to GitHub. The content of data files, however, need to be stored in some *Special remote* or have associated download URLs

git-annex 101

Just learn few new commands in addition to stock *git*. https://git-annex.branchable.com/walkthrough provides more details and examples

`git init; git annex init "repo name"`
Initialize both git and git-annex

`git annex add my_cool_big_file`
Add file(s) under control of git-annex

`git commit -a -m "message"`
Commit as you usually do with git

`git annex move my_cool_big_file --to usbdrive`
Offload the file to a repository you named "usbdrive"

`git annex whereis my_cool_big_file`
Check where file is available from

`git annex get my_cool_big_file`
Fetch file back to the local repository

Neuroimaging datasets to play with

Any dataset on `http://datasets.datalad.org` and some others you can `git clone ...`

`http://psydata.ovgu.de/forrest_gump/.git/`
Unique 7T MRI, fMRI, DTI dataset acquired during rich auditory stimulation. See http://studyforrest.org for more information

`git://github.com/datalad/nih--videocast`
git-annex "mirror" of http://videocast.nih.gov, created and updated by `datalad crawl`

`http://data.pymvpa.org/datasets/haxby2001/.git/`
Seminal work by Haxby et al. (2001) for testing and demonstrating MVPA techniques

How to install git-annex

```
apt-get install git-annex-standalone
```

from NeuroDebian, and see https://git-annex.branchable.com/install for more (OS X, Windows, Linux, Android)

How to get support

On Debian systems
`reportbug git-annex`

Community support
http://git-annex.branchable.com/bugs

IRC
#git-annex at OFTC network

http://datalad.org

DataLad

aims to simplify and thus facilitate discovery, delivery, management, and sharing of scientific data by establishing a federated data distribution. While initially aiming to deliver public neuroimaging datasets, DataLad itself is domain agnostic and could be used for any other field of endeavor.

DataLad FAQ

Federated?
It is impractical to distribute data through *classical* distribution mechanisms, where content is contained within packages available from the central location (or its mirrors). DataLad collects, unifies, monitors, and exposes through convenient interfaces data available across a wide range of data providers

Distributed?
DataLad uses distributed version control Git and built on top of it Git-annex for data logistics. Git-annex enables *distributed* operation where clones of the datasets could be made available across multiple sites and media without loosing track of data and meta-information (such as versioning)

Current coverage

Overall we already provide unified access to over 10TB of neural data (see datasets.datalad.org).

OpenfMRI.org
curated fMRI (and EEG) datasets

CRCNS.org
curated electrophysiological and neuroimaging datasets

INDI
a collation of various datasets and initiatives (functional connectome, *etc.*): ADHD200, HBNSSI, CoRR, etc

Planned and WiP

Rich metadata
next (0.10) release will feature per-file metadata interactions with metadata aggregated from variety of sources (audio files, BIDS, EXIF, NIfTI, etc)

OSF, Zenodo, and Figshare support
ability to upload or export (as archives) data to known portals of scientific data (see PR#1446, PR#1942)

XNAT
widely used imaging informatics platform used by HumanConnectome.org, NITRC-IR, OpenfMRI and others (see PR#1552)

NeuroDebian
Through the joint venture with the NeuroDebian project, DataLad will also expose itself as a Debian APT repository, making it possible to *install* and *upgrade* datasets using conventional tools such as `apt` and `aptitude`. Data *installation* would become as easy as software installation.

HumanConnectome.org
anatomical, functional, diffusion MRI data from 1,200 subjects

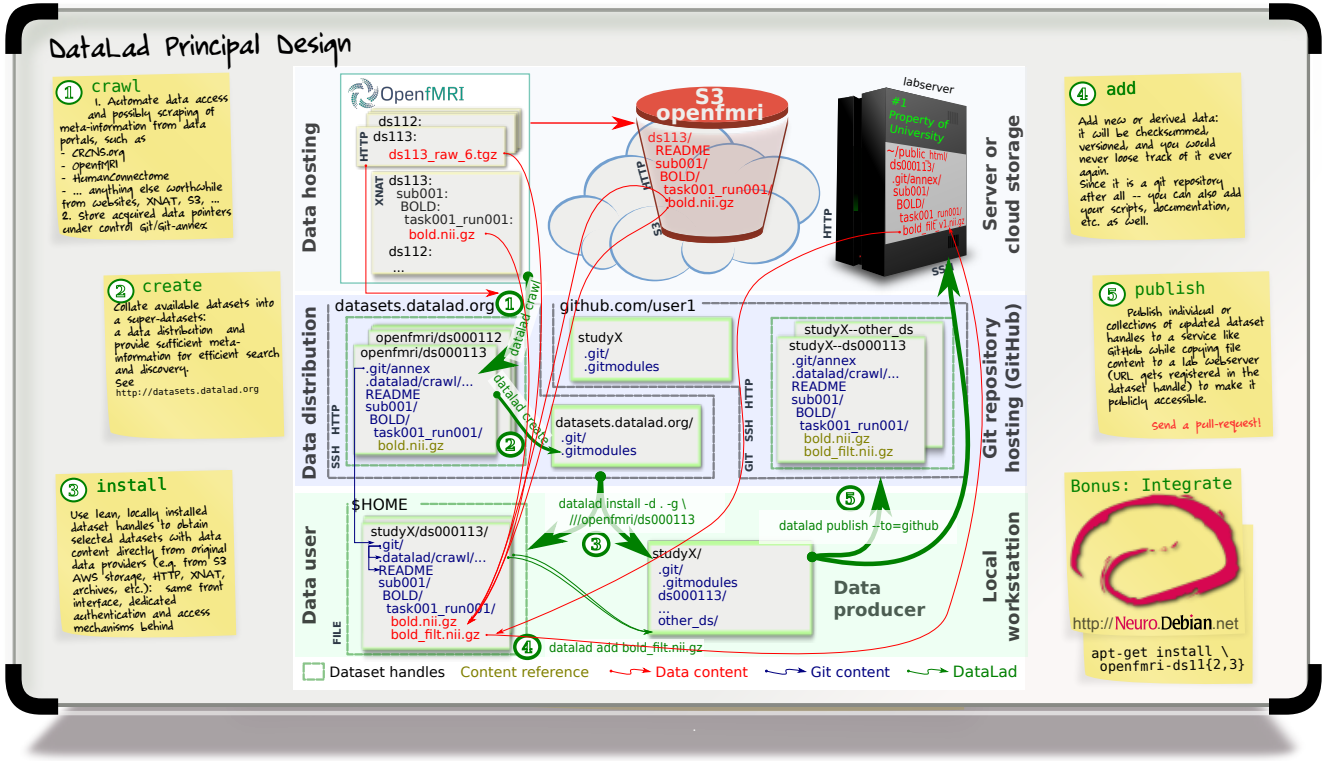
How could you help *both of us*?

Sharing scientific data is not yet as easy as it could and should be. Adhering to the following guidelines could help you to avoid unnecessary burden, thus making sharing easier and thus more rewarding. Overall motto is *Be Ready* and ...

Version your data:
even data close to bare origin might be *screwed* and require versioning:

- use Heudiconv to streamline DICOM to NIfTI/BIDS conversion with DataLad control
- use DataLad (and git and Git-annex) for managing your data or data you get from others. Visit `http://datalad.org/features.html` for demos.
- turn on versioning for your AWS S3 bucket(s) (and/or crawl it into a DataLad dataset)

Stay legit:
clear up and state ahead ownership (copyright) and (public domain dedication) license for your dataset. Provision public sharing in your consent forms **BEFORE** the data collection begins: `https://open-brain-consent.readthedocs.org`



Keep detail:
keep original detail – copies of acquisition protocols, exam cards, and the DICOMs (not only NIfTIs)

Be comprehensible:
adhere to a homogeneous files structure, adopt and extend if necessary some standard (*e.g.*, BIDS for neuroimaging or NWB for electrophysiological data)

Prepare to be reproduced:
analyze already pre-processed anonymized data

Think about longevity:
Deposit datasets to some public community repository (*e.g.*, OpenfMRI, figshare)

How could you help *DataLad*?

Contribute
`http://github.com/datalad/datalad/pulls`

Follow&Share
Twitter: `http://twitter.com/datalad`
Google+: `http://plus.google.com/+DataladOrg`
Blog: `http://datalad.org`

Discuss
`https://groups.google.com/forum/#!forum/datalad`

Complain&Suggest
`http://github.com/datalad/datalad/issues`
We are interested in use-cases, interesting datasets, feedback on design decisions, alpha-users, *etc.*

Acknowledgements

This project is co-funded by the US National Science Foundation (1429999) and the German Federal Ministry of Education and Research (BMBF 01GQ1411)

SPONSORED BY THE
Federal Ministry of Education and Research
BMBF 01GQ1411