


VERSION CONTROL

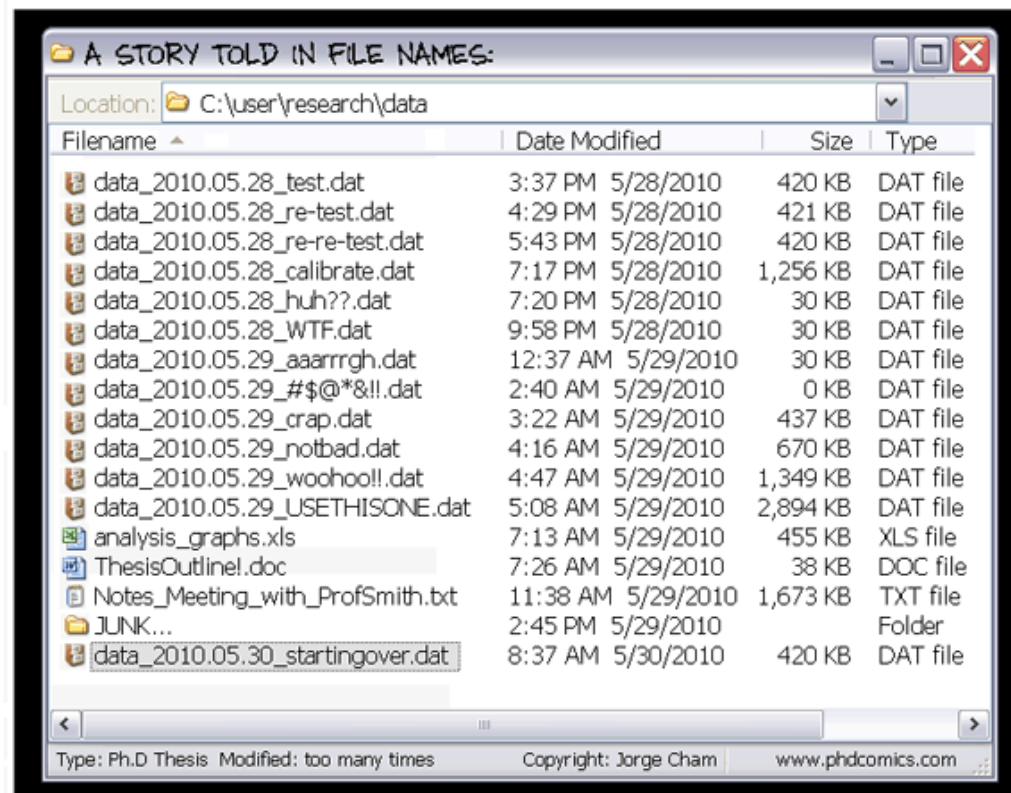
FOR DATA AND BEYOND

Adina Wagner

 @AdinaKrik

Psychoinformatics lab,
Institute of Neuroscience and Medicine, Brain & Behavior (INM-7)
Research Center Jülich
ReproNim/INCF fellow

Slides: [DOI 10.5281/zenodo.6346849](https://doi.org/10.5281/zenodo.6346849) (Scan the QR code)
Sources: github.com/adswa/talk-CIMeC



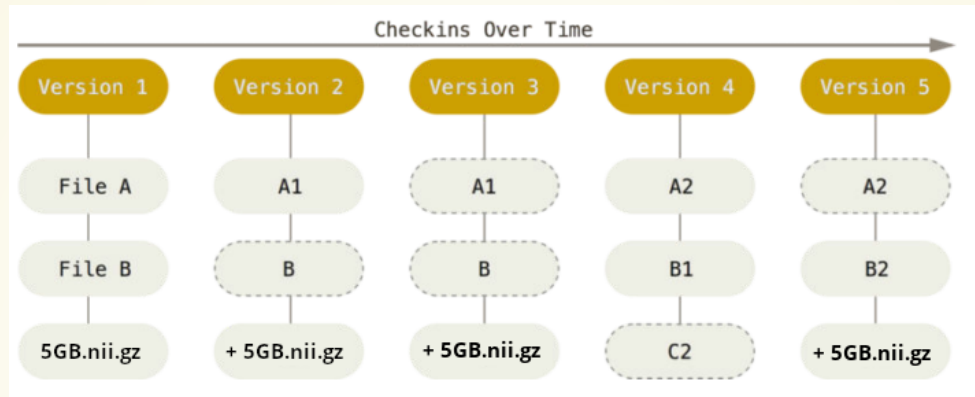
THE SAME, BUT FOR DATA:

```
--- /data/BnB1/DATA/download_data/eNKI -----  
    /..  
    5.2 TiB [#####] /eNKI_unzipped  
    3.3 TiB [##### ] /eNKI_redownload  
    3.2 TiB [##### ] /eNKI_BIDSdownload  
    724.2 GiB [#      ] /eNKI_20180806  
    218.8 GiB [      ] /eNKI_aus_Raw_Data
```

(Yes, 13 TB of data. Yes, real-life example)

HELP! GIT TO THE RESCUE?

Sadly, Git does not handle large files well.

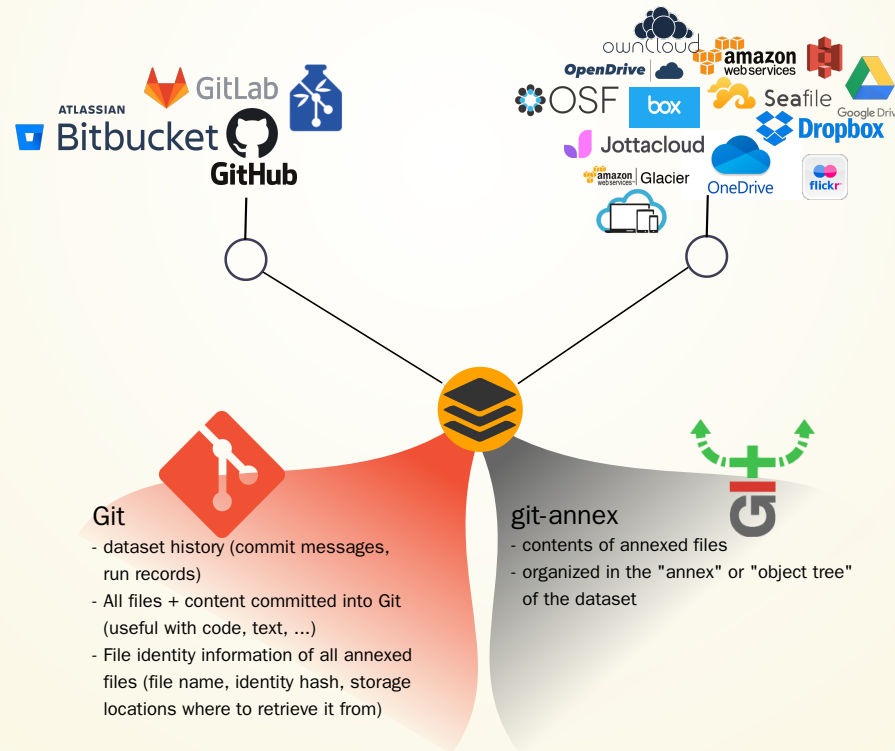


And repository hosting services refuse to handle large files:

```
adina@muninn in /tmp/myresearch on git:master
> git push gh-adswa master
Enumerating objects: 3, done.
Counting objects: 100% (3/3), done.
Delta compression using up to 8 threads
Compressing objects: 100% (2/2), done.
Writing objects: 100% (3/3), 497.87 KiB | 161.00 KiB/s, done.
Total 3 (delta 0), reused 0 (delta 0), pack-reused 0
remote: error: Trace: 64a78dd41ece8e5493fe33f97397a7a90ef9c91260ba32786970dbdcf5c4e0dd
remote: error: See http://git.io/iEPt8g for more information.
remote: error: File output.dat is 500.00 MB; this exceeds GitHub's file size limit of 100.00 MB
remote: error: GH001: Large files detected. You may want to try Git Large File Storage - https://git-lfs.github.com
To github.com:adswa/myresearch.git
! [remote rejected] master -> master (pre-receive hook declined)
error: failed to push some refs to 'github.com:adswa/myresearch.git'
```

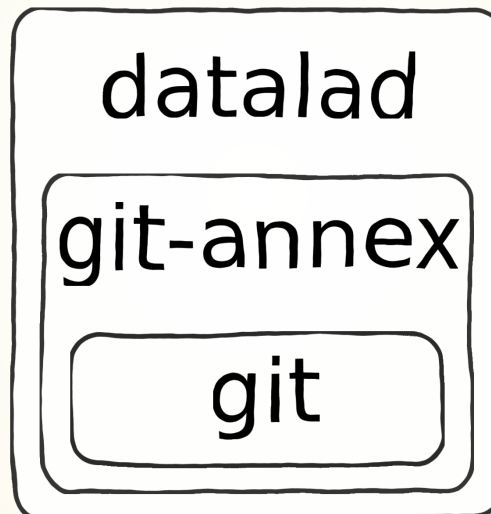


DISTRIBUTED VERSION CONTROL FOR DATA...



... AND TO IMPROVE SCIENTIFIC WORKFLOWS

"Share and treat data like software"



FURTHER INFORMATION

git-annex

- Source code: [git://git-annex.branchable.com/](https://git-annex.branchable.com/)
- Docs & Forum: git-annex.branchable.com/

DataLad

- Source code: github.com/datalad/datalad
- Technical docs: docs.datalad.org
- Video tutorials: Youtube channel "DataLad"
- Matrix channel: [DataLad](https://matrix.org/#/DataLad)
- User docs + tutorials: handbook.datalad.org



WHERE TO START

git-annex: "annex repo"

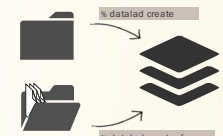
```
1 $ git init myrepo
2 Initialized empty Git repository in /tmp/myrepo

3 $ cd myrepo
4 $ git annex init
5 init ok
6 (recording state in git...)
```

datalad: "DataLad dataset"

```
1 $ datalad create mydataset
2 [INFO ] Creating a new annex
3 create(ok): /tmp/mydataset (da
```

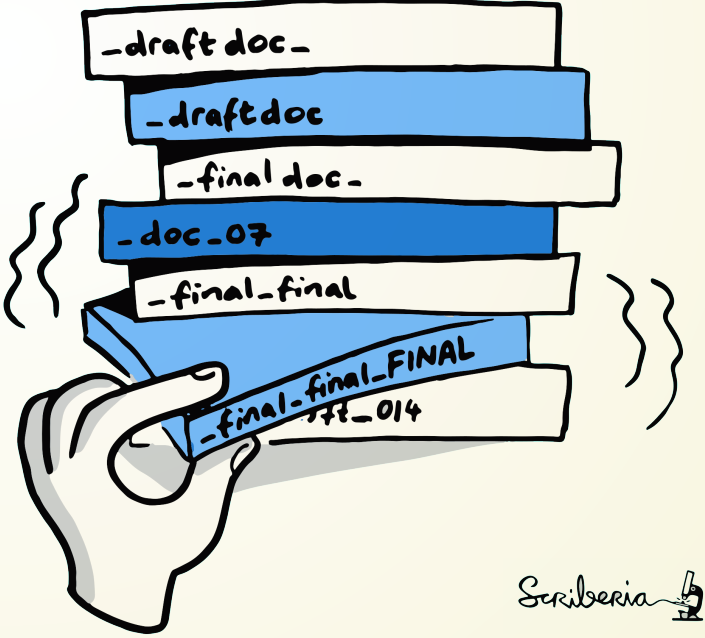
create new, empty datasets to populate...



... or transform existing directories into datasets

EXTENDING VERSION CONTROL ADVANTAGES TO DATA

1. TRANSPARENCY



1. TRANSPARENCY



Image credit: CC-BY Scriberia and The Turing Way

Scriberia 

Scriberia 

1. TRANSPARENCY

Git's revision history transparently lists all changes made in this collaboratively written paper:

```
2021-07-13 10:32 +0200 Adina Wagner M Merge pull request #8 from psychoinform
2021-07-13 10:31 +0200 Adina Wagner o [method] {origin/method} remove duplica
2021-07-13 10:21 +0200 Adina Wagner o Fix reference to listing
2021-07-13 10:20 +0200 Adina Wagner o Add HPC system to examples of singulari
2021-07-13 10:19 +0200 Adina Wagner o include participant_job.sh as a listing
2021-07-13 09:22 +0200 Adina Wagner M Merge branch 'master' of github.com:p
2021-07-13 08:53 +0200 Michael Hanke o Limit IQR figure size (450k instead o
2021-07-12 20:06 +0200 Adina Wagner M Merge branch 'master' of github.com:p
2021-07-12 19:42 +0200 Adina Wagner M Merge branch 'master' of github.com
2021-07-12 09:38 +0200 Laura Waite o add table 1 content
2021-07-13 09:21 +0200 Adina Wagner o Introduce the thought that software
2021-07-13 09:19 +0200 Adina Wagner o Refine information on on-demand con
2021-07-13 07:41 +0200 Adina Wagner o Properly cite missing references
2021-06-28 14:28 +0200 Adina Wagner o Restructure the first half of the m
2021-06-28 11:51 +0200 Adina Wagner o add a small post-workflow usage blu
2021-06-28 11:20 +0200 Adina Wagner o Shape existing contents into the ne
2021-06-28 11:19 +0200 Adina Wagner o Introduce a common structure in the
2021-06-28 11:16 +0200 Adina Wagner o Fix figure reference
2021-07-12 20:03 +0200 Adina Wagner o markup 'tar' as a software/command
2021-06-28 11:29 +0200 Adina Wagner o Add lost YODA reference
2021-07-12 09:38 +0200 Laura Waite o add table 1 content
2021-07-12 19:42 +0200 Adina Wagner o Cut slightly inaccurate detail from f
2021-07-08 15:25 +0200 Laura Waite o fix typos in the methods
2021-07-08 15:23 +0200 Laura Waite o a few tweaks/typos in the discussion
2021-07-08 15:20 +0200 Laura Waite o consistently hyphenate re-user(s)
[main] 2b2184810a33091353285ef8368e205c767c44ce - commit 521 of 600 88%

commit 2b2184810a33091353285ef8368e205c767c44ce
Refs:
Author: Adina Wagner <adina.wagner@online.de>
AuthorDate: Mon Jun 28 14:28:57 2021 +0200
Commit: Adina Wagner <adina.wagner@online.de>
CommitDate: Mon Jul 12 20:28:37 2021 +0200

    Restructure the first half of the method section, and provide a smaller grained s
---
manuscript/main.tex | 106 ++++++-----
1 file changed, 64 insertions(+), 42 deletions(-)
diff --git a/manuscript/main.tex b/manuscript/main.tex
index 8596876..0e0ace7 100644
--- a/manuscript/main.tex
+++ b/manuscript/main.tex
@@ -454,10 +454,17 @@ As a byproduct of up to complete computational reproducibility,
% Methods section provides all technical details necessary for the independent repro
\section*{Methods}

-\subsection*{Terminology}
+\subsection*{Generic workflow}
+
+The primary purpose of the workflow is the reproducible execution of a containerized
+We first introduce the general workflow and its elements in detail.
[diff] 2b2184810a33091353285ef8368e205c767c44ce - line 1 of 187 13%
```

1. TRANSPARENCY - FOR DATA

Why? Data changes,
too!

- Additional acquisitions
- Errors identified or fixed
- Restructuring to the latest BIDS standard
- ...

Example: The **ABCD** study identified several data issues in 2019, among others, flipped field maps for specific scanner types:

1. TRANSPARENCY - FOR DATA

Once you track changes to data with version control tools, you can find out *why* it changed, *what* has changed, *when* it changed, and *which version* of your data was used at which point in time.

```
2020-03-13 10:46 +0100 Adina Wagner o [DATALAD RUNCMD] add non-defaced
2020-03-13 10:29 +0100 Adina Wagner o [DATALAD RUNCMD] revert DICOM
2018-05-11 09:23 +0200 Michael Hanke o [master] {origin/HEAD} {origin/m
2018-05-11 09:19 +0200 Michael Hanke o Enable DataLad metadata extracto
2018-05-11 09:17 +0200 Michael Hanke o [DATALAD] new dataset
2018-05-11 09:17 +0200 Michael Hanke o [DATALAD] Set default backend fo
2018-01-19 14:19 +0100 Michael Hanke o <v1.5> Update changelog for 1.5
2018-01-19 14:09 +0100 Michael Hanke o BF: Re-import respiratory trace
2018-01-14 18:59 +0100 Michael Hanke o Fix type in physio log converter
2017-01-10 10:10 +0100 Michael Hanke o ENH: Report per-stimulus events
2016-12-10 20:18 +0100 Michael Hanke o Add BIDS-compatible stimuli/ dir
2016-11-15 07:04 +0100 Michael Hanke o Minor tweaks to gaze overlay scr
2016-10-30 11:03 +0100 Michael Hanke o Add "TaskName" meta data field f
2016-09-21 08:33 +0200 Michael Hanke o Add task-*_physio.json files
2016-09-21 08:23 +0200 Michael Hanke o BF: Fix task label in file names
2016-08-04 13:14 +0200 Michael Hanke o Update changelog
2016-08-03 22:22 +0200 Michael Hanke o Add cut position information to
2016-05-27 17:35 +0200 Michael Hanke o {origin/_} Mention openfMRI as d
2016-04-04 09:31 +0200 Michael Hanke o Update publication links
2016-03-31 11:26 +0200 Michael Hanke o Disable invalid test
[main] 6da25fb6fee2c698d35f52066698b6f94850f4d2 - commit 10 of 79 27%
commit 6da25fb6fee2c698d35f52066698b6f94850f4d2
Refs: v1.0-19-g6da25fb6
Author: Michael Hanke <michael.hanke@gmail.com>
AuthorDate: Fri Jan 19 14:09:53 2018 +0100
Commit: Michael Hanke <michael.hanke@gmail.com>
CommitDate: Fri Jan 19 14:11:23 2018 +0100

    BF: Re-import respiratory trace after bug fix in converter (fixes gh-
---
...er_task-movie-localizer_run-1_recording-cardresp_physio.tsv.gz | 2 +-
..._task-objectcategories_run-1_recording-cardresp_physio.tsv.gz | 2 +-
..._task-objectcategories_run-2_recording-cardresp_physio.tsv.gz | 2 +-
..._task-objectcategories_run-3_recording-cardresp_physio.tsv.gz | 2 +-
..._task-objectcategories_run-4_recording-cardresp_physio.tsv.gz | 2 +-
...calizer_task-retmapccw_run-1_recording-cardresp_physio.tsv.gz | 2 +-
...calizer_task-retmapclw_run-1_recording-cardresp_physio.tsv.gz | 2 +-
...calizer_task-retmapcon_run-1_recording-cardresp_physio.tsv.gz | 2 +-
...calizer_task-retmapexp_run-1_recording-cardresp_physio.tsv.gz | 2 +-
...2_ses-movie_task-movie_run-1_recording-cardresp_physio.tsv.gz | 2 +-
...2_ses-movie_task-movie_run-2_recording-cardresp_physio.tsv.gz | 2 +-
[diff] 6da25fb6fee2c698d35f52066698b6f94850f4d2 - line 1 of 2391 0%
```

TRANSPARENCY - FOR DATA

git-annex

```
1 $ git annex add mylargefile
2 add mylargefile
3 ok
4 (recording state in git...)
5 $ git commit -m "annexed a large file"
6 [master 0efa6cc] annexed a large file
7 1 file changed, 1 insertion(+)
8 create mode 120000 mylargefile
9
```

datalad

```
1 $ datalad save -m "annexed a large file"
2 add(ok): mylargefile (file)
3 save(ok): . (dataset)
4 action summary:
5   add (ok: 1)
6   save (ok: 1)
```

2. ACCESSIBILITY/AVAILABILITY

Springer Link Search Log in

Download PDF

Published: February 2010

An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data

The tool you really need for your research

expired email & left academia for good


[Behavior Research Methods](#) 42, 188–204 (2010) | [Cite this article](#)

12k Accesses | 304 Citations | 1 Altmetric | [Metrics](#)

Abstract

Event detection is used to classify recorded gaze points into periods of fixation, saccade, smooth pursuit, blink, and noise. Although there is an overall consensus that current algorithms for event detection have serious flaws and that a de facto standard for event detection does not exist, surprisingly little work has been done to remedy this problem. We present a new algorithm that addresses the known limitations into account. Moreover, the algorithm is designed to be used at the end of many saccades, a threshold that makes the event detection less sensitive to variations in noise level and the algorithm settings-free for the user. We demonstrate the performance of the new algorithm on eye movements recorded during scene perception and compare it with two of the most commonly used algorithms today. Unlike the currently used algorithms, fixations, saccades, and glissades are robustly identified. Using this algorithm, we found that glissades occur in about half of the saccades, during both scene perception, and that they have an average duration close to 24 msec. Due to the high prevalence of glissades, we argue that researchers must actively choose whether to assign the glissades to the choice affects dependent variables such as fixation and saccade duration significantly. C

"The toolbox is available from the corresponding author upon reasonable request"



2. ACCESSIBILITY/AVAILABILITY

The latest version of REMoDNaV can be installed from PyPi³ via `pip install remodnav`. The source code of the software can be found on Github.⁴ All reports on defects and enhancement can be submitted there. The analysis code underlying all results and figures presented in this paper, as well as the L^AT_EX sources, are located in another GitHub repository.⁵ All required input data, from Andersson et al. (2017) and the *studyforrest.org* project, are referenced in this repository at precise versions as DataLad⁶ subdatasets, and can be obtained on demand. The repository constitutes an automatically reproducible research object, and readers interested in verifying the results and claims of our paper can recompute and plot all results with a single command after cloning the repository.

Acknowledgements This work is based on an earlier Python implementation and evaluation of the original NH algorithm by Ulrike Schnaithmann and Isabel Dombrowe (Schnaithman, 2017).

²<https://www.michaeldorr.de/smoothpursuit/>

³<https://pypi.org/project/remodnav>

⁴<https://github.com/psychoinformatics-de/remodnav>

⁵<https://github.com/psychoinformatics-de/paper-remodnav/>

⁶<http://datalad.org>



2. ACCESSIBILITY/AVAILABILITY

Publicly shared code removes the bottleneck of an expired institutional e-mail address

The screenshot shows the GitHub repository page for 'remodnav' by 'psychoinformatics-de'. The repository is public and has 4 unwatchers, 10 forks, and 29 stars. The main content area displays a list of files and folders with their commit history. The README.md file is highlighted, showing the project title 'REMoDNaV - Robust Eye Movement Detection for Natural Viewing' and various badges for build status, code coverage, license, and release information.

File/Folder	Commit Message	Time Ago
.github/workflows	TST: Setup GH actions for CI, as travis ran out of credits	6 months ago
eval	Changed transparency of mainseq data points to 1 (opaque), changed ...	3 years ago
remodnav	Bump version	3 months ago
tools/ci	Switch to appveyor w/ cross-platform tests	13 months ago
.appveyor.yml	TST: Replace git annex url with snapshot, as suggested in #33 by @mih	6 months ago
.coveragerc	Coverage config	4 years ago
.gitignore	Small random fixes	3 years ago
.gitmodules	[DATA LAD] added content	4 years ago
CHANGELOG.md	Release 1.1: Add a changelog	3 months ago
CODE_OF_CONDUCT.md	Adopt the Contributor Covenant CoC	last month
CONTRIBUTORS	General project info	4 years ago
LICENSE	Rename license file to something more obvious	4 years ago
Makefile	BF: Fixup deps	4 years ago
README.md	Switch to appveyor w/ cross-platform tests	13 months ago
requirements-devel.txt	FIX dependency issue in travis	3 years ago
setup.py	Use pandoc to convert md to rst for nicer project display on PyPi	3 months ago

README.md

REMoDNaV - Robust Eye Movement Detection for Natural Viewing

build passing codecov 86% License MIT release v1.1.1 pypi package 1.1.1 DOI 10.5281/zenodo.5746931

About
Robust Eye Movement Detection for Natural Viewing
eye-tracking bids
Readme
View license
Code of conduct
29 stars
4 watching
10 forks

Releases 5
Even more safeguards Latest
on Dec 1, 2021
+ 4 releases

Packages
No packages published
Publish your first package

Used by 1
@christianwarmuth / neurogaze

Contributors 4
mih Michael Hanke
adswa Adina Wagner

2. ACCESSIBILITY/AVAILABILITY - FOR DATA

And the same can be true for data:

The screenshot shows a GitHub repository page for 'psychoinformatics-de / studyforrest-data-phase2'. The page includes a navigation bar with 'Pull requests', 'Issues', 'Marketplace', and 'Explore'. Below the repository name, there are buttons for 'Unwatch', 'Unstar', and 'Fork'. The main content area shows a list of files and folders, including '.datalad', 'code', 'src', 'stimuli', and subfolders 'sub-01' through 'sub-18'. Each file entry includes a description and a date. On the right side, there is an 'About' section with a description of the repository, a link to 'studyforrest.org', and a 'Releases' section with a 'First public release' button. The 'Contributors' section lists three contributors: mih Michael Hanke, dakot Daniel Kottke, and adswa Adina Wagner.

Search or jump to... Pull requests Issues Marketplace Explore

psychoinformatics-de / studyforrest-data-phase2

Unwatch 1 Unstar 6 Fork 8

Code Issues 4 Pull requests 1 Actions Projects Security Insights

master 2 branches 1 tag

Go to file Add file Code

mih Merge pull request #15 from adswa/ENH/README ... b5306e2 on May 7 77 commits

.datalad	[DATALAD] dataset aggregate metadata update	2 years ago
code	Fix type in physio log converter (fixes gh-11)	3 years ago
src	Recover lost segment from eyetracker (closes gh-3)	5 years ago
stimuli	Add BIDS-compatible stimuli/ directory (with symlinks)	4 years ago
sub-01	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-02	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-03	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-04	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-05	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-06	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-09	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-10	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-14	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-15	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-16	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-17	BF: Re-import respiratory trace after bug fix in converte...	3 years ago
sub-18	BF: Re-import respiratory trace after bug fix in converte...	3 years ago

About

studyforrest.org: Phase2 data (movie, eyetracking, retmapping, visual localizers) [BIDS]

studyforrest.org

Readme

View license

Releases 1

First public release Latest on Mar 26, 2016

Packages

No packages published

Contributors 3

mih Michael Hanke

dakot Daniel Kottke

adswa Adina Wagner

2. ACCESSIBILITY/AVAILABILITY - FOR DATA

On demand file access via git-annex/DataLad:

```
# clone the repository
$ git clone https://github.com/psychoinformatics-de/studyforrest-data-phase2.git
# get one or more files/directories/... on demand
$ git annex get file/directory/...
# or
$ datalad clone https://github.com/psychoinformatics-de/studyforrest-data-phase2.git
$ datalad get file/directory/...
```

Fortunate side-effect: Cloned repos/datasets are **small in size**,
but can be **browsed** for existing files and can provide **access** to
their content regardless of where it is hosted.
You can have access to more files than your computer has disk space!

3. SECURITY AND RELIABILITY

FOR MY LOST LAPTOP

I am a Rutgers Chemistry 5th year PhD student. On April 19th afternoon, my LENOVO THINKPAD T420S laptop was stolen from room 203 of Wright-Rieman building. If you stole my laptop and now you are reading this letter, I would like to say that you can keep the computer and I would like to pay you money for my data under D drive. The data is my FIVE-YEAR work. I really need the data under the D drive, there is a folder named RESEARCH, under RESEARCH folder, there is a THESIS folder. I only need that folder for my thesis defense, which is coming very soon. I would like to pay you \$1000 and use whatever way you offer to send you the money. The price is negotiable. My laptop password is 850713zd, my email address is [REDACTED] and phone number is [REDACTED]. PLEASE contact me and I would appreciate it so much!!!

3. SECURITY AND RELIABILITY

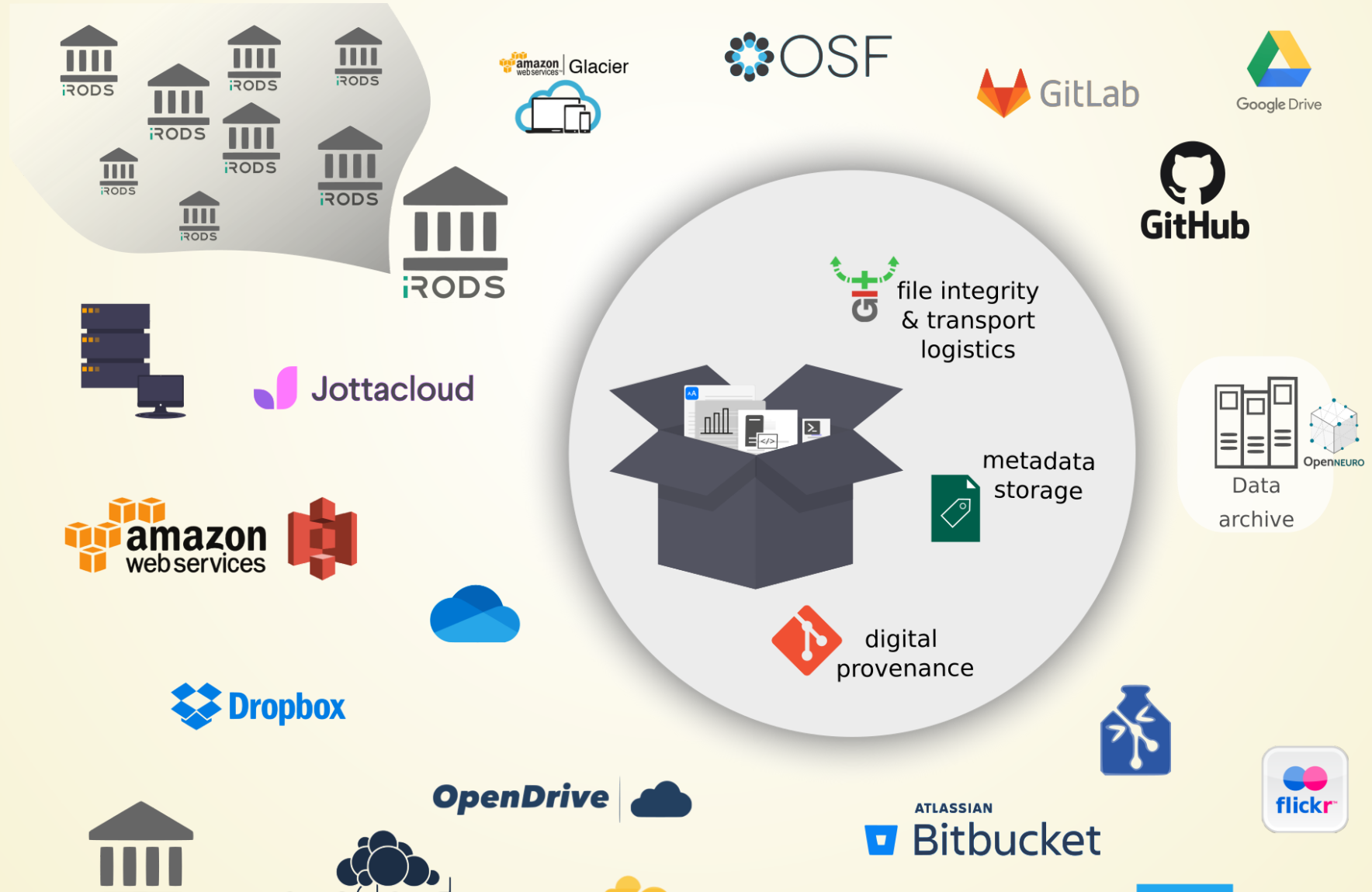
Git is great for keeping and synchronizing backups:

The screenshot shows a GitHub repository page for 'adswa / thesis'. The repository is public and has 1 watch, 0 forks, and 0 stars. The main branch is 'master'. The repository contains several files, including 'img', '.gitignore', 'Literatur.bib', 'Masterarbeit.bcf', 'Masterarbeit.bib', 'Masterarbeit.nlo', 'Masterarbeit.nls', 'Masterarbeit.tex', and 'README.md'. The 'README.md' file is selected, showing the title 'thesis' and the content: 'Backup for my thesis LaTeX project. This repo contains the LaTeX project of my Masters thesis. The main file that needs to be compiled with pdflatex is Masterarbeit.tex. There is and additional nomenclature nlo file which needs to be made with the command line call `makeindex Masterarbeit.nlo -s nomenc1.ist -o`'.

File	Description	Time
img	[DATALAD] added content	3 years ago
.gitignore	ENH: inital .gitignore for TeX projects	3 years ago
Literatur.bib	Fix Tab	3 years ago
Masterarbeit.bcf	have all dependent files in git as well	3 years ago
Masterarbeit.bib	alternative bib file should biber >2.7 get rele...	3 years ago
Masterarbeit.nlo	add file necesseary to generate nomenclature	3 years ago
Masterarbeit.nls	Nomenclature has FSL and SPM	3 years ago
Masterarbeit.tex	Add abomination of Code necessary for apa ...	3 years ago
README.md	Create README.md	3 years ago

3. SECURITY AND RELIABILITY - FOR DATA

Decentral version control for data integrates with a variety of services to let you store data in different places - creating a resilient network for data



3. SECURITY AND RELIABILITY - FOR DATA



Repository hosting

- usually no annex support & can't hold large data for free
- exposes Git history and files stored in Git
- datasets can be cloned from there

Storage hosting in a special remote

- usually no Git repository hosting service
- stores the object tree/ file contents
- datasets keep track of where data is stored, datalad get retrieves file contents from special remote

Git

- dataset history (commit messages, run records)
- All files + content committed into Git (useful with code, text, ...)
- File identity information of all annexed files (file name, identity hash, storage locations where to retrieve it from)

git-annex

- contents of annexed files
- organized in the "annex" or "object tree" of the dataset

3. SECURITY AND RELIABILITY - FOR DATA

Example: A Git repository with annexed data in a public S3 bucket

The screenshot shows a GitHub repository page for `OpenNeuroDatasets / ds003633`. The repository is public and has 3 watchers, 0 forks, and 0 stars. The main content area displays a list of files and folders, all recorded changes by [DATALAD]. The files and folders listed are:

- `.datalad` [DATALAD] new dataset 11 months ago
- `code` [DATALAD] Recorded changes 10 months ago
- `derivatives/preproc_...` [DATALAD] Recorded changes 10 months ago
- `sub-01/ses-movie` [DATALAD] Recorded changes 11 months ago
- `sub-02/ses-movie` [DATALAD] Recorded changes 11 months ago
- `sub-03/ses-movie` [DATALAD] Recorded changes 11 months ago
- `sub-04/ses-movie` [DATALAD] Recorded changes 11 months ago
- `sub-05/ses-movie` [DATALAD] Recorded changes 10 months ago
- `sub-06/ses-movie` [DATALAD] Recorded changes 11 months ago
- `sub-07/ses-movie` [DATALAD] Recorded changes 11 months ago
- `sub-08/ses-movie` [DATALAD] Recorded changes 11 months ago
- `sub-09/ses-movie` [DATALAD] Recorded changes 11 months ago
- `sub-10/ses-movie` [DATALAD] Recorded changes 11 months ago
- `sub-11/ses-movie` [DATALAD] Recorded changes 11 months ago
- `sub-emptyroom` [DATALAD] Recorded changes 11 months ago
- `.bidsignore` [DATALAD] Recorded changes 11 months ago
- `.gitattributes` [DATALAD] exclude paths from annex'ing 11 months ago
- `CHANGES` [DATALAD] Recorded changes 8 months ago
- `README` [DATALAD] Recorded changes 9 months ago

The right sidebar shows the repository's metadata, including the Readme, 0 stars, 3 watching, and 0 forks. The Releases section shows 4 tags, and the Packages section shows no packages published.

3. SECURITY AND RELIABILITY - FOR DATA



Repositories with annex support

- examples: GIN (gin.g-node.org), GitLab instances with enabled annex support
- can hold large data for free
- exposes Git history and all files + content
- datasets can be cloned from there



3. SECURITY AND RELIABILITY - FOR DATA

Publishing data to **Gin** - for example as a backup:

```
(handbook2) adina@muninn in /tmp/mydataset on git:master
> datalad create-sibling-gin published-dataset
create_sibling_gin(ok): [sibling repository 'gin' created at https://gin.g-node.org/adswa/published-dataset]
configure_sibling(ok): . (sibling)
action summary:
  configure_sibling (ok: 1)
  create_sibling_gin (ok: 1)
(handbook2) adina@muninn in /tmp/mydataset on git:master
> datalad push --to gin
copy(ok): mylargefile (file) [to gin...]
publish(ok): . (dataset) [refs/heads/git-annex->gin:refs/heads/git-annex 43395c0..f94feda]
publish(ok): . (dataset) [refs/heads/master->gin:refs/heads/master [new branch]]
action summary:
  copy (ok: 1)
  publish (ok: 2)
(handbook2) adina@muninn in /tmp/mydataset on git:master
> █
```

1

4. VISIBILITY AND REUSABILITY

Repository hosting services add cool features and integrations:

4. VISIBILITY AND REUSABILITY - FOR DATA

← → ↻ <https://doi.gin.g-node.org/10.12751/g-node.4ivuv8/> ☆

Published Data Keywords Public datasets on GIN

Dataset

Dynamics of fMRI patterns reflect sub-second activation sequences and reveal replay in human visual cortex - MRI data according to the Brain Imaging Data Structure (BIDS)

Lennart Wittkuhn, Nicolas W. Schuck

Max Planck Institute for Human Development, Max Planck UCL Centre for Computational Psychiatry and Ageing Research, Berlin, Germany

DOI: 10.12751/g-node.4ivuv8 [BROWSE REPOSITORY](#) [BROWSE ARCHIVE](#) [DOWNLOAD ARCHIVE \(ZIP 128 GiB\)](#)

Published 05 Dec. 2020 | License [Creative Commons Attribution-ShareAlike 4.0](#)

Description

Neural computations are often fast and anatomically localized. Yet, investigating such computations in humans is challenging because non-invasive methods have either high temporal or spatial resolution, but not both. Of particular relevance, fast neural replay is known to occur throughout the brain in a coordinated fashion about which little is known. We develop a multivariate analysis method for functional magnetic resonance imaging that makes it possible to study sequentially activated neural patterns separated by less than 100 ms with precise spatial resolution. Human participants viewed images individually and sequentially with speeds up to 32 ms between items. Probabilistic pattern classifiers were trained on activation patterns in visual and ventrottemporal cortex during individual image trials. Applied to sequence trials, probabilistic classifier time courses allow the detection of neural representations and their order. Order detection remains possible at speeds up to 32 ms between items. The frequency spectrum of the sequentiality metric distinguishes between sub-versus supra-second sequences. Importantly, applied to resting-state data our method reveals fast replay of task-related stimuli in visual cortex. This indicates that non-hippocampal replay occurs even after tasks without memory requirements and shows that our method can be used to detect such spontaneously occurring replay.

Keywords

[cognitive neuroscience](#) | [functional magnetic resonance imaging](#) | [hippocampal replay](#) |

References

Wittkuhn, L. and Schuck, N. W. (2020). Dynamics of fMRI patterns reflect sub-second activation sequences and reveal replay in human visual cortex. *Nature Communications*

Wittkuhn, L. and Schuck, N. W. (2020). Faster than thought: Detecting sub-second activation sequences with sequential fMRI pattern analysis. *bioRxiv*. doi:10.1101/2020.02.15.950667 <https://doi.org/10.1101/2020.02.15.950667>

Funding

Max Planck Society Independent Max Planck Research Group grant
European Union ERC Starting Grant ERC-2019-STG REPLAY-852669
Max Planck Institute for Human Development

Citation

Wittkuhn L, Schuck NW (2020) Dynamics of fMRI patterns reflect sub-second activation sequences and reveal replay in human visual cortex - MRI data according to the Brain Imaging Data Structure (BIDS). G-Node. <https://doi.org/10.12751/g-node.4ivuv8>

5. COLLABORATION



5. COLLABORATION

The screenshot shows the GitHub interface for the repository `datalad-datasets/human-connectome-project-openaccess`. The repository is public and has 7 issues, 1 pull request, 4 forks, and 21 stars. The navigation bar includes links for Code, Issues (7), Pull requests (1), Discussions, Actions, Projects, and Wiki. The search bar shows filters for `is:pr is:closed`. Below the search bar, there are buttons for Labels (10) and Milestones (0), and a green button for 'New pull request'. The main content area displays a list of pull requests, each with a title, author, merge date, and status. The pull requests are:

- Update after Nov 2021 release** (#31) by mih, merged 13 days ago, Approved. 2 reviews, 1 comment.
- BF: Prioritize subdataset clone configuration** (#28) by adswa, merged on Oct 18, 2021, Approved. 2 comments.
- DOC: Provide link to the HCP S1200 project** (#27) by adswa, merged on Sep 6, 2021. 1 review.
- Add missing WM task bold file from ConnectomeDB** (#25) by adswa, merged on May 1, 2021, 1 task done. 1 review, 1 comment.
- add a range of MNINonLinear/Results/ files for subjects 193239 467351 705341** (#18) by adswa, merged on Feb 22, 2021, 1 task done. 1 review, 1 comment.
- add rfMRI_REST1_LR.nii.gz for 200614 and 205119 that are only available via REST** (#11) by adswa, merged on Aug 26, 2020, 1 task done. 3 comments.
- Add missing data from three subjects** (#10) by mih, merged on Jul 28, 2020. 1 review.
- ENH: minor tune ups to README.md** (#6) by yarikoptic, merged on Apr 9, 2020, Approved.

Psst! github.com/datalad-datasets/human-connectome-project-openaccess provides fine-grained access to the HCP dataset

5. COLLABORATION

Teamscience on more than code:

The screenshot shows a GitHub issue page with a dark theme. At the top, the issue title is "Q: Adjustment to tiny labeling mistake (ref Zemblys, 2018)? #2". Below the title, a purple pill indicates the issue is "Closed", and text says "adswa opened this issue on Mar 8, 2019 · 1 comment".

The main content is a comment from user "adswa" dated "Mar 8, 2019". The comment text is as follows:

Zemblys et al., 2018, report a minor labeling mistake in one image file [here](#) (or a pay-wall free version [here](#)) in Appendix 2:

We found an obvious labeling mistake in the one of the validation trials (file UH29_img_Europe_labelled_MN. We fixed this error by reassigning 75 samples, [3197,3272) (zero-based index), from the saccade to the fixation class.

I checked the data file in question and it appears to still contain the erroneous saccade labels. Just to reconfirm: this labeling error has not been fixed in the data file in this repository, correct?

If you wish, I can PR a fixed file, the issue at hand is intended to just reconfirm my assumption.

Thanks in advance!

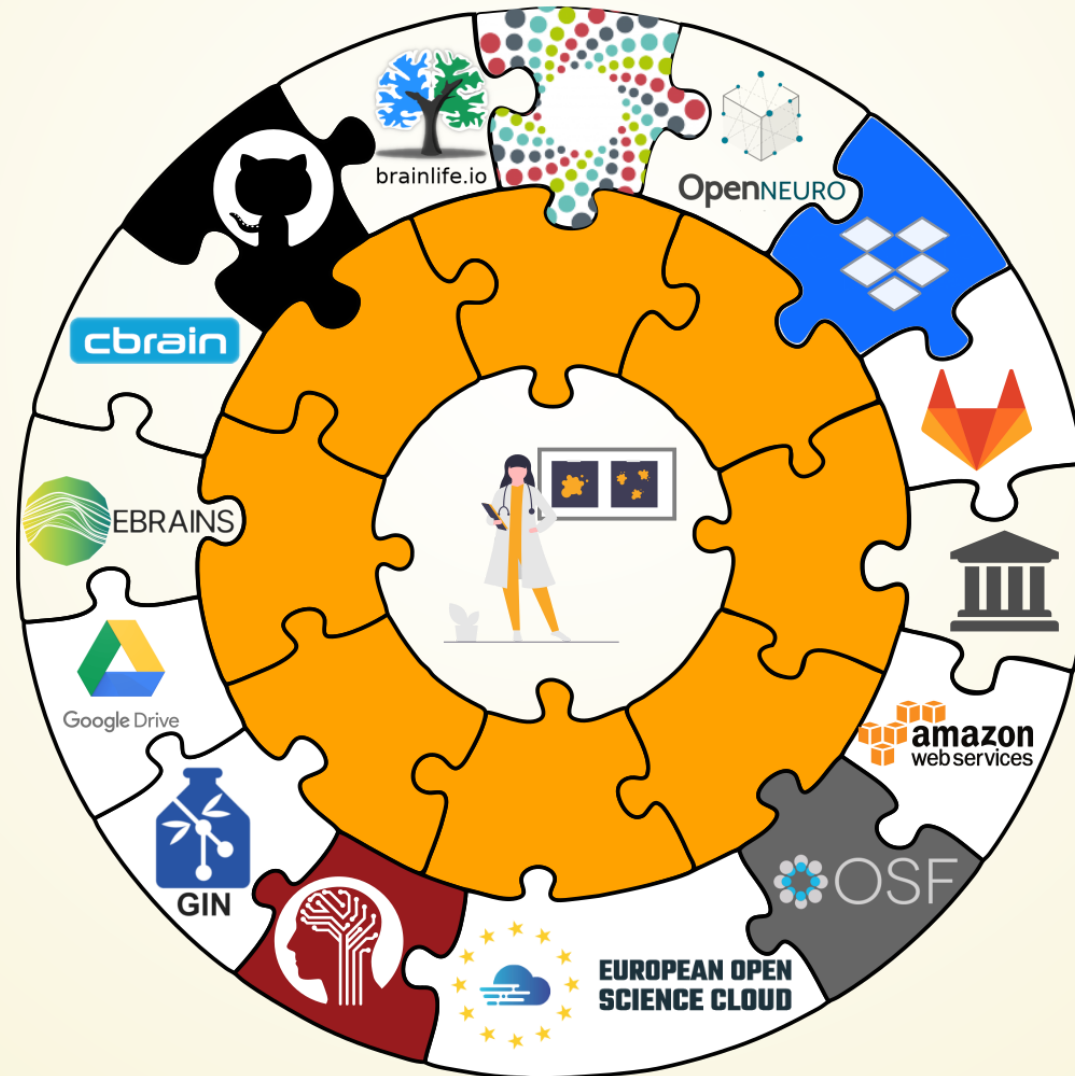
Below the comment, there is a commit entry: "adswa added a commit to adswa/remodnav that referenced this issue on Mar 8, 2019". The commit message is "ENH/FIX: use file with fixed labels." and the commit hash is "1b2b162".

6. EFFICIENCY



6. EFFICIENCY

DataLad is built to maximize interoperability and use with hosting and storage technology you already use



6. EFFICIENCY

```
(handbook2) adina@muninn in /tmp/mylargefile/mydataset on git:master
```

0 AM (+0100 UTC). Thank you for your patience.

hereismyresearch

Contributors: [Adina Svenja Wagner](#)
Date created: 2022-03-10 08:57 PM | Last Updated: 2022-03-10 08:58 PM
Category: Data

Description: This component was built from a DataLad dataset using the datalad-osf extension (<https://github.com/datalad/datalad-osf>). With this extension installed, this component can be git or datalad cloned from a 'osf://ID' URL, where 'ID' is the OSF node ID that shown in the OSF HTTP URL, e.g. <https://osf.io/q8xnk> can be cloned from osf://q8xnk. This particular project can be cloned using 'datalad clone osf://4fr7n'

Name ^ v	Modified ^ v
hereismyresearch	
OSF Storage (United States)	
.git	
MD5E-s6--8509e857aec759b1085d5010f0f6f677	2022-03-10 08:57 PM

Citation

Tags

- 493fb01c-4c99-4420-949d-ea72d600a551 DataLad dataset

Recent Activity

- Adina Svenja Wagner added file .git/repo.zip to OSF Storage in hereismyresearch 2022-03-10 08:58 PM
- Adina Svenja Wagner added file .git/refs to OSF Storage in hereismyresearch 2022-03-10 08:58 PM
- Adina Svenja Wagner created folder .git in OSF Storage in hereismyresearch 2022-03-10 08:58 PM

7. REPRODUCIBILITY

Science has many different building blocks

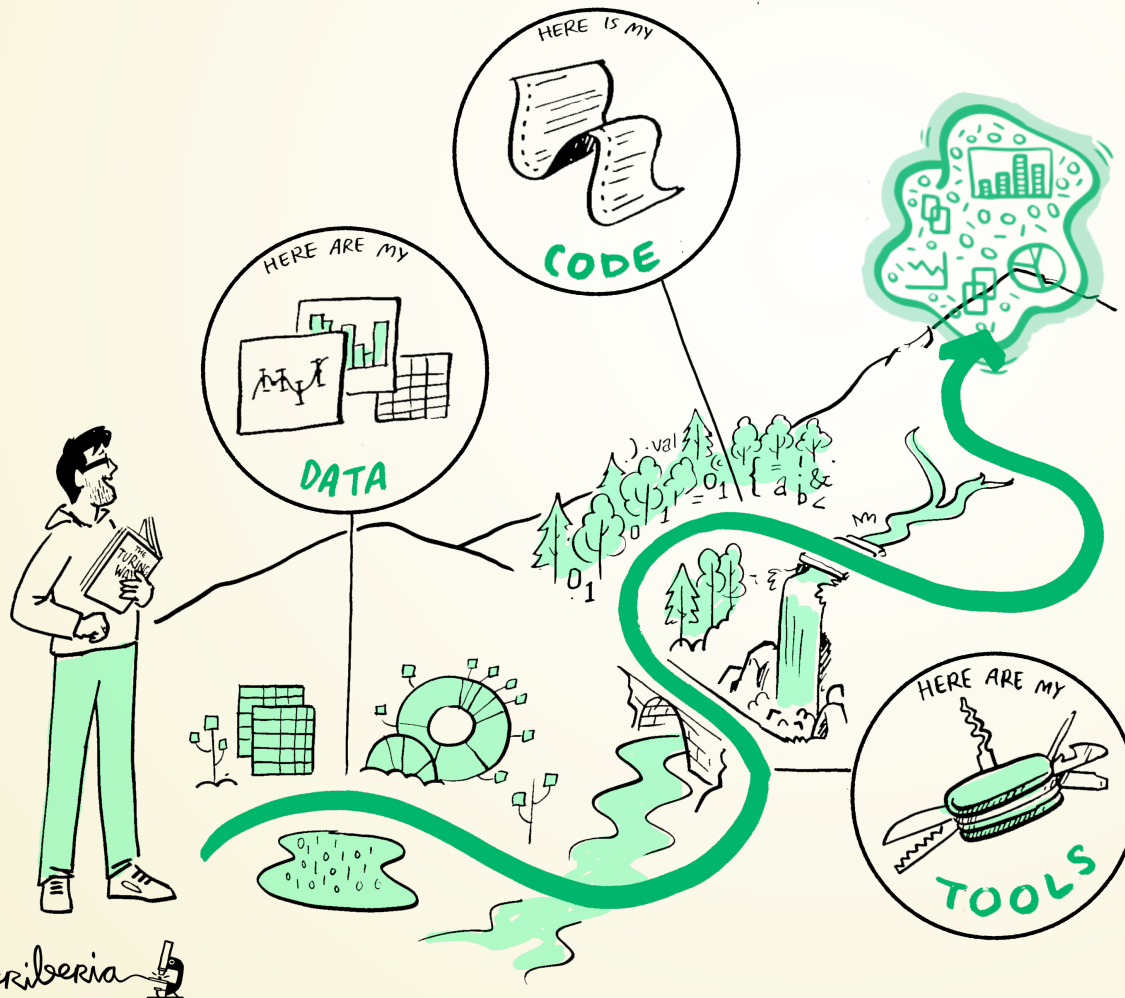


Code, software, and data produce research outputs:



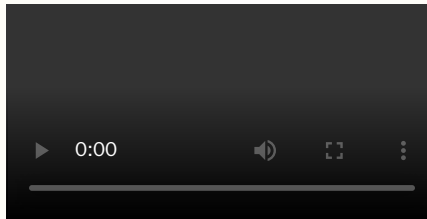
7. REPRODUCIBILITY

The more you share, the more likely can others reproduce your results



7. REPRODUCIBILITY

VC tools for data let you keep all ingredients also next to each other:



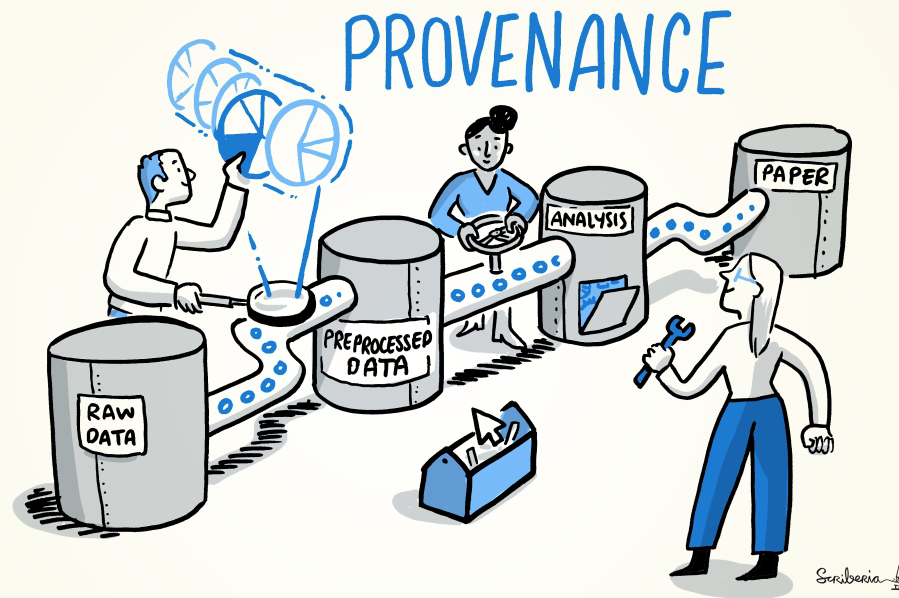
You can find this recording on YouTube: <https://www.youtube.com/watch?v=nhLqmF58SLQ> and a Walkthrough at handbook.datalad.org

```
1 from datalad.api import get as datalad_get
2 [...]
3 infiles = [op.join('data', 'raw_eyegaze', 'sub-32', 'beh', 'sub-32_task-movie_run-5_r
4 for f in infiles:
5     datalad get(f)
6     data = np.recfromcsv(f)
```

UNIQUE ADDITIONAL ADVANTAGES FOR SCIENCE

MODULARITY VIA DATASET NESTING

- Typically, Git repositories are cumbersome to link to each other. DataLad provides seamless nesting mechanisms:

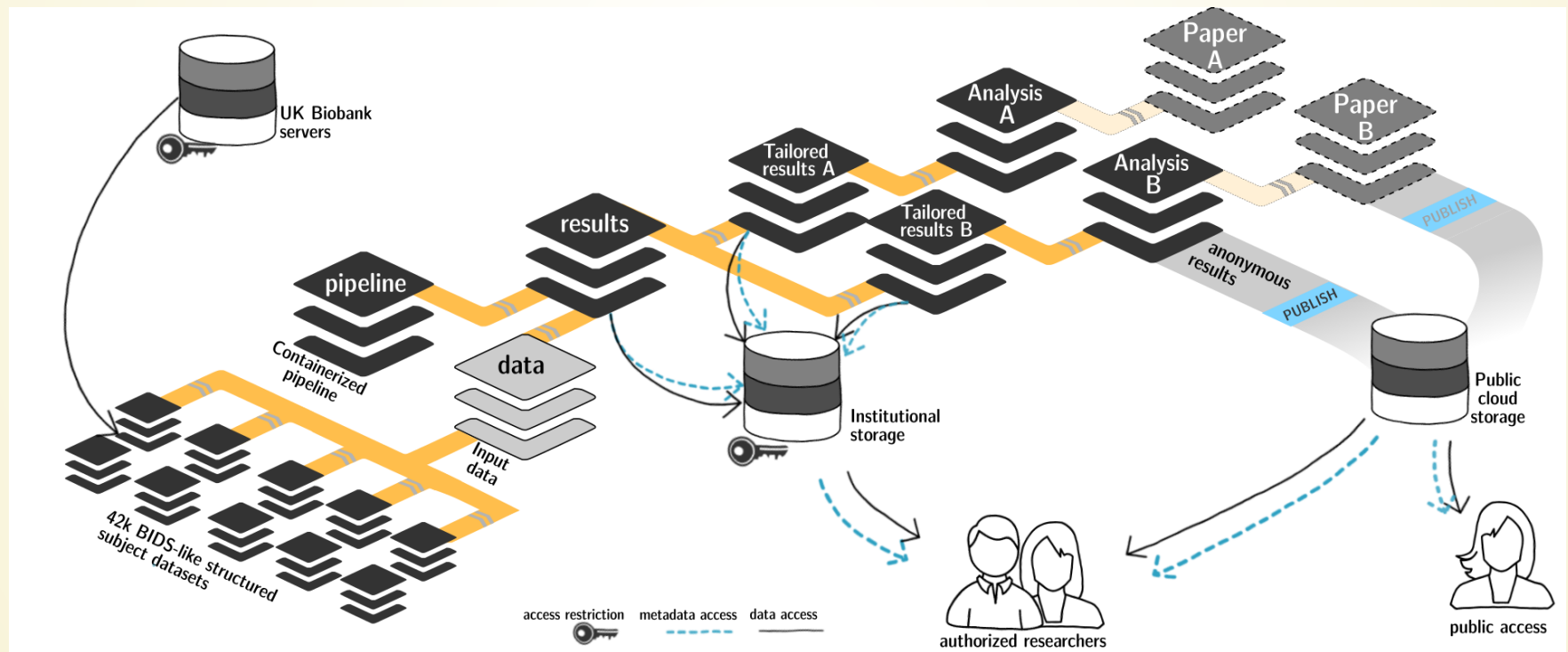


- Modularizes research components for transparency, reuse, and access management
- Overcomes scaling issues with large amounts of files

```
adina@bulk1 in /ds/hcp/super on git:master) datalad status --annex -r
15530572 annex'd files (77.9 TB recorded total size)
nothing to save, working tree clean
```

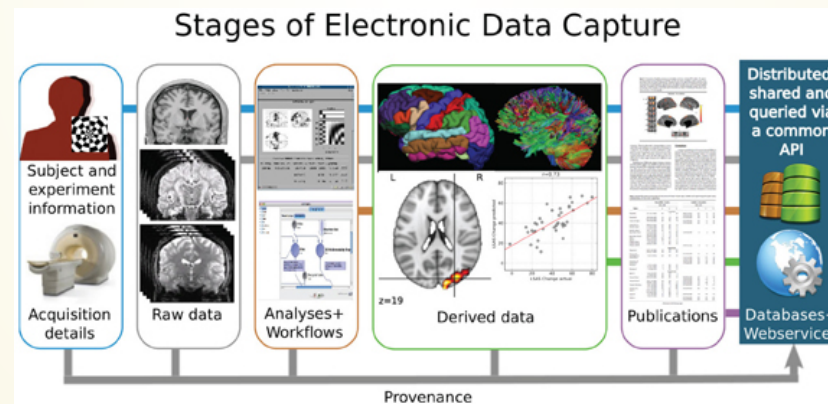
PRIVACY

fine-grained privacy decisions: git-annex or your choice of hosting service allow to keep files or datasets accessible only to you or authorized users.



MACHINE-READABLE, RE-EXECUTABLE PROVENANCE

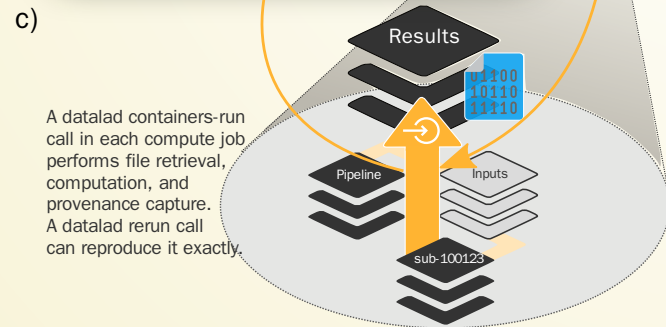
Much of neuroscientific research is computationally intensive, with complex workflows from raw data to result, and plenty of researchers degrees of freedom



MACHINE-READABLE, RE-EXECUTABLE PROVENANCE

```
a)
# perform and capture a computational execution
$ datalad containers-run \
-m "Compute subject ${subid}" \
-n cat \
--input "inputs/${subid}/*T1w.nii.gz" \
--output "${subid}" \
"<arguments for container invocation>"
```

```
b)
# recompute a previous computation
$ datalad rerun e035f896s45c9
```



```
d)
commit e035f896s45c9fac70cn7cc4dbd0dad43907755p
Author: Jane Doe <j.doe@fz-juelich.de>
AuthorDate: Wed Feb 10 18:05:30 2021 +0100
Commit: Jane Doe <j.doe@fz-juelich.de>
CommitDate: Wed Feb 10 18:05:30 2021 +0100

[DATALAD RUNCMD] Compute sub-6025043/ses-2

=== Do not change lines below ===
{
"chain": [],
"cmd": "singularity exec -B {pwd} --cleanenv code/pipeline/.datalad/
environments/cat/image sh -e -u -x -c [...]"
"dsid": "8938de76-0302-45b5-9825-3c6ce3f3ffe",
"exit": 0,
"extra_inputs": [
"code/pipeline/.datalad/environments/cat/image"
],
"inputs": [
"inputs/ukb/sub-6025043/ses-2/anat/sub-6025043_ses-2_T1w.nii.gz",
"code/cat_standalone_batch.txt",
"code/finalize_job_outputs.sh"
],
"outputs": [
"sub-6025043/ses-2"
],
"pwd": "."
}
^^^ Do not change lines above ^^^
---
sub-6025043/ses-2/inforoi.tar.gz | 1 +
sub-6025043/ses-2/native.tar.gz | 1 +
sub-6025043/ses-2/surface.tar.gz | 1 +
sub-6025043/ses-2/vbm.tar.gz | 1 +
4 files changed, 4 insertions(+)
```

	Basic commit metadata
	Author, Agent, Date, Time, and Commit Message
	Transformations
	Command call/ Container parametrization
	Software container image
	Origin: http://containers.ds.inm7.de/ Version: dfa6d975ea888ed33bf714
	Input data
	Origin: http://ukb.ds.inm7.de/ Version: 0c7f0b45140dde1d7291b1f
	Expected output data/folder
	Captured output data
	Path, Content hash

TAKE HOME MESSAGES

Data deserves version control

It changes and evolves just like code

git-annex and DataLad extend the advantages of Git and hosting services to your data

Increased transparency, better reproducibility, easier accessibility, efficiency through automation and collaboration, streamlined procedures for synchronizing and updating your work, ...

git-annex and DataLad have unique additional advantages

Have access to more data than you have disk space

Who needs short-term memory when you can have automatic provenance capture?

Link versioned data to your analysis at no disk-space cost

ACKNOWLEDGEMENTS

Funders



EUROPEAN UNION
European Regional Development Fund



Software

- Joey Hess (git-annex)
- The DataLad team & contributors

Illustrations

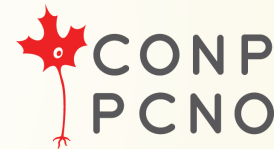
- The Turing Way project & Scriberia



Collaborators



Human Brain Project



VirtualBrainCloud



OpenNEURO



brainlife.io