



### How could I get data?

- \* Data "Installation":

1. Go to website(s) and



2. Extract

- \* Data "upgrades"

1. Read news

2. goto!

- \* Data "sharing"

(I have heard about

Upload button(s))

### What about software?

- \* Linux distros

```
apt-get install anything
```

```
apt-get upgrade
```

- \* Git-inspired folks

```
git clone git://anything
```

```
git pull
```

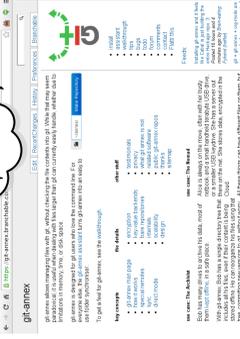
```
git push public
```

```
# send a pull request
```

Like that for data please?



Try git-annex



<http://datalad.org>

# Converging catalogues, warehouses, and deployment logistics into a federated 'data distribution'

Yaroslav O. Halchenko<sup>1</sup>, Michael Hanke<sup>2</sup>

- 1) Dept. of Psychological and Brain Sciences, Dartmouth College, NH, USA
- 2) Dept. of Experimental Psychology, University of Magdeburg, Germany

### What DataLad?

- 1 **crawl**

1. Automate data access and possibly scraping of meta-information from data portals, such as

- CRKNS.org
- HumanConnectome
- OpenfMRI
- ... anything else conceivable from XNAT, S3, ZONIS, ...

2. Store acquired data pointers under control git-annex

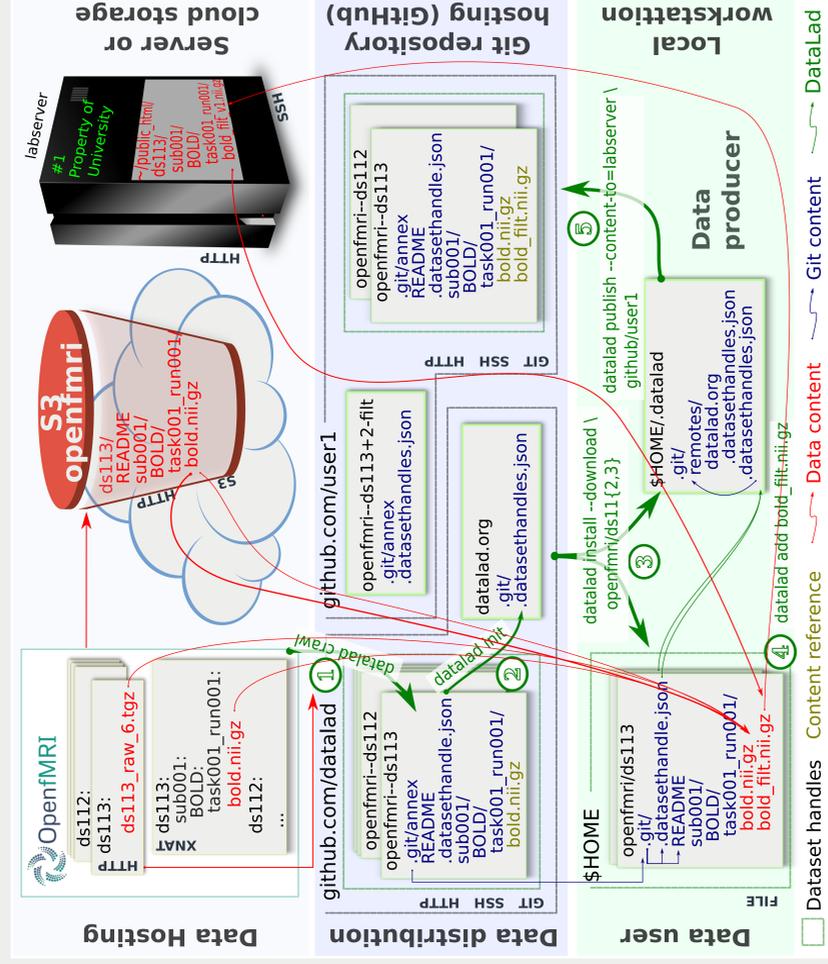
- 2 **init**

Collate available datasets into a collection; data distribution and provide sufficient meta-information for efficient search and discovery. See <http://dataprotocols.org/date-packages>



- 3 **install**

Use lean, locally installed dataset handles to obtain selected datasets with data content directly from original data providers (e.g. from S3 AWS storage, HTTP, XNAT, archives, etc.); same front interface, dedicated authentication and access mechanisms behind



- 4 **add**

Add new or derived data: it will be checked, versioned and you could never lose track of it ever again. Since it is a git repository after all -- you can also add your scripts, documentation, etc. as well.

- 5 **publish**

Publish individual or collections of updated dataset handles to a service like GitHub while copying file content to a local web-server (URL gets registered in the dataset handle) to make it publicly accessible. **Send a pull-request!**

**Bonus: Integrate**

<http://NeuroDebian.net>  
apt-get install \n openfmri-ds11{2,3}

### Take home messages from DataLads

- 1 **be legit**

- provision data sharing in the consistent forms  
- decide on copyright/license

- 2 **eat your own...**

- analyze data in a layout ready for sharing and anonymized (where possible)

- 3 **be anal**

- try to collect as much detail about your acquisition as possible (sequence protocols, BICOMs, associated BVA data, ...)  
- don't trust -- test your data!

- 4 **be "standard"**

- try to adhere to some convention on the hierarchy, e.g. of opening/ing  
- consider providing a dataset descriptor

- 5 **think forward**

- version your data!  
- dates in the filenames  
- or git-annex right away!  
- upload your data to some public data hosting portal (e.g. ccrns.org, openfmri.org, etc.)

Add your words of wisdom here

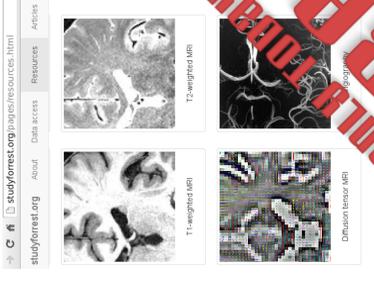
Add your words of wisdom here

Add your words of wisdom here

- 6 **Try git-annex**

git clone \n [http://psydata.ovgu.de/forrest\\_gump/git](http://psydata.ovgu.de/forrest_gump/git)

[studyforest.org](http://studyforest.org/psydata/resources.html)



### Acknowledgements!

- \* Linus Torvalds and Git team
- \* Joey Hess (git-annex)
- \* NSF grant #1429999



- \* <http://zerich.deviantart.com> (Joey's picture)
- \* [opendipart.org](http://opendipart.org) contributors

**SfN 2014**  
Visit us  
at booth  
#3419

### When DataLad?

