ADOPTING DATALAD FOR SHARING NEUROIMAGING DATA

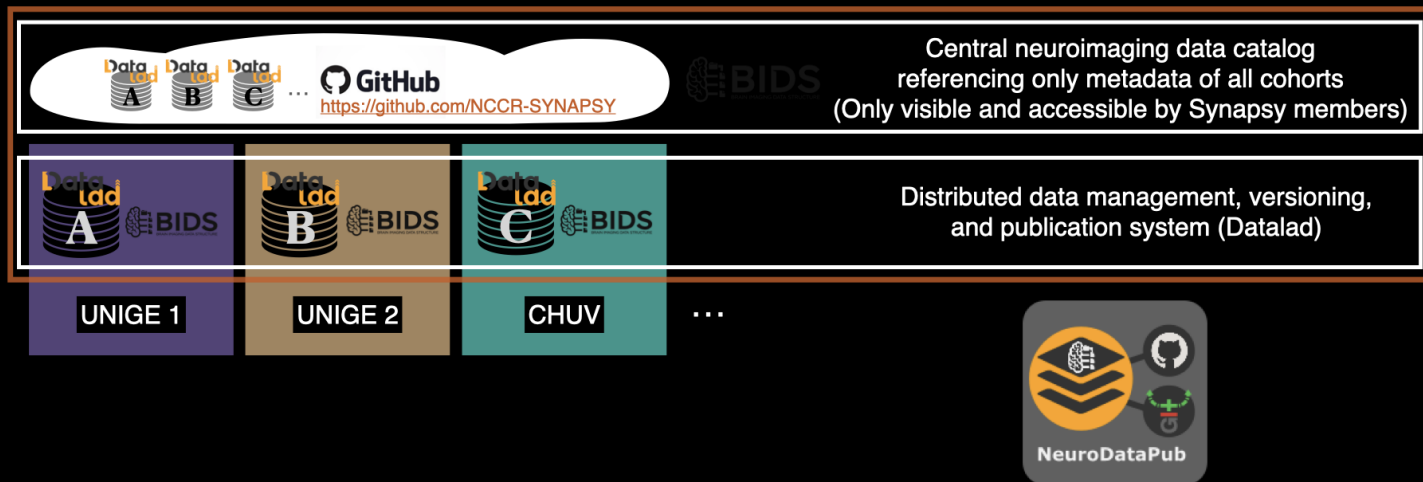Introduction to NeuroDataPub

by Sebastien Tourbier

31 08 2021

WHAT IS NEURODATAPUB?

NEURODATAPUB: NCCR-SYNAPSY NEUROIMAGING DATA PUBLISHING TOOL

THE AIMS OF NEURODATAPUB

- Support neuroimaging dataset catalogue creation and lower the barriers to adopt Datalad in order to fully implement Synapsy Neuroimaging Data Management Plan.
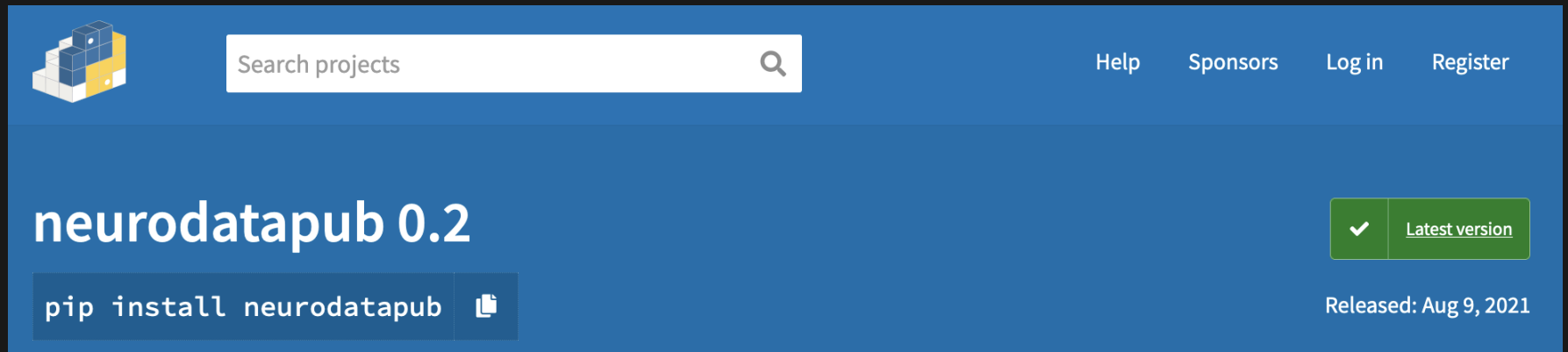
EASY TO INSTALL

- Provide a Conda environment with all dependencies installed (such as Datalad and extensions, git-annex) that can be easily created as follows:

```
$ conda env create -f "https://github.com/NCCR-SYNAPSY/neurodatapub/raw/main/conda/environment.yml"
```

- Neurodatapub is available on the Python Package Index (PyPI) and can be easily installed.
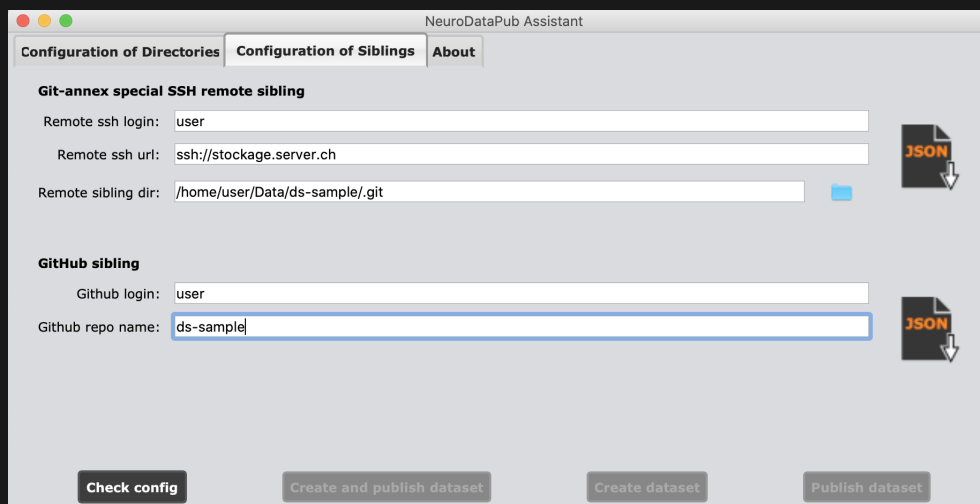
EASY TO USE

- Come with a graphical user interface, aka the NeuroDataPub Assistant

```
$ conda activate neurodatapub-env
(neurodatapub-env)$ neurodatapub --gui
```


- Created to facilitate:
  - sibling configuration,
  - creation of the JSON sibling configuration files,
  - execution of NeuroDataPub

REDUCE THE RISK OF ERRORS

- Execution of NeuroDataPub in 3 modes:
  - creation and publication of a datalad dataset,
  - creation of a datalad dataset only,
  - publication of an existing datalad dataset only.

- Control the configurations before enabling execution.
- Use JSON schema to validate the JSON sibling configuration files.

HOW TO GET STARTED

- Instructions for installation and usage available on ReadTheDocs

https://neurodatapub.readthedocs.io
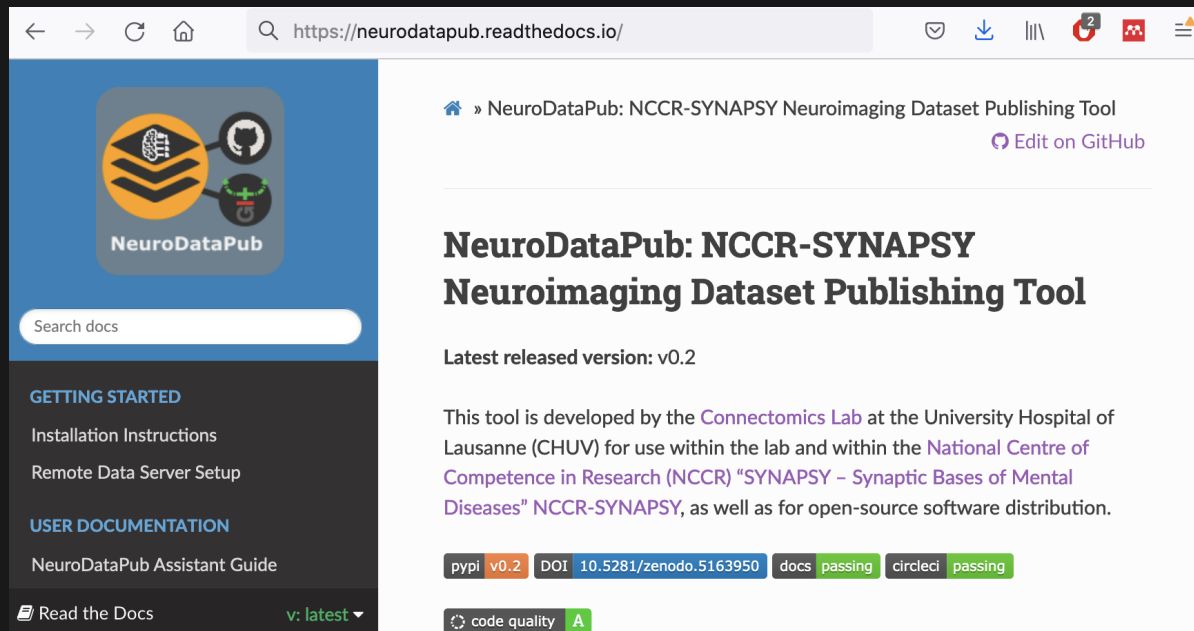
HOW TO INTERACT WITH A DATASET PUBLISHED BY NEURODATAPUB

INSTALL A LOCAL COPY OF THE PUBLISHED DATASET (NO CONTENT)

- Configure ssh by adding the following lines to the ~/.ssh/config file:

```
Host stockage.server.ch
        HostName stockage.server.ch
        User user
```

This will indicate automatically to ssh which user login to use to connect.
- Clone with Datalad the GitHub dataset repository:

```
(neurodatapub-env)$ datalad clone "git@github.com:NCCR-SYNAPSY/ds-NeuroDataPubTest13.git"
```

- **Note:** At this stage, the dataset does not contain any annexed files. The ssh information will be used when establishing the ssh connection to download the actual annexed content from the data storage server.

DOWNLOAD THE ANNEXED FILE CONTENT ON DEMAND

- For instance, one can download the annexed T1w image from the SSH data storage server with **datalad get** command the GitHub dataset repository:

```
(neurodatapub-env)$ datalad get -d "ds-NeuroDataPubTest13" "sub-01/ses-01/anat/*T1w.nii.gz"
```

- **Note:** If annexed files are published on a SSH data storage server, one might have an account to access it and be able to download the image.

HOW TO UPDATE THE PUBLISH A NEW VERSION OF THE DATASET

- Update the CHANGES of the BIDS dataset where you will refer to the new dataset version, for instance:

```
1.0.2 2021-08-27
        - Fix participants.tsv
```

- Save the state of the dataset with **datalad save** with the **-t** tag which specifies the new version and push the changes to GitHub sibling with **datalad push**:

```
(neurodatapub-env)$ datalad save -d "ds-NeuroDataPubTest13" -m 'Brief change description' -t 1.0.2
(neurodatapub-env)$ datalad push -d "ds-NeuroDataPubTest13" --to github
```

THANK YOU FOR YOUR ATTENTION!