# Adding an Open Science Framework (OSF) Dataset to Khanlab-Datasets

Jason Kai

May 13, 2019

## Disclaimer

This documentation is written primarily for the use of converting datasets within the Khanlab for further distribution via Datalad with the option to host through the Canadian Open Neuroscience Platform (CONP) portal.

The documentation written assumes the dataset is already hosted on Open Science Frameowrk (OSF). No assumptions are made regarding how the data may be organized.

## Datalad-osf

To more easily convert a dataset from OSF to Datalad, `datalad-osf`, the utility scripts created by TemplateFlow have been modified and made available at the following Github repository (link to be added).

A command-line version of the tool has been created and can be called using the command `datalad-osf` following installation. This tool may be included in existing Khanlab containers in the future. Setup of this utility requires Python 3 and a number of Python dependencies (see `requirements.txt` contained in the repository).

### Usage

```
datalad-osf <project key> [-s <subset>]
```

`<project key>` - The set of characters following the project's URL. For example, the project key for `BigBrainHippoUnfold` is `x542s` (https://osf.io/x542s)

`<subset>` - An optional argument; converts only a subset of data hosted within osf. This can be a path to any valid directory within the OSF project relative to the top level of the project

## Dataset Conversion

*Note: Conversion of the dataset will store a (temporary) copy of the data locally on the machine being used. Please ensure there is sufficient space.*

1. Create the Datalad dataset locally and navigate to the directory.

   ```
   datalad create -d <local directory>
   ```

2. Download the OSF (subset of) data to the datalad dataset.

   ```
   datalad-osf <key>
   ```

3. Using `datalad`, create the repository on Github. *(For Khanlab members, data will be hosted through the Khanlab-datasets organization. If you are not yet a member of the organization, please ask to be added)*

4. If you are hosting the data within your own repository, use the following command:

```
datalad create-sibling-github -d <local directory> <project_name>
```

   If data is to be hosted through an Github organization, use the following command:

```
datalad create-sibling-github --github-organization <org name> -d <local
   directory> <project_name>
```

5. Save and publish the changes.

```
datalad save
datalad publish --to github
```

## Host on CONP

6. Add `descriptor.json` file to the repository. Contents of the descriptor can be found below:

```
{
"title": "A descriptive title",
"description": "A description of the dataset - can include available
   modalities, goal of project, or link to paper"
"authors": "Authors of Dataset or Owning Organization",
"licenses": "License for distribution (e.g. CC-By Attribution 4.0
   International)",
"contact": "contact@email.com"
}
```

   Add the descriptor to the datalad dataset and publish to the repository.

```
datalad add --to-git ./descriptor.json
```

7. Create and add a README.md file directly in the repository if one does not already exist.

```
datalad add --to-git ./README.md
```

8. Save and publish modifications

```
datalad save
datalad publish --to github
```

***The remaining steps should be done by a single person / data manager of organization.***

9. Fork and install the conp-dataset repository. *(Forking of the dataset is not required if publishing through khanlab-datasets)*

```
datalad install git@github.com:<username>/conp-dataset
```

   If adding datasets from `khanlab-datasets`, first pull any changes to forked repository to ensure copy is up to date. Install from (https://github.com/khanlab-datasets/conp-dataset)

```
datalad install git@github.com/khanlab-datasets/conp-dataset
```

10. Navigate to the installed conp-dataset. Add dataset to be shared as a submodule to the conp-dataset fork.

```
git submodule add http://github.com/khanlab-datasets/<project_name>.git
   investigators/Khanlab<project_name>
```

11. Save and publish the modifications to fork.

```
datalad add --to-git .gitmodules
datalad save
datalad publish
```

12. Create pull request to merge dataset into main conp-dataset repository with additional dataset.