September 6, 2019

CISBAT 2019 workshop
“Open science”

Sharing your code

A quick guide to Git

Dr Roberto Castello
EPFL Solar Energy and Building Physics Laboratory (LESO-PB)
roberto.castello@epfl.ch

Git

A free and open source distributed version control system

- Git documentation:
  - https://git-scm.com/

- Installing Git:
  https://git-scm.com/book/en/v2/Getting-Started-Installing-Git

create a new directory, open it and perform a

```
git init
```

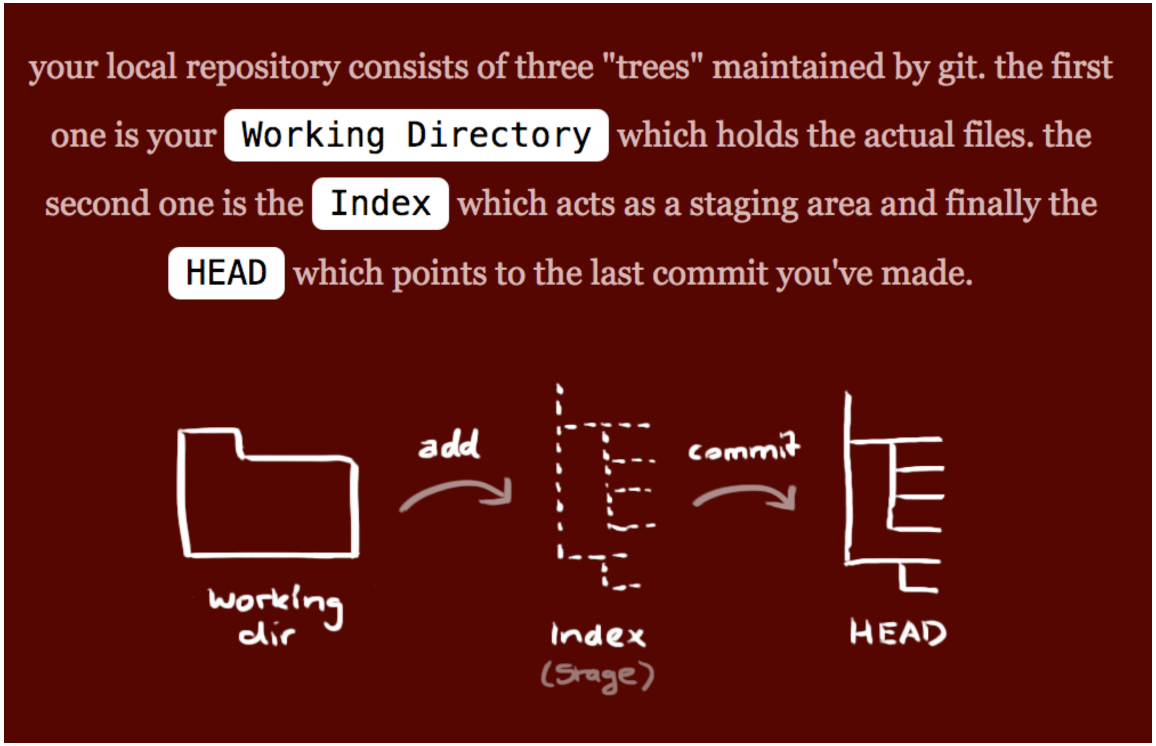
to create a new git repository.

(*) Most of the material is taken from https://rogerdudler.github.io/git-guide/

Checkout a repository

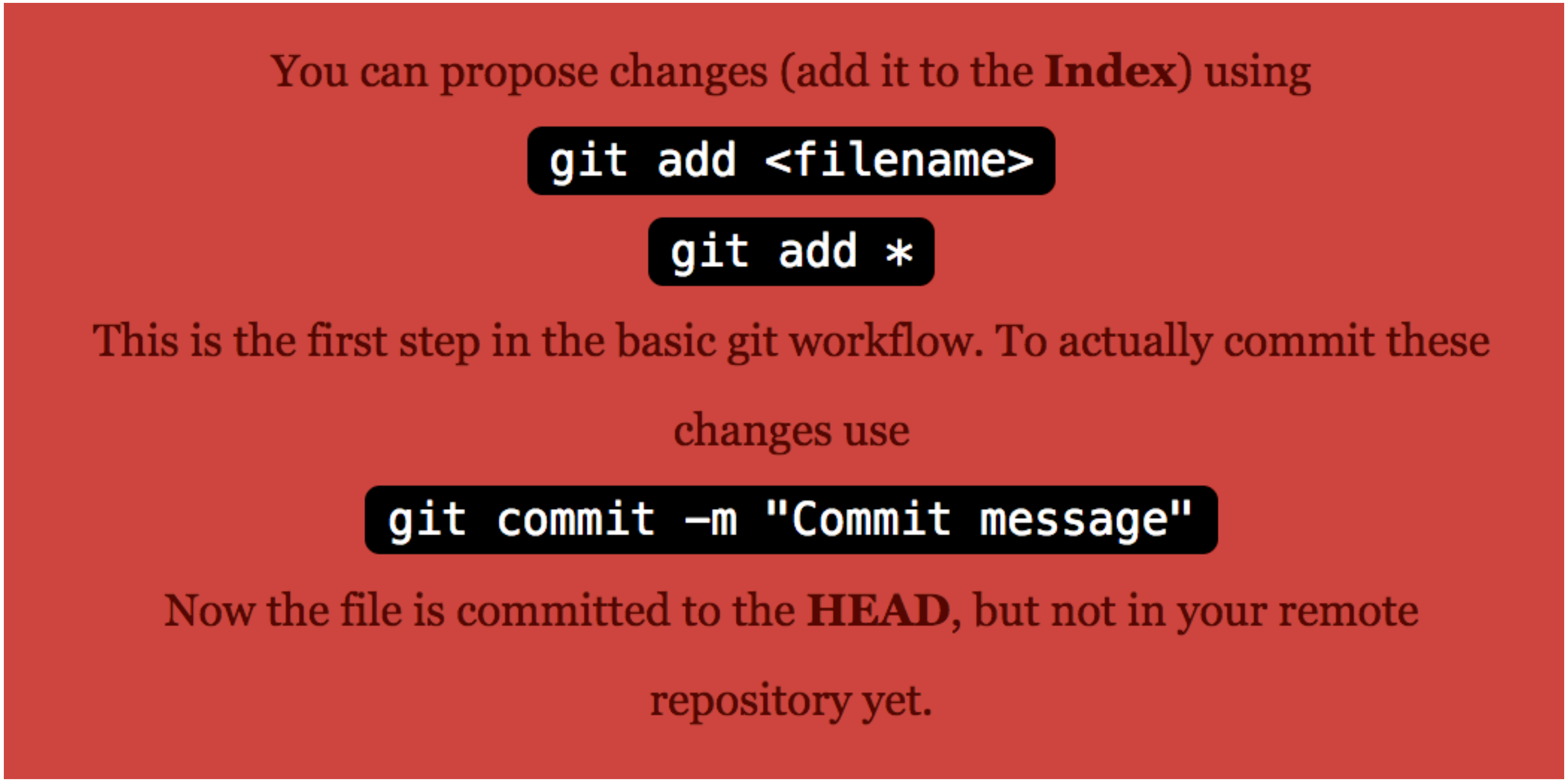
create a working copy of a local repository by running the command

```
git clone /path/to/repository
```



your local repository consists of three "trees" maintained by git. the first one is your `Working Directory` which holds the actual files. the second one is the `Index` which acts as a staging area and finally the `HEAD` which points to the last commit you've made.

Add files and commit



You can propose changes (add it to the **Index**) using

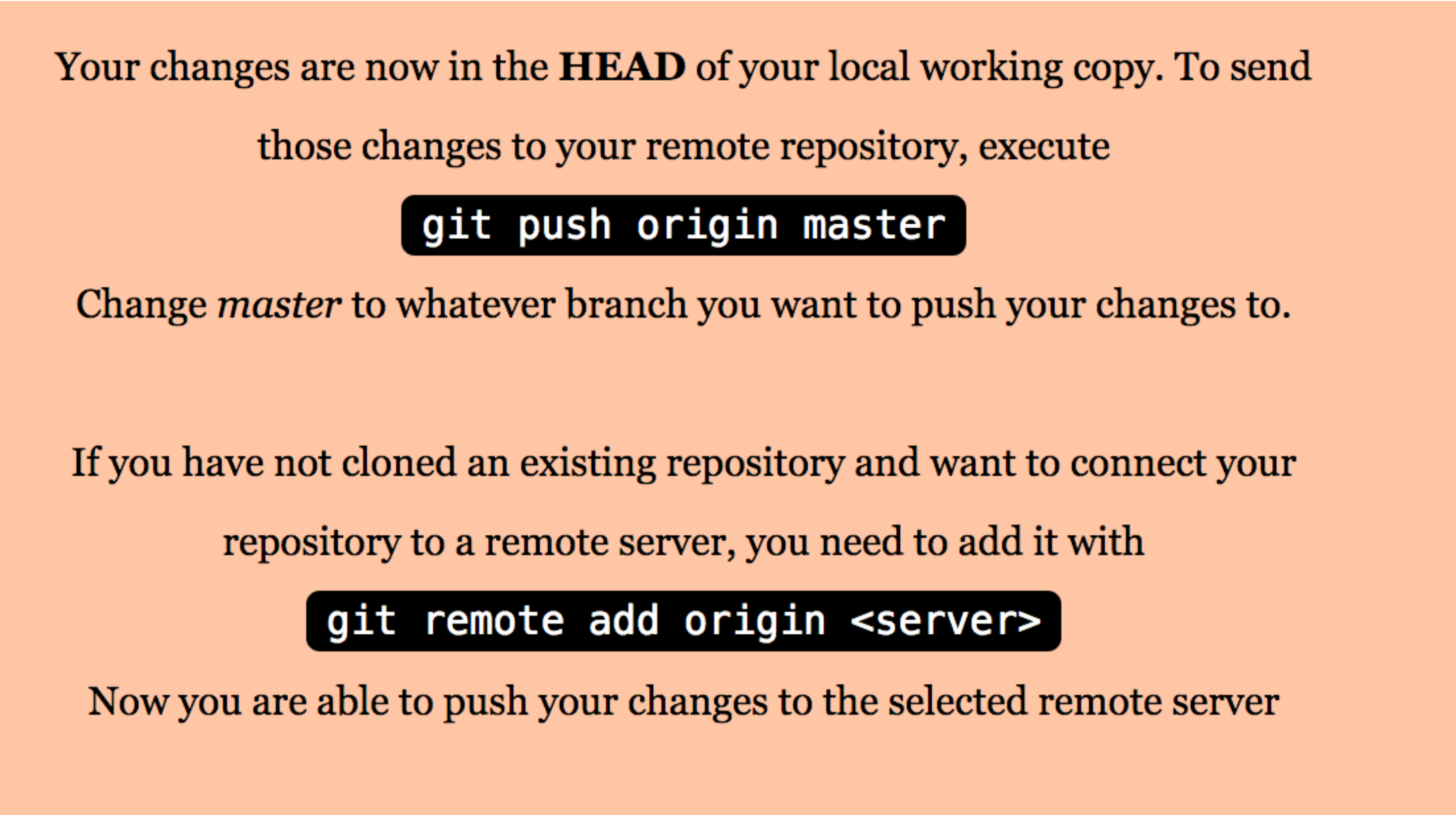
```
git add <filename>
git add *
```

This is the first step in the basic git workflow. To actually commit these changes use

```
git commit -m "Commit message"
```

Now the file is committed to the **HEAD**, but not in your remote repository yet.

Pushing changes



Your changes are now in the **HEAD** of your local working copy. To send those changes to your remote repository, execute

```
git push origin master
```

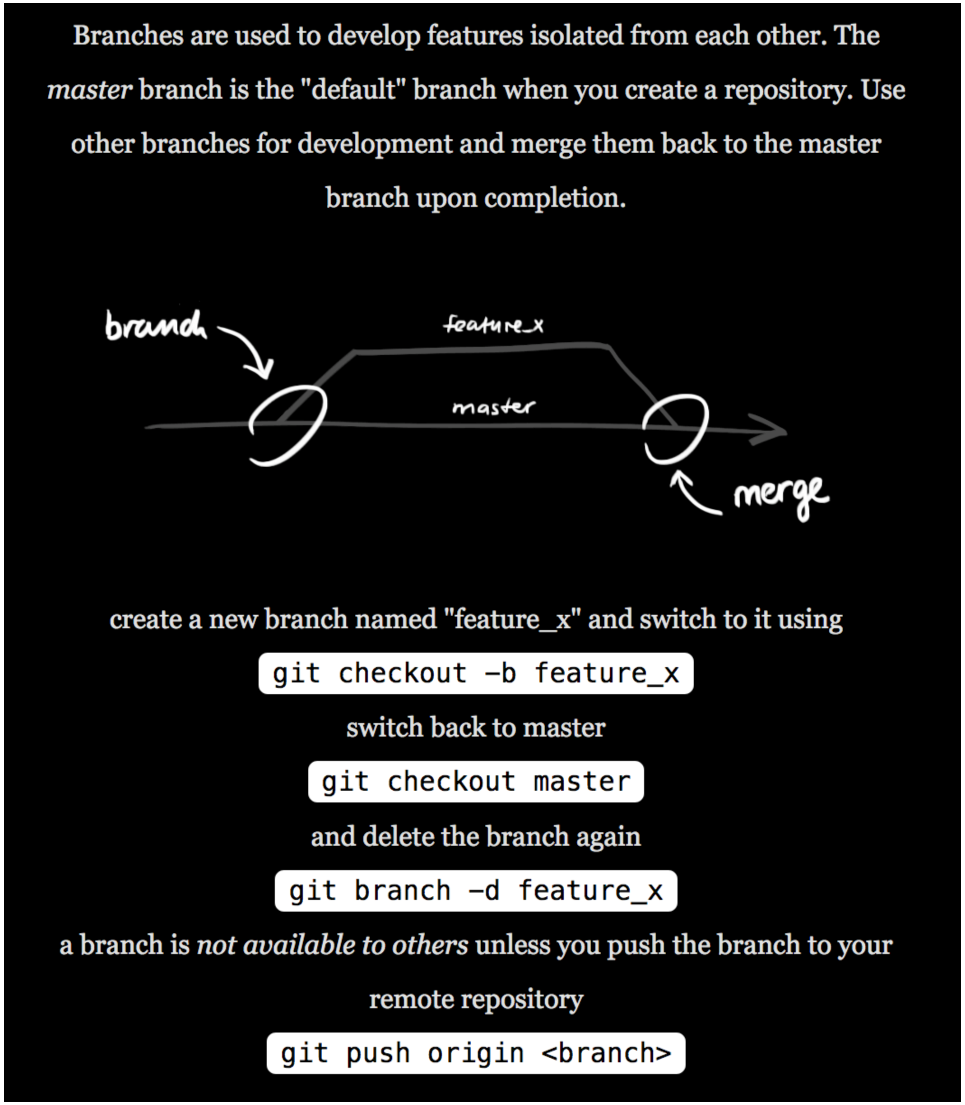
Change *master* to whatever branch you want to push your changes to.

If you have not cloned an existing repository and want to connect your repository to a remote server, you need to add it with

```
git remote add origin <server>
```

Now you are able to push your changes to the selected remote server

Changing branch



Branches are used to develop features isolated from each other. The *master* branch is the "default" branch when you create a repository. Use other branches for development and merge them back to the master branch upon completion.

create a new branch named "feature_x" and switch to it using

```
git checkout -b feature_x
```

switch back to master

```
git checkout master
```

and delete the branch again

```
git branch -d feature_x
```

a branch is *not available to others* unless you push the branch to your remote repository

```
git push origin <branch>
```

A usual workflow

1. First create an account on GitHub: https://github.com/
2. Once you have an account, Initialize your workspace

   ```
   git init
   ```
3. Clone your favorite repository

   ```
   git clone <my_repo>
   ```
4. Create a new branch for your own development

   ```
   git checkout -b <my branch>
   ```
   - N.B. 3. and 4. are needed only the first time. Otherwise just get in sync with master (or your branch) `git pull origin master` and then switch to your branch `git checkout <my branch>`
5. Edit your file(s) and stage the file(s) for commit

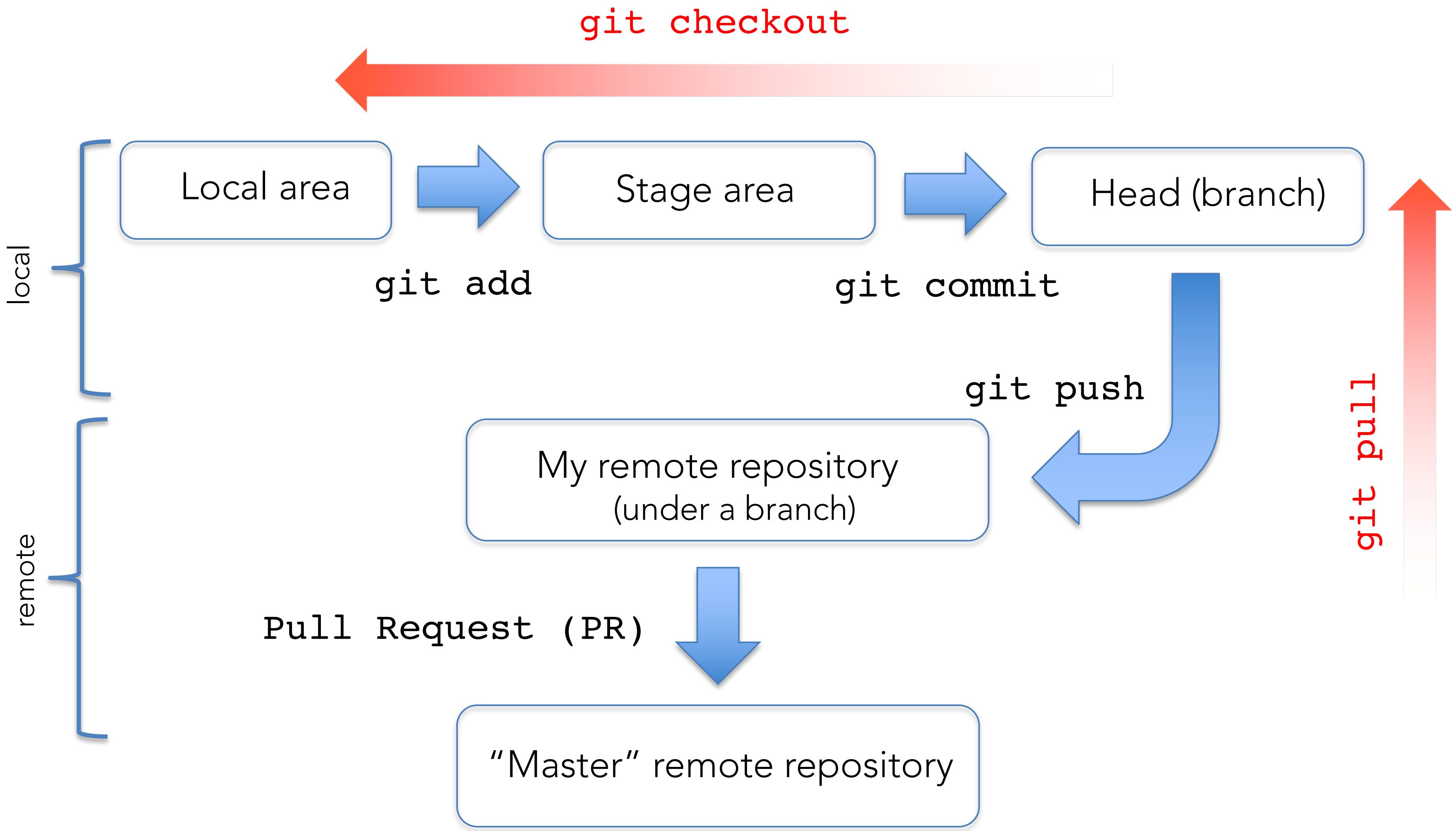
   ```
   git add <file>
   ```
6. Always check what is the status

   ```
   git status
   ```
7. Let's commit (whatever has been staged) to the Head

   ```
   git commit -m “Write a comment here”
   ```
8. Push commits to the remote repository

   ```
   git push origin <my branch>
   ```

The Git flow



git checkout
local
Local area
git add
Stage area
git commit
Head (branch)
git push
git pull
remote
My remote repository
(under a branch)
Pull Request (PR)
"Master" remote repository

September 6, 2019

CISBAT 2019 workshop
“Open science”

Data publication

How to share your data and make them reproducible

Dr Roberto Castello
EPFL Solar Energy and Building Physics Laboratory (LESO-PB)
roberto.castello@epfl.ch

FAIR Principles

A dataset is FAIR if it's:

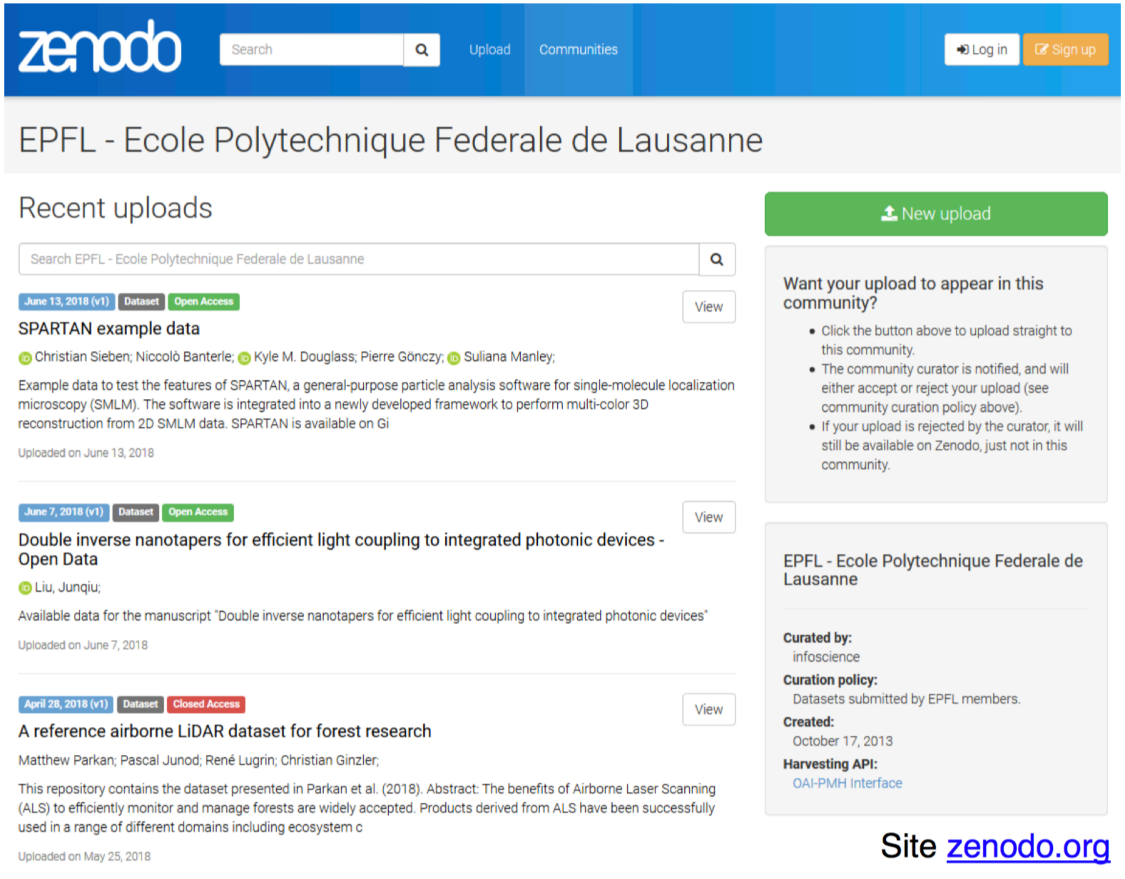
- Findable:
  Data and metadata are easy to find by both humans & computers.
  - Use metadata
  - Deposit (meta)data in repository/registry
  - Assign a persistent identifier (eg. DOI, HANDL, URN)
- Accessible:
  Machines & humans can readily access or download (meta)data.
  - As-open-as-possible access to your data (licensing, …)
  - Services with user-friendly interfaces
  - Make the metadata available after data deletion
- Interoperable:
  Data from different datasets are ready to be exchanged or combined.
  - Use open file format(s), whenever possible
  - Use standardized vocabularies/tags
  - Use cross-linking as much as possible
- Reusable:
  (Meta)data are easily replicated / combined in future research.
  - Attach standardized license to your data (CC, GPL, …)
  - Capture provenance information as precisely as possible

Data repositories

NAME	DISCIPLINE	NON-PROFIT / INSTITUT.	COUNTRY	FREE	MAX VOLUME	LICENSING
zenodo	Generic	✓ (CERN)	🇨🇭	✓	50GB/dataset, ∞ datasets	CC, GNU, BSD
MATERIALSCLOUD	STI / Materials	✓ (EPFL)	🇨🇭	✓	5GB General / 50GB AiiDa DB	CC-BY (MIT for Aiida)
figshare	Generic	✗ (Holtzbrinck Group)	🇺🇸	Freemium	1 TB per dataset	CC0, CC-BY
DRYAD	Bio / Medical	✓ (?)	🇺🇸	✗	?	CC0
The Dataverse Project	Generic	✓ (Harvard University)	🇺🇸	✓	?	?
EUDAT	Generic	✓ (HORIZON 2020)	🇪🇺	✓	?	CC (DARUP)

The SNSF encourages the use of **re3data.org**[52]. Also check the data repositories recommended by the ERC Scientific Council[86]

Data repository example



EPFL - Ecole Polytechnique Federale de Lausanne

Recent uploads

Search EPFL - Ecole Polytechnique Federale de Lausanne

June 13, 2018 (v1) Dataset Open Access

SPARTAN example data

Christian Sieben; Niccolò Banterle; Kyle M. Douglass; Pierre Gönczy; Suliana Manley;

Example data to test the features of SPARTAN, a general-purpose particle analysis software for single-molecule localization microscopy (SMLM). The software is integrated into a newly developed framework to perform multi-color 3D reconstruction from 2D SMLM data. SPARTAN is available on Gi

Uploaded on June 13, 2018

June 7, 2018 (v1) Dataset Open Access

Double inverse nanotapers for efficient light coupling to integrated photonic devices - Open Data

Liu, Junqiu;

Available data for the manuscript "Double inverse nanotapers for efficient light coupling to integrated photonic devices"

Uploaded on June 7, 2018

April 28, 2018 (v1) Dataset Closed Access

A reference airborne LiDAR dataset for forest research

Matthew Parkan; Pascal Junod; René Lugrin; Christian Ginzler;

This repository contains the dataset presented in Parkan et al. (2018). Abstract: The benefits of Airborne Laser Scanning (ALS) to efficiently monitor and manage forests are widely accepted. Products derived from ALS have been successfully used in a range of different domains including ecosystem c

Uploaded on May 25, 2018

Want your upload to appear in this community?

- Click the button above to upload straight to this community.
- The community curator is notified, and will either accept or reject your upload (see community curation policy above).
- If your upload is rejected by the curator, it will still be available on Zenodo, just not in this community.

EPFL - Ecole Polytechnique Federale de Lausanne

Curated by:
infoscience

Curation policy:
Datasets submitted by EPFL members.

Created:
October 17, 2013

Harvesting API:
OAI-PMH Interface

Site zenodo.org

- Hosted by the CERN
- Free of charges
- Unlimited datasets (max 50GB/dataset)
- Metadata harvesting
- All file formats accepted

- OpenAIRE integration (EC reporting)
- Automated DOI assignement
- Usage statistics interface
- GitHub integration
- ORCID integration

- OAI-PMH protocol (content harvesting)
- 18 petabytes disk cluster
- Each file has 2 replicas on different servers
- 2 independent MD5 checksums per file
- Metadata 12-hourly backup cycle

- ... etc. ...

SDSC

WHY RENKU?

Reproducibility

A real Git for data science, RENKU fosters reproducible research by enabling scientists to retrieve history and data provenance, and go back in time to every step of published science.

Reusability and repetition

The platform facilitates the sharing and reuse of data and algorithms, and empowers specialists to use other people's work in their own projects and execute them in an infrastructure agnostic environment. Attributions are therefore also consistently guaranteed.

Collaboration

RENKU supports a collaborative environment for dynamic and interactive prototyping by enabling content-rich discussions.

Security

Data is safe with RENKU. It makes use of state of the art security and privacy preserving technologies and best practices. It will give fine grained control over who accesses any data, from where and how.

Federation

RENKU is designed to connect independently administered platforms and positions itself as a unique one-stop shop for high quality data by allowing a federated access across institutions, giving each the freedom to enforce its own access controls over resources.

Discovery

Thanks to its automatically maintained and enriched knowledge graph, the platform supports targeted exploration as well as unforeseen discoveries by giving scientists access to the big picture through interconnected metadata. Science is thus described in a straightforward, intelligible manner.