Curate and publish your FAIR dataset with Solidipes

Son Pham-Ba[1] & Guillaume Anciaux[2]

[1]ENAC-IT4Research
[2]Computational Solid Mechanics Laboratory

École polytechnique fédérale de Lausanne

Funded by ETH BOARD
Open Research Data (ORD) Program

Introduction

I am a **researcher**, and want to **publish my dataset**.

How can I ensure that:
- My dataset can be cross-referenced in an article?
- My dataset is openly accessible?
- Its files can be opened and viewed by anyone?
- Its creation process is documented and easily reproducible?

Solidipes is a Python package that supports the **curation**, **publication**, and **sharing** of **research data**. It is named after *Armillaria solidipes*, the species of the largest living organism on Earth, a fungus forming an underground **network** spanning 9 km².

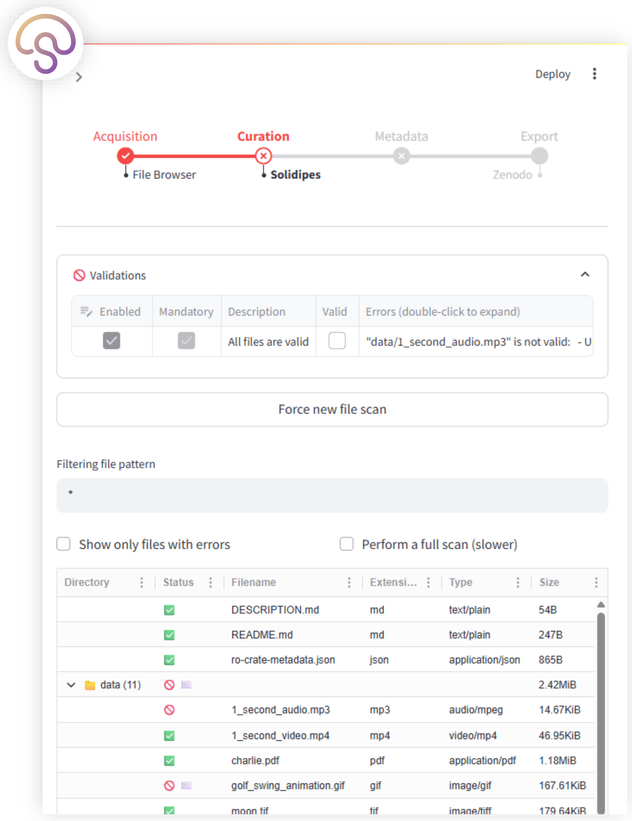
Getting started

To install Solidipes, run the following command:
(better in a Python virtual environment!)

```
$ pip install solidipes
```

Launch a curation session in your dataset:

```
$ solidipes report web
```



1 Acquisition

Run Solidipes on your dataset or create one from scratch to get started with publishing.

Get the most out of Solidipes' features by using version control (*e.g.* Git).

Version control GitLab

Data from various **cloud storage** services can accessed and edited seamlessly without copying.

NFS, S3, SMB, SSH, dtool...

Any already published dataset (along with its metadata) can be **retrieved** for exploration or curation.

2 Curation

Organize, clean, annotate your dataset, ensuring relevance and reusability. In short: make it **FAIR**.

A set of **validations** is performed on each file:
- Identify encoding/file format
- Extract metadata (header, properties)
- Attempt to (partially) load the file
- Linting on source code or scripts

A **data curator** can simultaneously access the platform to comment on files.

Field-specific files can be **loaded**, **viewed**, and curated using custom or **community-made** Solidipes **plugins**.

Solidipes

On your computer OR On an online instance *e.g.* dcsm.epfl.ch or renku

3 Metadata

Essential **metadata** includes:
- Authors, affiliations, ORCID
- Title
- Keywords
- License
- Related publications

A detailed **description** should include:
- Author contributions
- Directory structure
- Data collection method with details for reproduction
- Funding sources

The metadata and description are automatically merged into a comprehensive formatted **README.md** file.

4 Export

The curated dataset can be exported to **long-term storage** platforms (*e.g.* Zenodo or institutional repositories) with an associated **permanent identifier** (DOI).

At any stage, the dataset can be exported to an online instance of Solidipes (*e.g.* on Renku), for **sharing** a live preview of the dataset with a data curator or other scientists.

FAIR principles for datasets

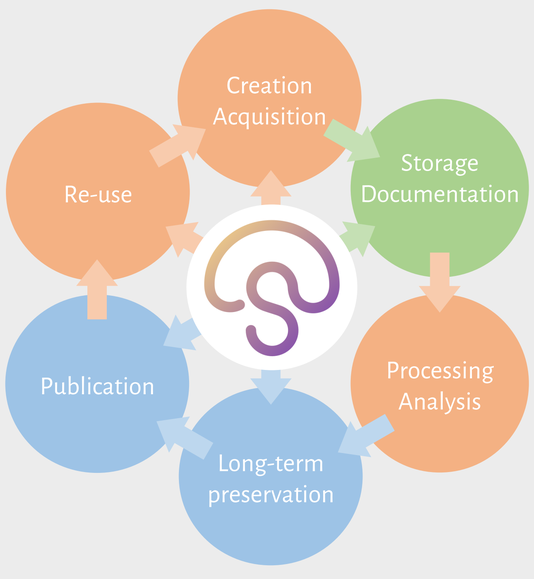
Findable
- Annotation, metadata: authors, keywords, cross-links, ownership, *etc.*
- Digital Object Identifier (DOI)

Accessible
- Open data repository
- Guaranteed etention time

Interoperable
- Use standards: file formats, metadata, vocabularies, ontologies, *etc.*

Reusable
- Open (source) license
- Environnent: software versions, dependencies, *etc.*

Research data lifecycle



Solidipes is made with Python 3 and Streamlit

Solidipes is used in the dataset curation process for the **Diamond open-access** Journal of Theoretical, Computational and Applied Mechanics JTCAM

Next steps

Your contribution is welcome!
- Integrate **ontologies** for FAIR datasets, ideally field-specific
- Show **workflows**, *i.e.* the steps and environment needed to generate the dataset
- Provide guidance on *[store output data]* versus *[only keep scripts to regenerate data]*, to minimize **CO_2 impact**

Scan to explore a live demo!
https://go.epfl.ch/solidipes-demo