

Data curation procedure at JTCAM
On the usage of **Solidipes** and Jupyter Notebooks

Guillaume Anciaux
LSMS, IIC, ENAC, EPFL



<https://gitlab.com/dcsn/solidipes>

Definitions: CODATA RDM Terminology

Data Curation ensures usefulness for discovery and reuse.

- *validation of the data (encoding, file formats)*
- *description completeness (documentation, annotation)*
- *discipline dependent soundness*
- *made by journals, universities (arXiv, HAL) and specialized archivists (e.g. astronomy with NASA, ESA, neurologists and brain MRI)*

Definitions: CODATA RDM Terminology

Data Curation ensures usefulness for discovery and reuse.

- *validation of the data (encoding, file formats)*
- *description completeness (documentation, annotation)*
- *discipline dependent soundness*
- *made by journals, universities (arXiv, HAL) and specialized archivists (e.g. astronomy with NASA, ESA, neurologists and brain MRI)*

Data curation is a central pillar in fostering scientific discoveries.

JTCAM dataset curation policy

The following criteria are required in order to accept a submission to the JTCAM community:

- Must be Open Access
- Ownership described in depth
- Detailed description (using standard ontologies or controlled vocabularies)
- Cross-linked reference must be added
- Software permanent links (Software Heritage)
- Acknowledged grants
- Cleaned (no unnecessary files/folders or redundancy)
- Permissive licenses are required (CC0, CC-BY-4.0)
- Files formats are open
- Workflow description

<https://zenodo.org/communities/jtcam/curation-policy>

Publishing and curating software

Mature development tools

- The distributed **git** version control system
- Developer platforms (**GitHub**, **GitLab**) allow discussions, reviews, suggestions and merging (e.g. the publication procedure of JOSS)
- ⚠ Permanent IDs: **Software Heritage**

Publishing and curating software

Mature development tools

- The distributed **git** version control system
- Developer platforms (**GitHub**, **GitLab**) allow discussions, reviews, suggestions and merging (e.g. the publication procedure of JOSS)
- ⚠ Permanent IDs: **Software Heritage**

Reproducibility and regression tests

- + Continuous integration and unit tests → allows check for reproducibility of workflows
- + Needs to save context (Operating systems, dependencies, versions) → Docker containers
- **Needs to be conceived by authors**
- + Open source software warrants access and perenity
- + Machine learning community is pushing in this direction

Publishing datasets

Where? **Zenodo** ?

- Long term preservation ⇒ **for 20 more years from now**
- Generalist ⇒ stores anything, no checks
- Read only datasets after publication
- < 50 GB limit
- Modifiable metadata
- **Metadata descriptions** and **Communities** (e.g. **JTCAM Zenodo community**)

What content ? Who checks ?

- At the moment **no one but the owner** (Can publish anything)
- + Curation perspective (e.g. **Horizon-Zen project**)

Difficulty: Convincing researchers



<https://zenodo.org/records/10108736>

Difficulty: Convincing researchers



<https://zenodo.org/records/10108736>

Need for helper tools

DCSM Project

Project **Dissemination of Computational Solid Mechanics**(DCSM)

- Fund by **Open Research Data (ORD)**
- G. Anciaux (dev and supervision@EPFL), S. Pham-Ba (developer@EPFL)

Goals

- Provide a **cloud based** repository/storage/tool for **solidmechanics** community
- Simplify the **verification, analysis and annotation** (curation) of datasets
- Stand-alone tool for researchers to manipulate data **on their personal computer**
- Web platform <https://dcsm.epfl.ch>
- Used at JTCAM for **data reviews**

Solidipes



***Amillaria Solidipes** grows and spreads primarily underground, and is possibly the largest living organism on Earth by mass, area, and volume and is colloquially called the "Humongous fungus".* [[Wikipedia](#)]

Solidipes



Amillaria Solidipes grows and spreads primarily underground, and is possibly the largest living organism on Earth by mass, area, and volume and is colloquially called the "Humongous fungus". [[Wikipedia](#)]

... Nothing to do with **solids**

Solidipes



Amillaria Solidipes grows and spreads primarily underground, and is possibly the largest living organism on Earth by mass, area, and volume and is colloquially called the "Humongous fungus". [[Wikipedia](#)]

... Nothing to do with **solids**

```
pip install solidipes
```

Solidipes: analysis and curation tool

- 1 Access remote data-storage/repository (S3, ssh, Windows share)

Solidipes: analysis and curation tool

- 1 Access remote data-storage/repository (S3, ssh, Windows share)
- 2 Scan files

Solidipes: analysis and curation tool

- 1 Access remote data-storage/repository (S3, ssh, Windows share)
- 2 Scan files
- 3 For each file
 - Identify the encoding/file format
 - Extract the metadata (CSV headers, image properties, finite element field descriptions)
 - Attempt a (partial) loading of the file
 - If any perform additional validation checks

Solidipes: analysis and curation tool

- 1 Access remote data-storage/repository (S3, ssh, Windows share)
- 2 Scan files
- 3 For each file
 - Identify the encoding/file format
 - Extract the metadata (CSV headers, image properties, finite element field descriptions)
 - Attempt a (partial) loading of the file
 - If any perform additional validation checks
- 4 Generates a validating report in either
 - text mode (terminal)
 - Jupyter notebook
 - WebApp allowing to graphical scrutiny (images, interactive 3D rendering, ...)

Solidipes: analysis and curation tool

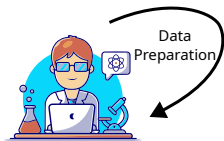
- 1 Access remote data-storage/repository (S3, ssh, Windows share)
- 2 Scan files
- 3 For each file
 - Identify the encoding/file format
 - Extract the metadata (CSV headers, image properties, finite element field descriptions)
 - Attempt a (partial) loading of the file
 - If any perform additional validation checks
- 4 Generates a validating report in either
 - text mode (terminal)
 - Jupyter notebook
 - WebApp allowing to graphical scrutiny (images, interactive 3D rendering, ...)
- 5 If validated: enables export to Zenodo/Renku

Demonstration

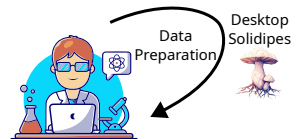
Solidipes: analysis and curation tool



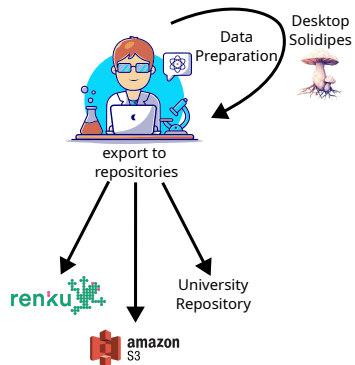
Solidipes: analysis and curation tool



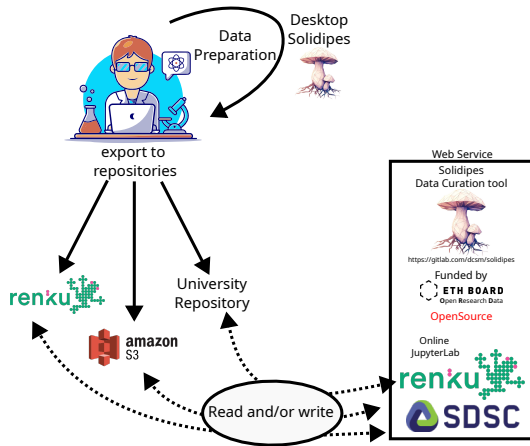
Solidipes: analysis and curation tool



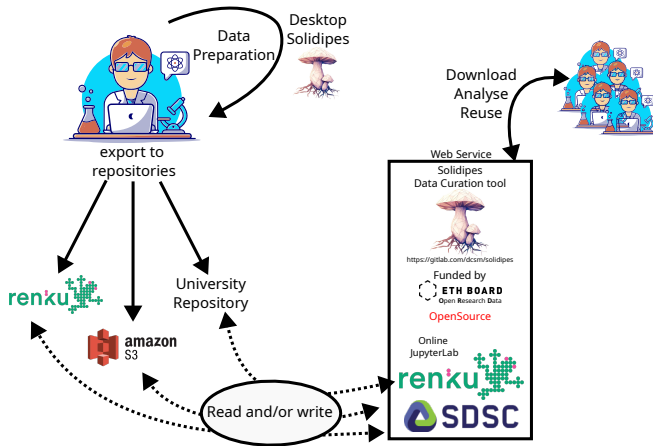
Solidipes: analysis and curation tool



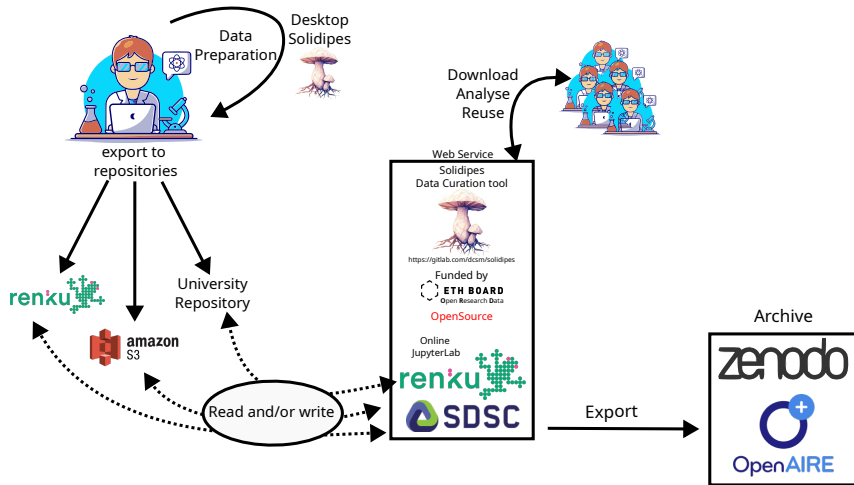
Solidipes: analysis and curation tool



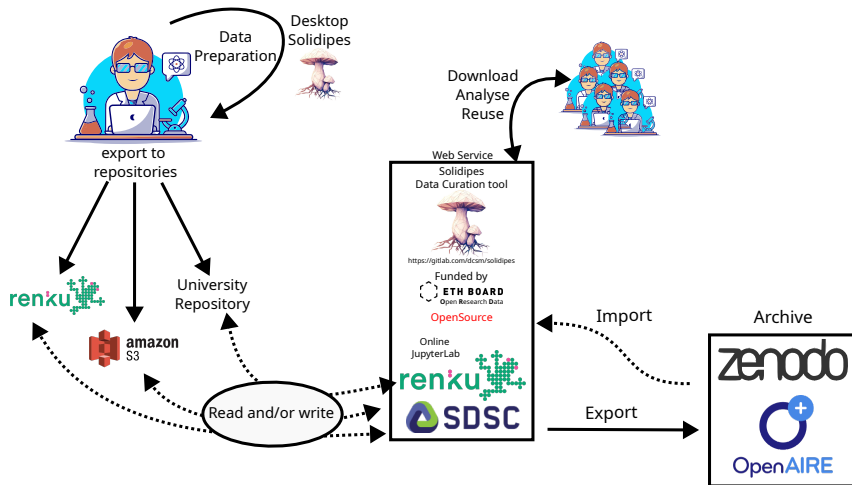
Solidipes: analysis and curation tool



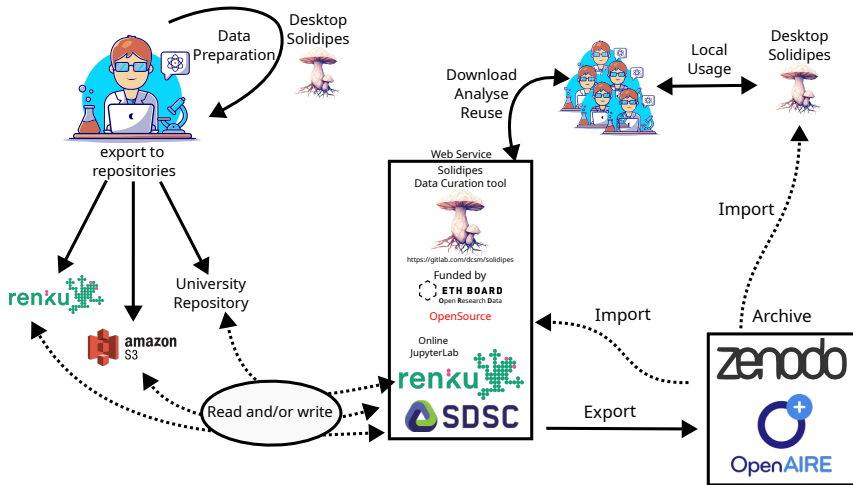
Solidipes: analysis and curation tool



Solidipes: analysis and curation tool



Solidipes: analysis and curation tool



Solidipes: analysis and curation tool

Features


- Analysis: Jupyterlabs and context preserving
- Curation: dedicated readers&viewers (web oriented)
- Export/Import/Mount (S3, samba, nfs, Zenodo repositories)
- Operating Context saved (**Docker** containerization)

Demo

- E. Eid, R. Seghir, & J. Réthoré. Accompanying data for the paper "Crack branching at low tip speeds: spilling the T"
- [Zenodo](#)
- [Renku](#)
- [Curation session](#)


Web Service

Solidipes
Data Curation tool



<https://gitlab.com/dcsn/solidipes>


Funded by



ETH BOARD
Open Research Data

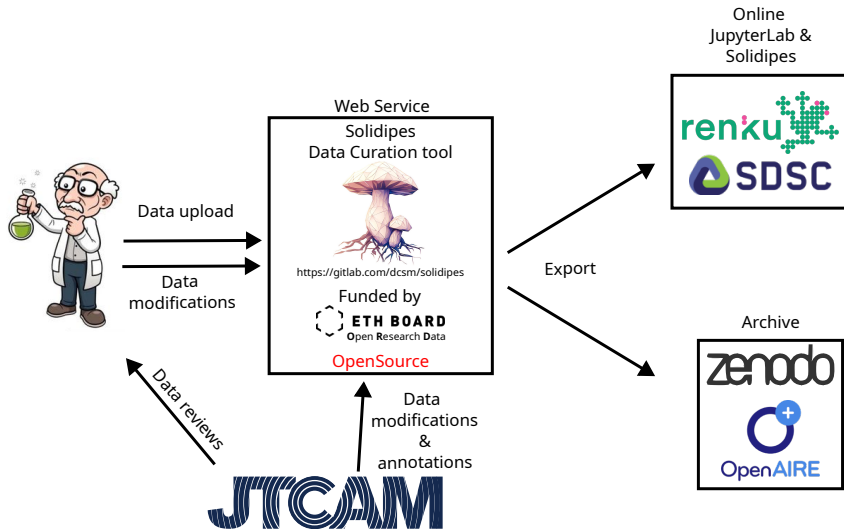
OpenSource

Online
JupyterLab



SDSC

Dataset Curation Management@JTCAM



Conclusion

Where we are

- **Data curation** with loaders/viewers (mostly for continuum solidmechanics)

Conclusion

Where we are

- **Data curation** with loaders/viewers (mostly for continuum solidmechanics)
- **Flexible** (remote) data storage

Conclusion

Where we are

- **Data curation** with loaders/viewers (mostly for continuum solidmechanics)
- **Flexible** (remote) data storage
- Publication and archive on **Zenodo/Renku**

Conclusion

Where we are

- **Data curation** with loaders/viewers (mostly for continuum solidmechanics)
- **Flexible** (remote) data storage
- Publication and archive on **Zenodo/Renku**
- **JTCAM curation policy** enforced with Solidipes@DCSM already

Conclusion

Where we are

- **Data curation** with loaders/viewers (mostly for continuum solidmechanics)
- **Flexible** (remote) data storage
- Publication and archive on **Zenodo/Renku**
- **JTCAM curation policy** enforced with Solidipes@DCSM already
⇒ Brings good principles to this Diamond open access initiative
- User **Documentation**

Conclusion

Where we are

- **Data curation** with loaders/viewers (mostly for continuum solidmechanics)
- **Flexible** (remote) data storage
- Publication and archive on **Zenodo/Renku**
- **JTCAM curation policy** enforced with Solidipes@DCSM already
⇒ Brings good principles to this Diamond open access initiative
- User **Documentation**

Next steps

- Ontologies ⇒ **Automatic and Robust** validation&recognition (reviewer friendly)
- Complete workflow remains a **manual** task ⇒ guaranty reproducibility
- **Guide** researchers for CO₂ consideration of data storage costs
- **Plugins** to become multi-disciplinary

Conclusion

Where we are

- **Data curation** with loaders/viewers (mostly for continuum solidmechanics)
- **Flexible** (remote) data storage
- Publication and archive on **Zenodo/Renku**
- **JTCAM curation policy** enforced with Solidipes@DCSM already
⇒ Brings good principles to this Diamond open access initiative
- User **Documentation**

Next steps

- Ontologies ⇒ **Automatic and Robust** validation&recognition (reviewer friendly)
- Complete workflow remains a **manual** task ⇒ guaranty reproducibility
- **Guide** researchers for CO₂ consideration of data storage costs
- **Plugins** to become multi-disciplinary

Community effort ?