On the necessity of curation of datasets to achieve *FAIR* standard goals in scientific publications

Guillaume Anciaux
LSMS, IIC, ENAC, EPFL

Graphic from PHD Comics

Introduction

Researcher-Reader Needs

- Comprehensive
- Trustworthy
- Searchable
- Reproducible
- Easy access

Introduction

Researcher-Reader Needs

- Comprehensive
- Trustworthy
- Searchable
- Reproducible
- Easy access

Researcher-Author Needs

- Visible
- Supports Carreer
- Good/Fair evaluations
- Technical support

Introduction

Researcher-Reader Needs

- Comprehensive
- Trustworthy
- Searchable
- Reproducible
- Easy access

Researcher-Author Needs

- Visible
- Supports Carreer
- Good/Fair evaluations
- Technical support

In all cases **O**pen **A**ccess is a plus...

History

First scientific press:

- 1450: Printing Press (in europe)
- 1534: Foundation of Cambridge University Press
- 1665: *Journal des Sçavans* (France), *Philosophical Transactions of the Royal Society* (UK)

History

First scientific press:

- 1450: Printing Press (in europe)
- 1534: Foundation of Cambridge University Press
- 1665: *Journal des Sçavans* (France), *Philosophical Transactions of the Royal Society* (UK)

Defined the purpose of scientific journals:

- registration: authorship/priority claim
- certification: usually peer-review
- dissemination: provide (targeted) access
- archiving: permanent access link (citable)

Open Access Models

Credits to oabooks-toolkit

- Gold
  - Immediate open access publication
  - Managed by the publisher (**APCs**)
  - **licence allowing reuse (e.g. Creative Commons)**

Open Access Models

Credits to oabooks-toolkit

- Gold
  - Immediate open access publication
  - Managed by the publisher (**APCs**)
  - **licence allowing reuse (e.g. Creative Commons)**
- Green (self-archiving)
  - Publication archived online in an **Open-access repository** (arXiv, HAL, infoscience).
  - **No publisher work** (copy-editing, proofreading, typesetting, indexing, metadata tagging, marketing or distribution).
  - Not listed by publishers (**no metrics**)

Open Access Models

Credits to oabooks-toolkit

- Gold
  - Immediate open access publication
  - Managed by the publisher (**APCs**)
  - **licence allowing reuse (e.g. Creative Commons)**
- Green (self-archiving)
  - Publication archived online in an **Open-access repository** (arXiv, HAL, infoscience).
  - **No publisher work** (copy-editing, proofreading, typesetting, indexing, metadata tagging, marketing or distribution).
  - Not listed by publishers (**no metrics**)
- **Diamond**

  Diamond open access *refers to academic texts (such as monographs, edited collections, and journal articles) published/distributed/preserved with* ***no fees to either reader or author.***

Open Access Models

Credits to oabooks-toolkit

- Gold
  - Immediate open access publication
  - Managed by the publisher (**APCs**)
  - **licence allowing reuse (e.g. Creative Commons)**
- Green (self-archiving)
  - Publication archived online in an **Open-access repository** (arXiv, HAL, infoscience).
  - **No publisher work** (copy-editing, proofreading, typesetting, indexing, metadata tagging, marketing or distribution).
  - Not listed by publishers (**no metrics**)
- **Diamond**

  Diamond open access *refers to academic texts (such as monographs, edited collections, and journal articles) published/distributed/preserved with* ***no fees to either reader or author.***
- **Overlay journal**

  An ***open access*** *academic* ***overlay journal*** *does not produce its own content, but selects from texts that are* ***already freely available online****.*

Diamond Open Access, Overlay journal

V. Acary
INRIA Grenoble
France

M. Legrand
McGill Univ
Canada

G. Anciaux
EPFL
Suisse






M. Montagnat
IGE Grenoble, CNRS
France

F. Gibier
U Montpellier
France

V. Yastrebov
MINES Paris, CNRS
France

JTCAM
Journal of Theoretical, Computational and Applied Mechanics

- **Overlay Journal**
  - Always a preprint shared on Open Archives (even for refused papers)
  - Diamond Open Access
  - FAIR open access (**F**indable, **A**ccessible, **I**nteroperable, **R**eusable)

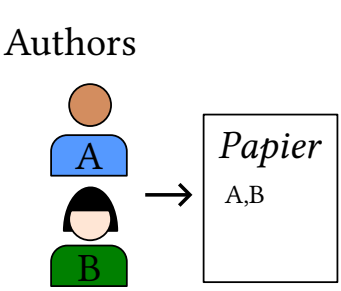
- **Overlay Journal**
  - Always a preprint shared on Open Archives (even for refused papers)
  - Diamond Open Access
  - FAIR open access (**F**indable, **A**ccessible, **I**nteroperable, **R**eusable)
- **Team**
  - Technical board: creators of the journal + data/software editor
  - Scientific Board: invited
  - Editorial board: elected
  - Collegial decisions

- **Overlay Journal**
  - Always a preprint shared on Open Archives (even for refused papers)
  - Diamond Open Access
  - FAIR open access (**F**indable, **A**ccessible, **I**nteroperable, **R**eusable)
- **Team**
  - Technical board: creators of the journal + data/software editor
  - Scientific Board: invited
  - Editorial board: elected
  - Collegial decisions
- **Peer Reviewed**
  - Publish reviewers' work as Open Reviews

- **Overlay Journal**
  - Always a preprint shared on Open Archives (even for refused papers)
  - Diamond Open Access
  - FAIR open access (**F**indable, **A**ccessible, **I**nteroperable, **R**eusable)
- **Team**
  - Technical board: creators of the journal + data/software editor
  - Scientific Board: invited
  - Editorial board: elected
  - Collegial decisions
- **Peer Reviewed**
  - Publish reviewers' work as Open Reviews
- **Data&Software Review/Curation**
  - Strong incentive for **reproducibility** (ongoing)

- **Overlay Journal**
  - Always a preprint shared on Open Archives (even for refused papers)
  - Diamond Open Access
  - FAIR open access (**F**indable, **A**ccessible, **I**nteroperable, **R**eusable)
- **Team**
  - Technical board: creators of the journal + data/software editor
  - Scientific Board: invited
  - Editorial board: elected
  - Collegial decisions
- **Peer Reviewed**
  - Publish reviewers' work as Open Reviews
- **Data&Software Review/Curation**
  - Strong incentive for **reproducibility** (ongoing)
- **Copy-editing**
  - Very high quality

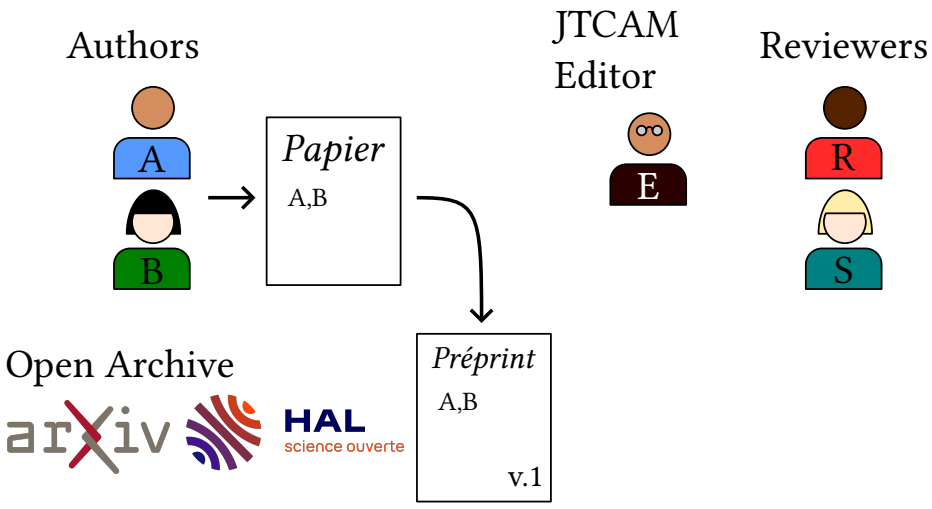
JTCAM: publication process



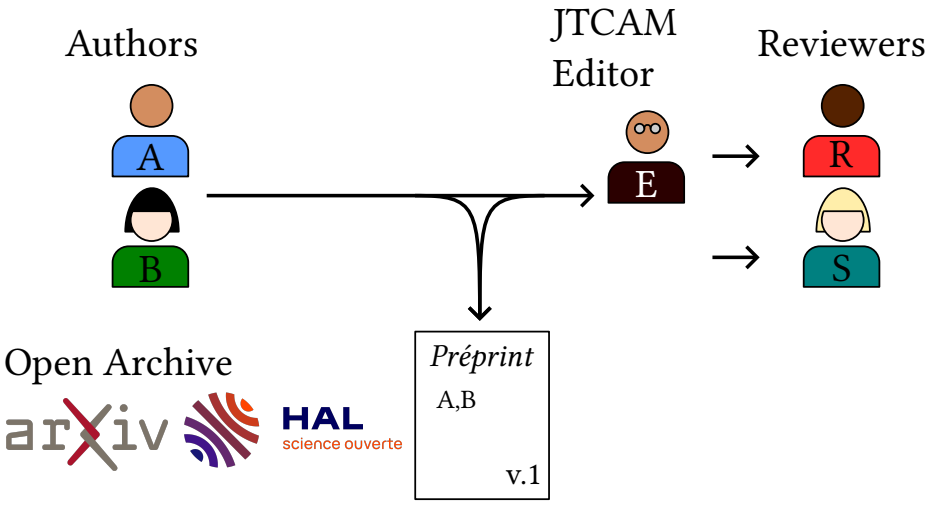
Open Archive

arXiv HAL science ouverte

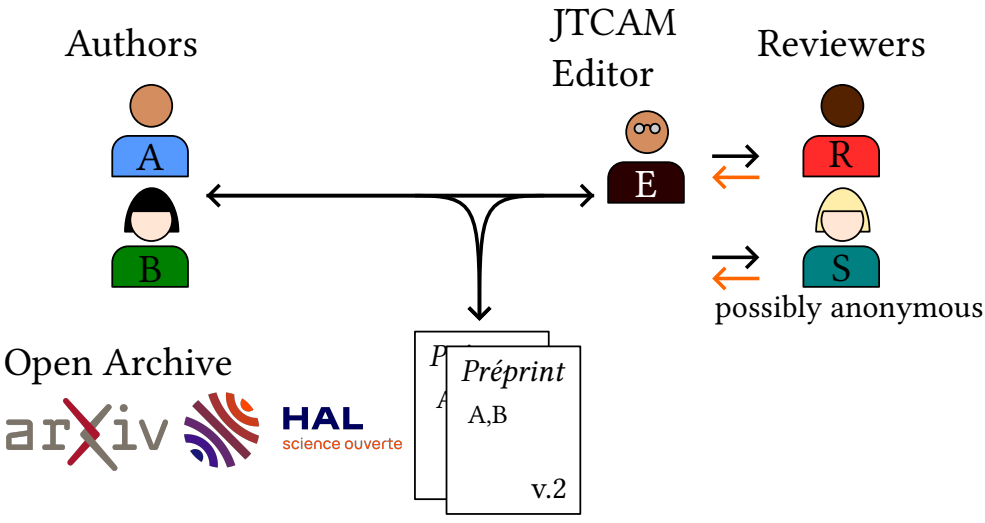
JTCAM: publication process



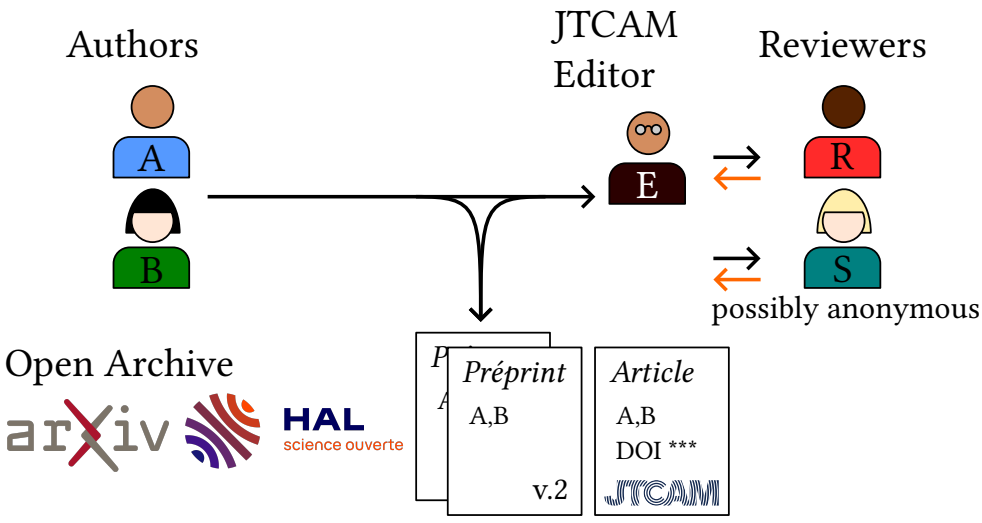
JTCAM: publication process



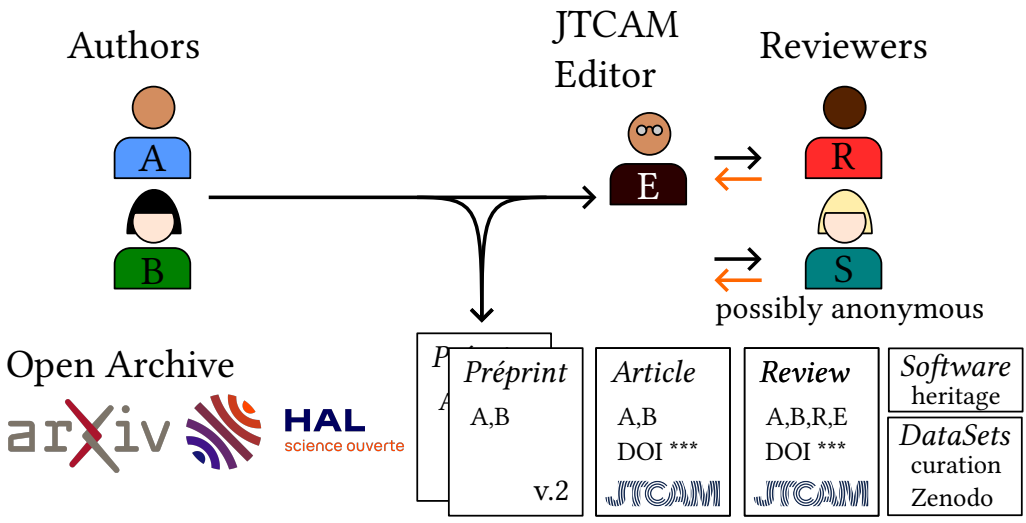
JTCAM: publication process



JTCAM: publication process



JTCAM: publication process



JTCAM FAIR principles

Findable by Journal indexation

- *Directory of Open Access Journals* (DOAJ), *Free Journal Network* (FJN), *International Standard Serial Number International Center* (ISSN), Mir@bel

Accessible

- OpenSource Episcience CMS (funded by French CCSD through CNRS, INRIA, INRAE, OpenAIRE, FNSO)
- Overlay Journal: articles stored in open repositories (arXiv, HAL)
- Curated/Reviewed Datasets with DOI @Zenodo (Curation with ETH-ORD)
- CC-BY license

Interoperable

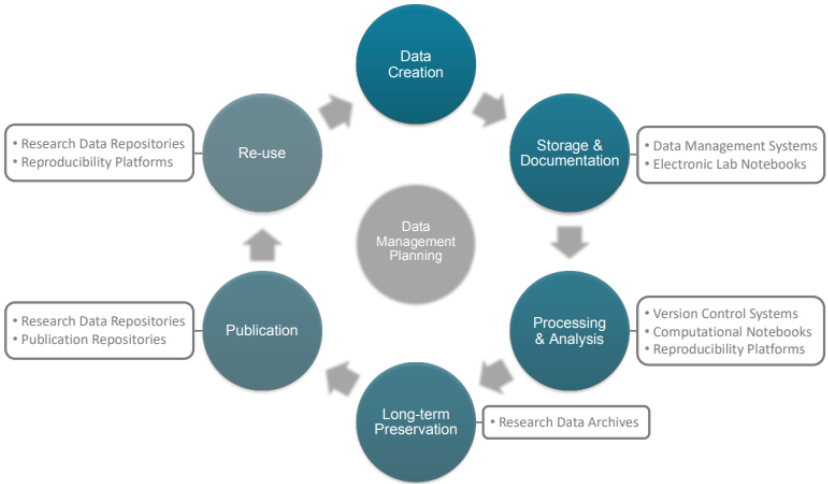
- Provided by the repositories with metadata

Reusable

- Saving Software revision @Software Heritage (SWHID ~ DOI for software) complement datasets

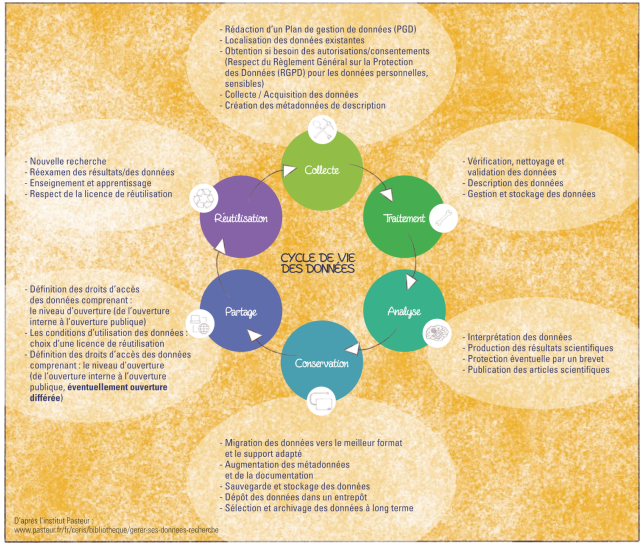
Open Data

Data cycle


Data Creation
Storage & Documentation
• Data Management Systems
• Electronic Lab Notebooks
Processing & Analysis
• Version Control Systems
• Computational Notebooks
• Reproducibility Platforms
Long-term Preservation
• Research Data Archives
Publication
• Research Data Repositories
• Publication Repositories
Re-use
• Research Data Repositories
• Reproducibility Platforms
Data Management Planning

On fait le point sur les données...

https://zenodo.org/records/10108736



What makes (open) datasets useful: FAIR

Findable

- Need for searchable databases
- Annotation (keywords, cross-links, ownership, ...)
- *Digital Object Identifier* (DOI)

Accessible

- Retention times
- Choice of data repository
- Question of storage costs

Interoperable

- Compatibility issues (beetween repositories, packages, file formats, ...)

Reusable

- Operating context (Software versions, dependencies, ...)
- Open (source) licences

Need for annotation

JTCAM dataset curation policy

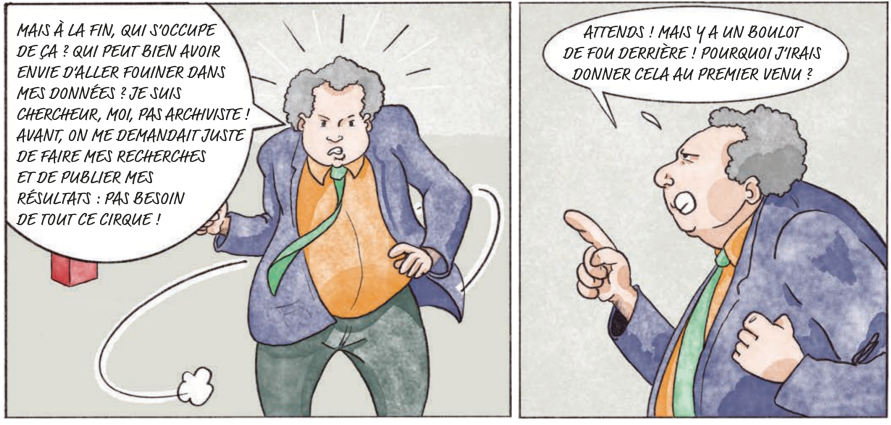
The following criteria are required in order to accept a submission to the JTCAM community:

- 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 redundency)
- Permissive licenses are required (CC0, CC-BY-4.0)
- Files formats are open
- Workflow description

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

Difficulty

Convincing researchers



DCSM Project

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

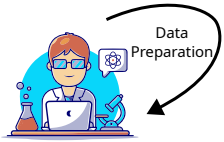
- Fund by Open Research Data (ORD)
- G. Anciaux (dev and supervision@EPFL), S. Pham-Ba (developer@EPFL)

Goals

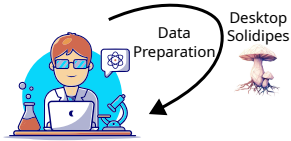
- Provide a **cloud based** repository/storage/tool for **solidmechanics** community
- Simplify the **verification, analysis and annotation** (curation) of datasets
- Stand-alone tool for researchers to manipulate data **on their personal computer**
- Web tool to do the same, online
  $\Rightarrow$ Prototype Web out **end of January 2024**
- Used at JTCAM for **data reviews**

Solidipes: analysis and curation tool

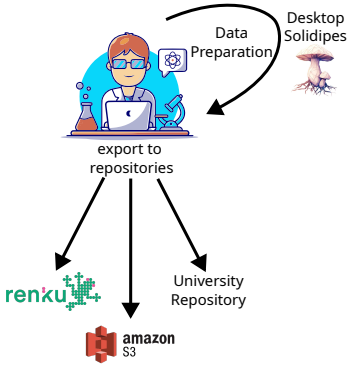
Solidipes: analysis and curation tool



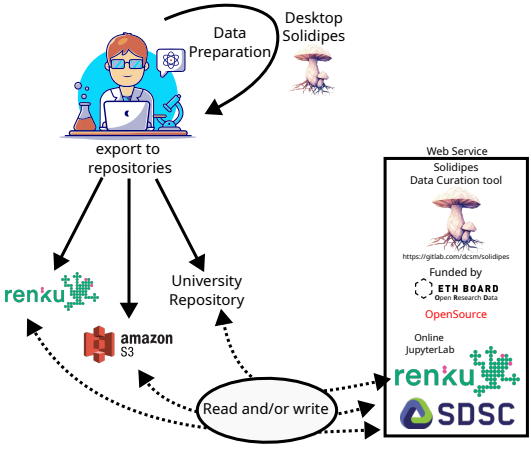
Solidipes: analysis and curation tool



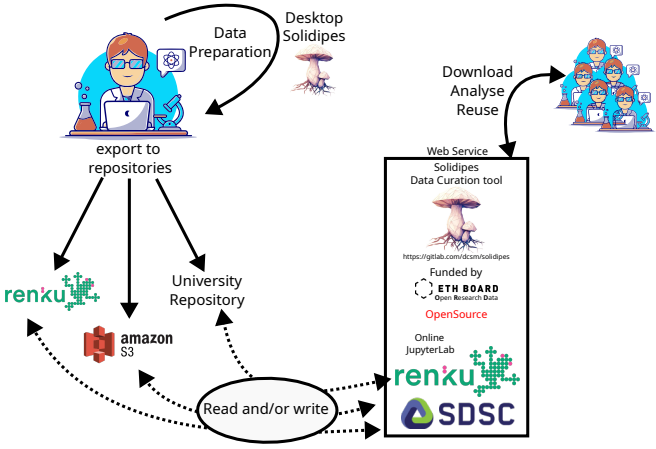
Solidipes: analysis and curation tool



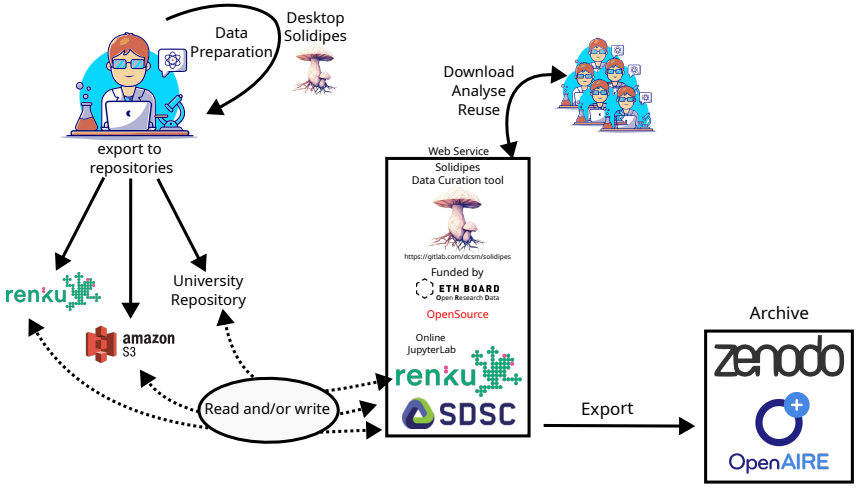
Solidipes: analysis and curation tool



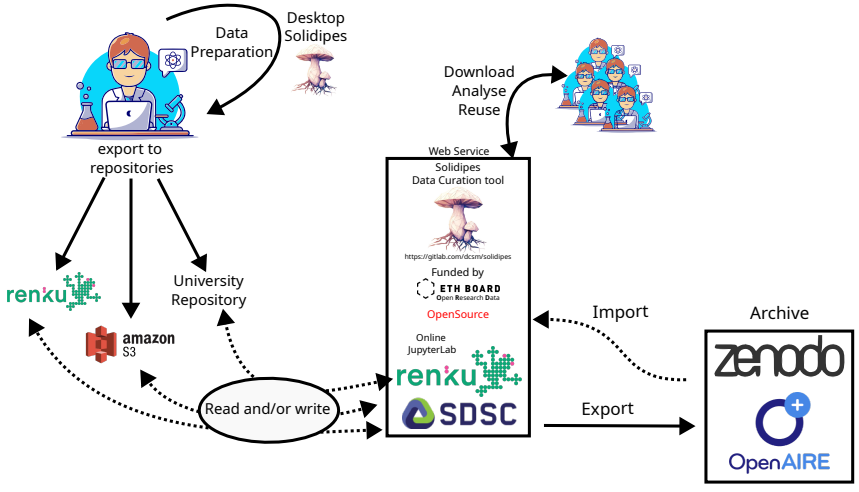
Solidipes: analysis and curation tool



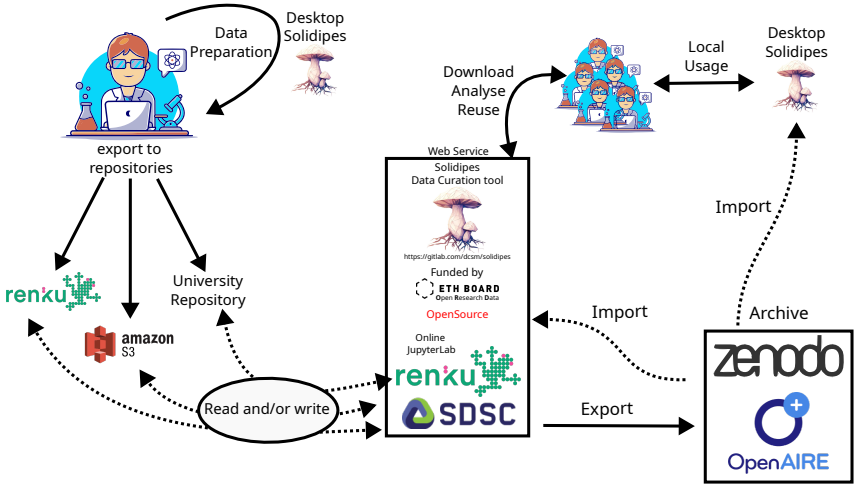
Solidipes: analysis and curation tool



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/dcsm/solidipes

Funded by

ETH BOARD
Open Research Data

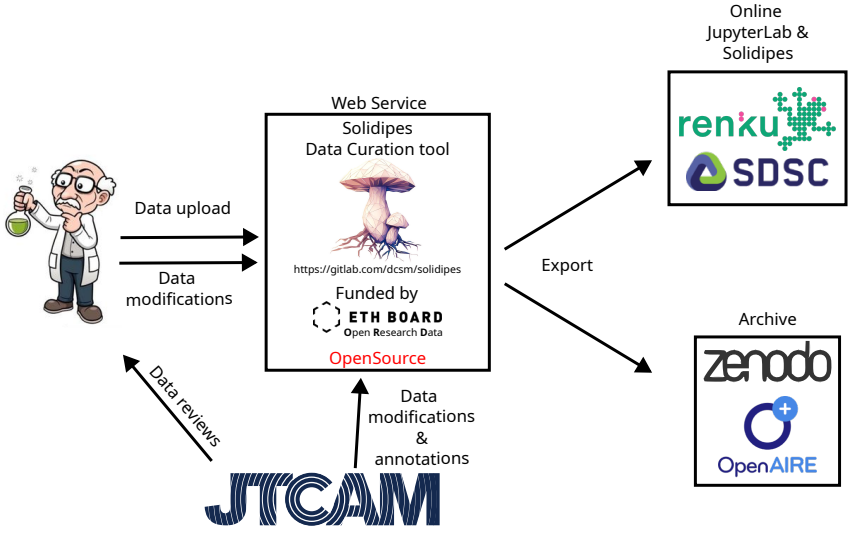
OpenSource

Online
JupyterLab

renku

SDSC

Dataset Curation Management@JTCAM



Conclusion

Where we are

- Data curation tools for solidmechanics on its way

Conclusion

Where we are

- Data curation tools for solidmechanics on its way
- Flexible data storage

Conclusion

Where we are

- Data curation tools for solidmechanics on its way
- Flexible data storage
- Archive on Zenodo (follows EPFL policy)

Conclusion

Where we are

- Data curation tools for solidmechanics on its way
- Flexible data storage
- Archive on Zenodo (follows EPFL policy)
- JTCAM curation policy enforced with Solidipes@DCSM already

Conclusion

Where we are

- Data curation tools for solidmechanics on its way
- Flexible data storage
- Archive on Zenodo (follows EPFL policy)
- JTCAM curation policy enforced with Solidipes@DCSM already
  $\Rightarrow$ Brings good principles to this Diamond open access initiative
- User Documentation provided

Conclusion

Where we are

- Data curation tools for solidmechanics on its way
- Flexible data storage
- Archive on Zenodo (follows EPFL policy)
- JTCAM curation policy enforced with Solidipes@DCSM already
  $\Rightarrow$ Brings good principles to this Diamond open access initiative
- User Documentation provided

Next steps

- Finilize the open prototype (private access at the moment)
- Complete workflow (program runs) is still manual and difficult
  $\Rightarrow$ Necessary to guaranty reproducibility
- Guide the researchers for when and what quantity deserves to be saved for long times
  $\Rightarrow$ CO2 emissions?