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

Role of scientific journals

*What is the **Role** played by **Journals** and **Publishers** ?*

- **Registration**: authorship/priority claim
- **Certification**: usually peer-review
- **Dissemination**: provide (targeted) access
- **Archiving**: permanent access link (citable)

*What is **FAIR** data ?*

- **Findable**: permanent links
- **Accessible**: online services
- **Interoperable**: standard formats, access to data
- **Reusable**: Applicable to journals ?

(short) History of academic press

You have to know the past to understand the present.

Carl Sagan, Astronomer

You can't know where you are going until you know where you have been.

Maya Angelou, Poetress

(short) History of academic press

You have to know the past to understand the present.

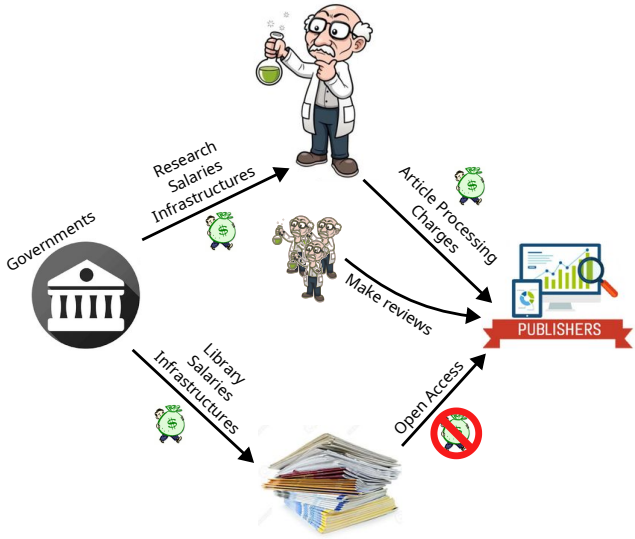
Carl Sagan, Astronomer

You can't know where you are going until you know where you have been.

Maya Angelou, Poetress

- S. Buranyi, Is the staggeringly profitable business of scientific publishing bad for science? The Guardian (2017).
- Against Parasite Publishers: Making Journals Free (2022)

Publishing system: problematic?



Cost of APCs?

Grossmann, A. & Brembs, B. Current market rates for scholarly publishing services. (2021)

> *[...] conservative estimates show that the publication cost for a representative scholarly article is around $400.*

Cost of APCs?

Grossmann, A. & Brembs, B. Current market rates for scholarly publishing services. (2021)

> *[...] conservative estimates show that the publication cost for a representative scholarly article is around $400.*

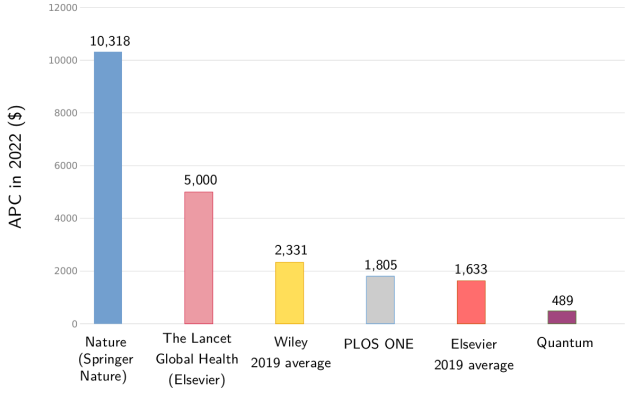
Yet APCs scale with impact factor

Cost of APCs?

Grossmann, A. & Brembs, B. Current market rates for scholarly publishing services. (2021)

> *[...] conservative estimates show that the publication cost for a representative scholarly article is around $400.*

Yet APCs scale with impact factor



Cost of APCs?

***Nature*, doi:10.1038/d41586-023-01391-5**

ISTOCK/GETTY

Neuroimaging research is at the centre of a row about open-access publishing fees.

EDITORS QUIT BRAIN RESEARCH JOURNALS TO PROTEST AGAINST FEES

They say the charges to publish articles open access are unsustainable.

of $6,300; the fee at *Nature Neuroscience*, published by Springer Nature, is $11,690; and *Human Brain Mapping*, published by Wiley, charges $3,850. (*Nature*'s news team is editorially independent of *Nature Neuroscience* and of Springer Nature.)

The *NeuroImage* editors say that the fees exclude many scholars who are based in countries where research is not well funded. They think that the charges don't reflect direct article costs, and say it is wrong for publishers to make profit from science that they haven't funded.

Elsevier says that it is committed to advancing open access to research and has schemes to support researchers in poorer countries. Davis says Elsevier helps researchers in 120 low- and middle-income countries to receive affordable access to nearly 100,400 peer-reviewed resources, through a public–private partnership called Research4Life. He adds that Elsevier automatically applies waivers or discounts on fees to publish articles in fully open-access journals when all authors are in a low-income country.

Open-access transition

NeuroImage launched in 1992, and became open access in 2020 with an APC of $3,000, which has been raised twice. *NeuroImage: Reports* launched in 2021 to publish results, including null findings, and methods. In June last year, the editors, led by Smith, asked Elsevier to lower *NeuroImage*'s APC to less

Elsevier agreement

Negociations with Elsevier – update 14th june 2024

- illimited access
- illimited open publications
- all journals
- All possible AI usage
- cost ?

The **Diamond Open Access** alternative

is a ***Diamond open access*** *journal, i.e. published with* ***no fees to either reader or author.***

and is an ***overlay*** *journal, i.e. that does not produce its own content, but selects from texts that are* ***already freely available online*** *(thanks to Episciences!).*

is a ***Diamond open access*** *journal, i.e. published with* ***no fees to either reader or author.***

and is an ***overlay*** *journal, i.e. that does not produce its own content, but selects from texts that are* ***already freely available online*** *(thanks to Episciences!).*

- Editorial process entirely controlled by researchers (Reviews, Copy-editing)
- Supported by *Centre for Direct Scientific Communication* (CCSD@CNRS/INRAE/Inria)
- Wide spectrum: theoretical, applied, numerical, experimental

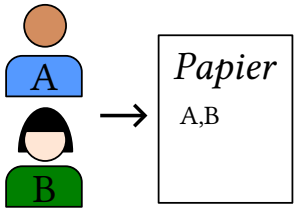
is a ***Diamond open access*** *journal, i.e. published with* ***no fees to either reader or author.***

and is an ***overlay*** *journal, i.e. that does not produce its own content, but selects from texts that are* ***already freely available online*** *(thanks to Episciences!).*

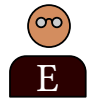
- Editorial process entirely controlled by researchers (Reviews, Copy-editing)
- Supported by *Centre for Direct Scientific Communication* (CCSD@CNRS/INRAE/Inria)
- Wide spectrum: theoretical, applied, numerical, experimental
- Publish **Open Reviews**
- Include **Dataset curation** (*ETH-Board* support)

JTCAM: publication process

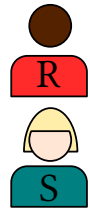
Authors



JTCAM Editor



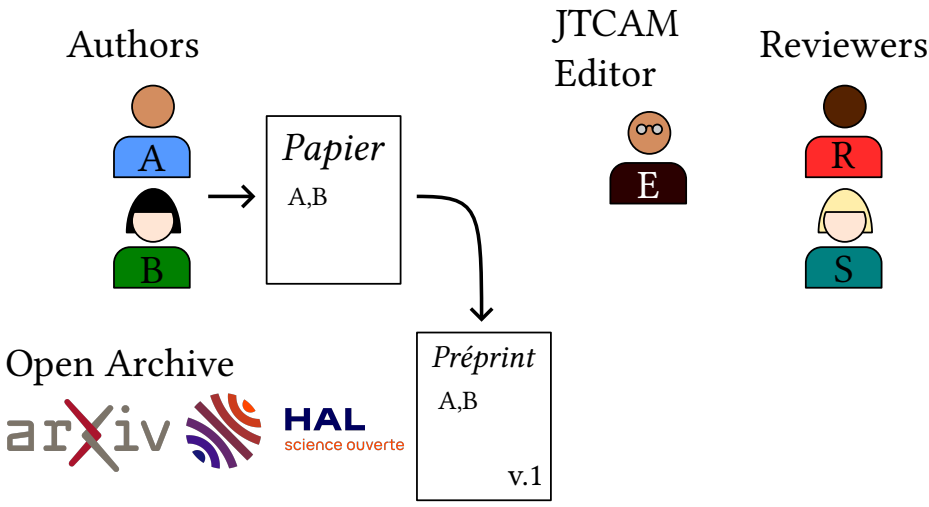
Reviewers



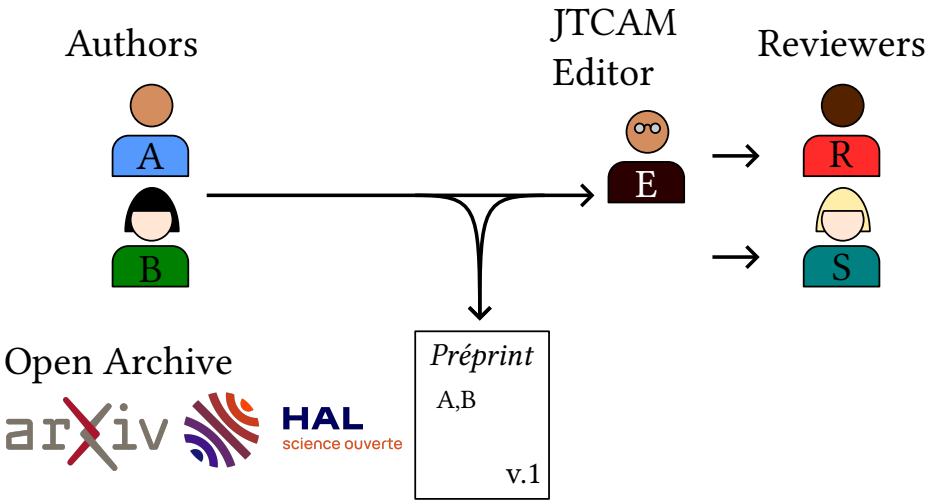
Open Archive

arXiv HAL science ouverte

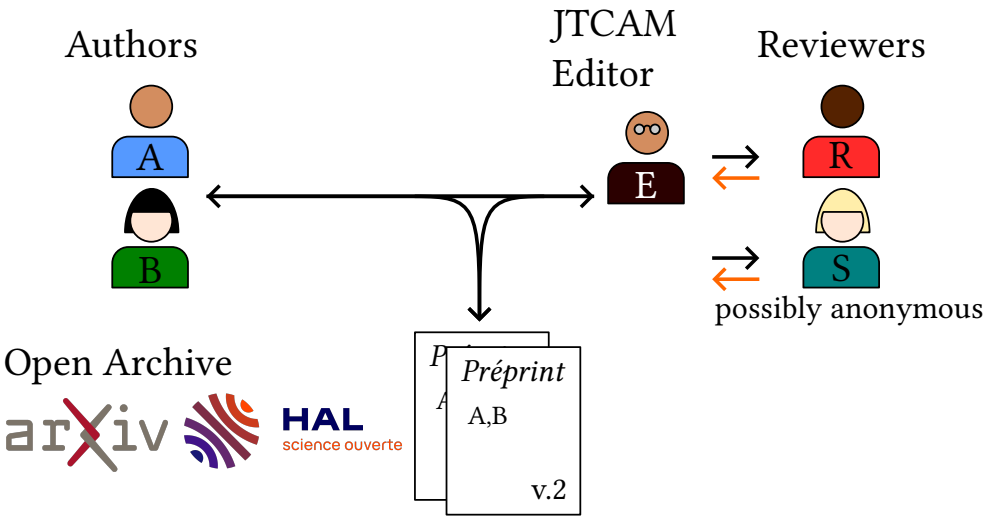
JTCAM: publication process



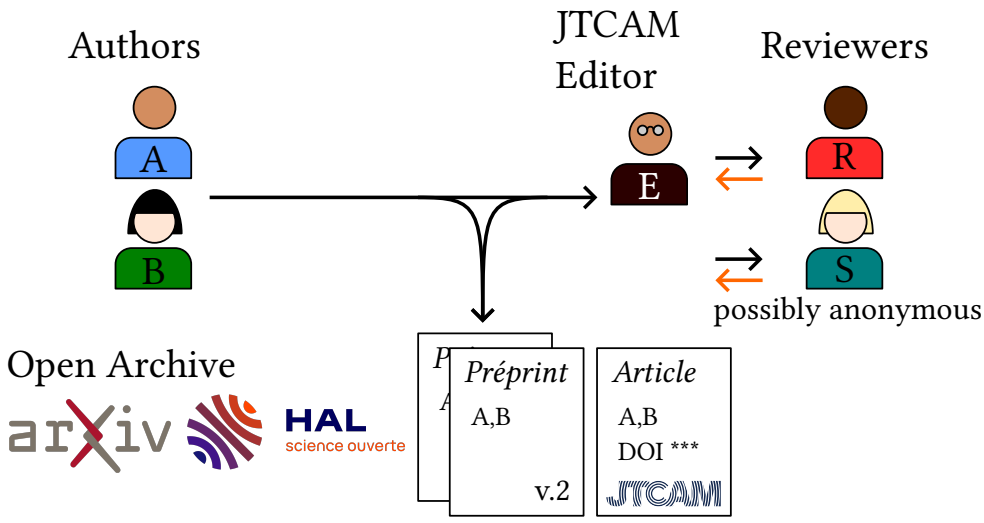
JTCAM: publication process



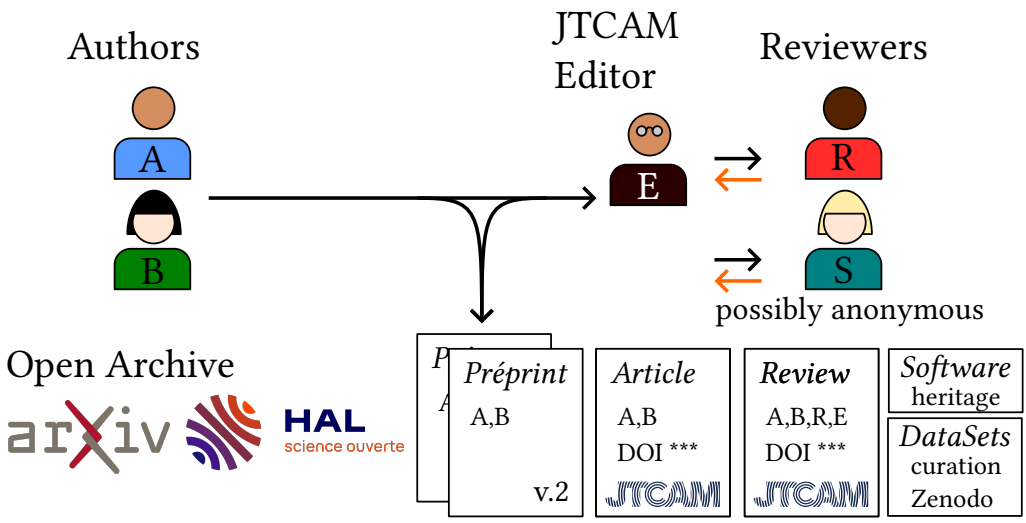
JTCAM: publication process



JTCAM: publication process



JTCAM: publication process



What makes (open) datasets useful: FAIR

Findable

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

Accessible

- Retention times?
- Data repository?
- Storage costs?

Interoperable

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

Reusable

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

Journals policies with datasets

e.g. the Springer research data policy

Classification of journals

Type 1	Data sharing and data citation is encouraged
Type 2	Data sharing and evidence of data sharing encouraged
Type 3	Data sharing encouraged and statements of data availability required
Type 4	Data sharing, evidence of data sharing and peer review of data required

Journals policies with datasets

e.g. the Springer research data policy

Classification of journals

Type 1	Data sharing and data citation is encouraged
Type 2	Data sharing and evidence of data sharing encouraged
Type 3	Data sharing encouraged and statements of data availability required
Type 4	Data sharing, evidence of data sharing and peer review of data required

$\Rightarrow$ **(Computational) mechanics should be of type 4**

Journals policies with datasets

e.g. the Springer research data policy

Classification of journals

Type 1	Data sharing and data citation is encouraged
Type 2	Data sharing and evidence of data sharing encouraged
Type 3	Data sharing encouraged and statements of data availability required
Type 4	Data sharing, evidence of data sharing and peer review of data required

⇒ **(Computational) mechanics should be of type 4**

- Springer journals created to review datasets (Scientific Data, BMC series, Discover)
- The Journal of Open Source Software (JOSS), Open Source Initiative

Journals policies with datasets

e.g. the Springer research data policy

Classification of journals

Type 1	Data sharing and data citation is encouraged
Type 2	Data sharing and evidence of data sharing encouraged
Type 3	Data sharing encouraged and statements of data availability required
Type 4	Data sharing, evidence of data sharing and peer review of data required

⇒ **(Computational) mechanics should be of type 4**

- Springer journals created to review datasets (Scientific Data, BMC series, Discover)
- The Journal of Open Source Software (JOSS), Open Source Initiative
- and now JTCAM ⇒ ambitious goals

Journals policies with datasets

e.g. the Springer research data policy

Classification of journals

Type 1	Data sharing and data citation is encouraged
Type 2	Data sharing and evidence of data sharing encouraged
Type 3	Data sharing encouraged and statements of data availability required
Type 4	Data sharing, evidence of data sharing and peer review of data required

⇒ **(Computational) mechanics should be of type 4**

- Springer journals created to review datasets (Scientific Data, BMC series, Discover)
- The Journal of Open Source Software (JOSS), Open Source Initiative
- and now JTCAM ⇒ ambitious goals

Makes sense to gather the review of paper and datasets

Journals policies with datasets

e.g. the Springer research data policy

Classification of journals

Type 1	Data sharing and data citation is encouraged
Type 2	Data sharing and evidence of data sharing encouraged
Type 3	Data sharing encouraged and statements of data availability required
Type 4	Data sharing, evidence of data sharing and peer review of data required

⇒ **(Computational) mechanics should be of type 4**

- Springer journals created to review datasets (Scientific Data, BMC series, Discover)
- The Journal of Open Source Software (JOSS), Open Source Initiative
- and now JTCAM ⇒ ambitious goals

Makes sense to gather the review of paper and datasets

My personal belief is that journals should play a role

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

Ideal curation

- **Versatile** data storage
- **Various** disciplines (experimental, theoretical, numerical, fluid mechanics, solid mechanics, ...)
- Collaborative curation (**concurrent editing**)
- Robust descriptions (**ontologies**)
- Reproducibility (**workflows**)

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)
- Young project (18 months)
- Use lots of project dependencies

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 service: `https://dcsm.epfl.ch`
- Used at JTCAM for **data reviews**
- *"Overleaf"* for datasets

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 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 graphical scrutiny (images, interactive 3D rendering, ...)
5. If validated: enables export to Zenodo/Renku

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 (*a platform and tools for reproducible and collaborative data analysis*)
- Curation session

Web Service

Solidipes
Data Curation tool

https://gitlab.com/dcsm/solidipes

Funded by

ETH BOARD
Open Research Data

OpenSource

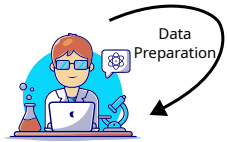
Online
JupyterLab

renku

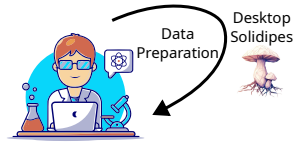
SDSC

Solidipes: analysis and curation tool

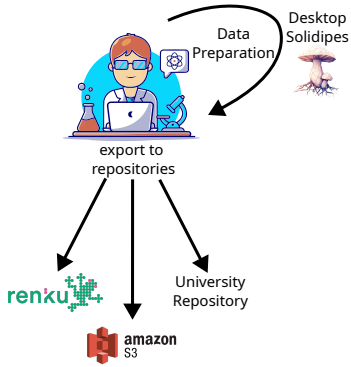
Solidipes: analysis and curation tool



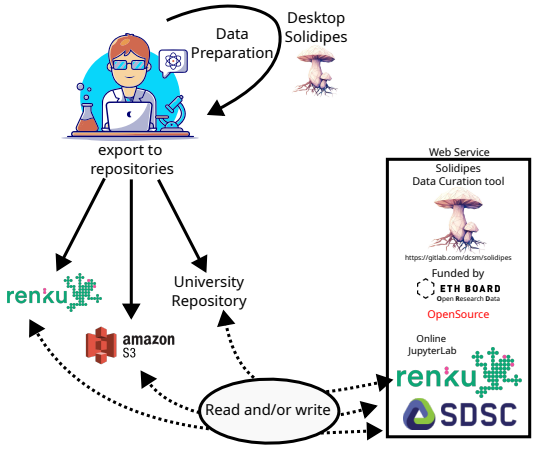
Solidipes: analysis and curation tool



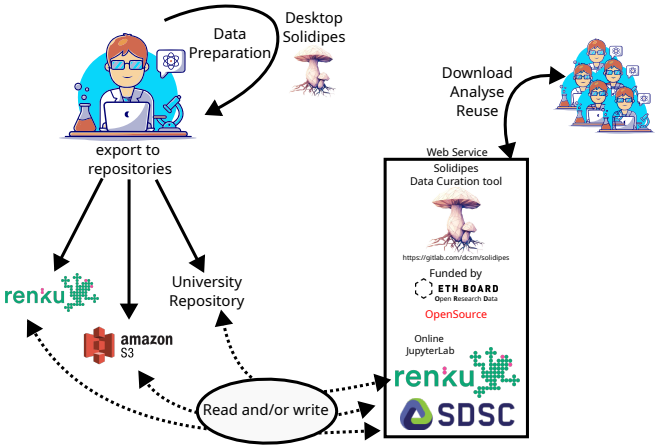
Solidipes: analysis and curation tool



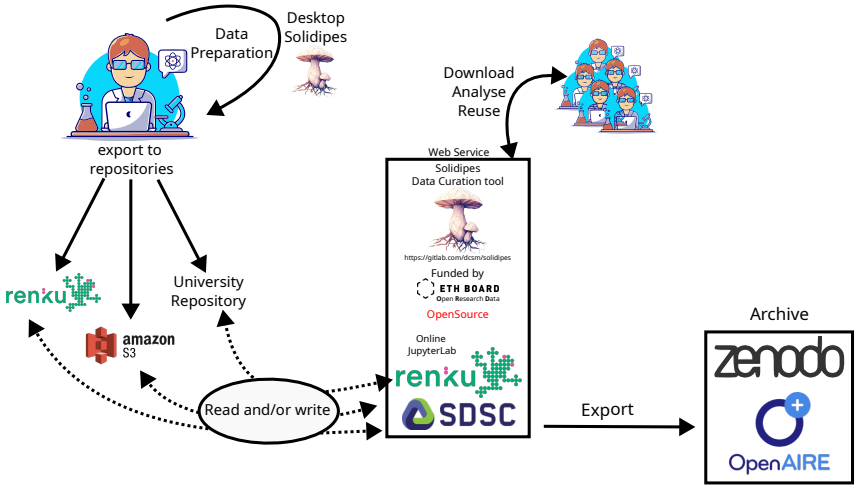
Solidipes: analysis and curation tool



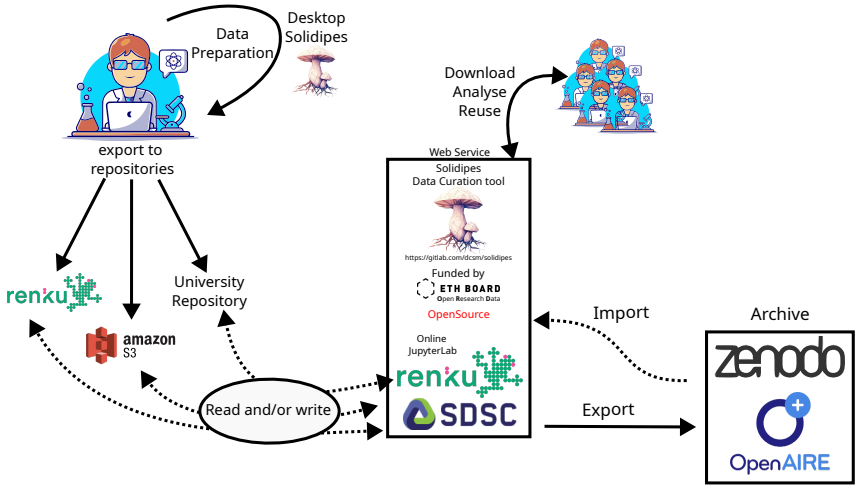
Solidipes: analysis and curation tool



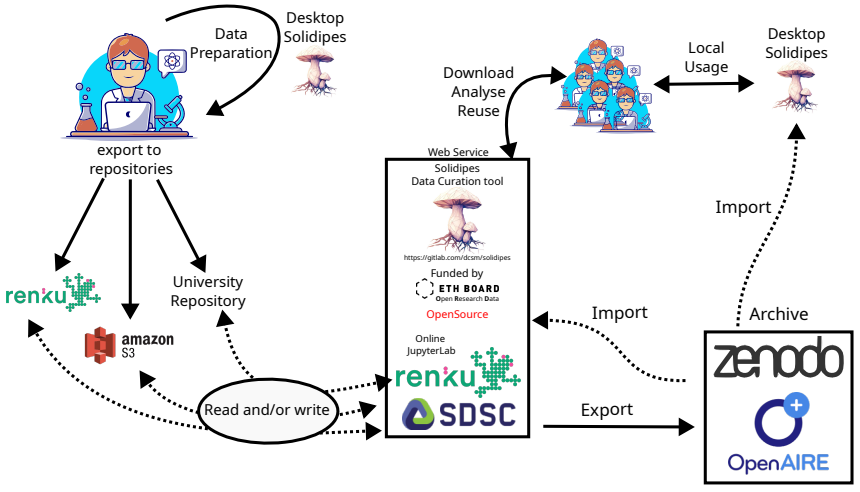
Solidipes: analysis and curation tool



Solidipes: analysis and curation tool



Solidipes: analysis and curation tool



Solidipes project

Solidipes: an interoperable tool for curation of research data

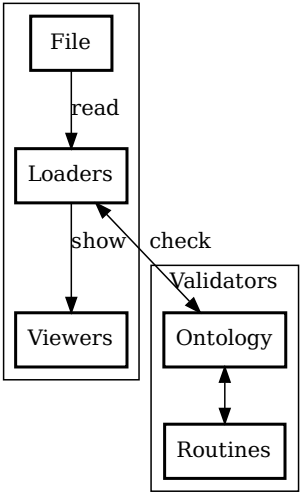
- Fund by Open Research Data (ORD)
- G. Anciaux (LSMS), S. Pham-Ba (ENAC-IT4R), A. Borel (EPFL Library), V. Savchenko and A. Neronov (LASTRO)
- 18 more months

Goals

- **Generalizing** Solidipes for virtually all fields
- **Extending** to view workflows
- **Integrating** with other storage repositories (import/export)
- **Integrating** with new services (Galaxy, MMODA)
- **Distribute** plugins contributed by distinct scientific communities.

How to create a scientific field

Solidipes components



Loader&Viewers

- image, notebook, text, pdf, code_snippet, video, hdf5, python_pickle, table, xml
- meshio, abaqus

Validations

- Mime type **matches** extension
- CSV **have** headers
- General: data **fall** into a category

Ontologies

Ontology (adapted from Wikipedia)

*an ontology encompasses **definitions** of the **categories**, **properties**, and **relations** between the **data entities** of a **(scientific) topic**.*

Ontologies

Ontology (adapted from Wikipedia)

> *an ontology encompasses* ***definitions*** *of the* ***categories***, ***properties***, *and* ***relations*** *between the* ***data entities*** *of a* ***(scientific) topic***.

In practice it is an **annotated graph** described with the *Resource Description Framework* (RDF)

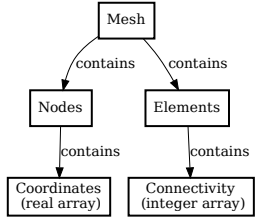
Ontologies

Ontology (adapted from Wikipedia)

*an ontology encompasses **definitions** of the **categories**, **properties**, and **relations** between the **data entities** of a **(scientific) topic**.*

In practice it is an **annotated graph** described with the *Resource Description Framework* (RDF)

e.g. a **valid** mesh file **must** contain nodes and elements.



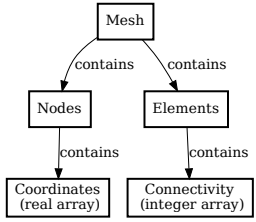
Ontologies

Ontology (adapted from Wikipedia)

> *an ontology encompasses* ***definitions*** *of the* ***categories***, ***properties***, *and* ***relations*** *between the* ***data entities*** *of a* ***(scientific) topic***.

In practice it is an **annotated graph** described with the *Resource Description Framework* (RDF)

e.g. a **valid** mesh file **must** contain nodes and elements.



- XDMF is a XML file (detected by linux) containing *nodes and elements* tags
- **.inp** is the **Abaqus** input file format, which may, or may not include meshing information

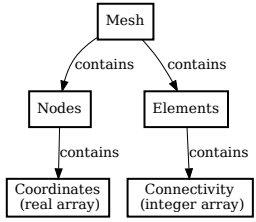
Ontologies

Ontology (adapted from Wikipedia)

> *an ontology encompasses* ***definitions*** *of the* ***categories***, ***properties***, *and* ***relations*** *between the* ***data entities*** *of a* ***(scientific) topic***.

In practice it is an **annotated graph** described with the *Resource Description Framework* (RDF)

e.g. a **valid** mesh file **must** contain nodes and elements.



- XDMF is a XML file (detected by linux) containing *nodes and elements* tags
- **.inp** is the **Abaqus** input file format, which may, or may not include meshing information
- Will allow semi-automatic validation, and content check for specific purpose

Ontologies

ROcrate

- Adopt “ro-crate-metadata.json”.
- ROcrate Metadata
- Dublin Core standard
- ROHub

Ontology Viewers ?

- RDF: (xml or Turtle-ttl)
- RDFLib (python library)
- JSON-LD https://json-ld.org/ (Schema.org)

Ontologies in mechanics

J.-L. Hippolyte, P. Duncan, M. Bevilacqua and M. Chrubasik. *Ontologies for Experimental Mechanics*. British Society for strain measurement, National Physical Laboratory, Hampton Rd, Teddington TW11 0LW, UK. 2022 [link]

Marcin Skulimowski. *An OWL Ontology for Quantum Mechanics*. Faculty of Physics and Applied Informatics, University of Lodz. Pomorska 149/153, 90-236 Lodz, Poland. 2002 [link]

H.A. Preisig, T.F. Hagelien, J.Friis, P. Klein, N. Konchakova. *Ontologies in Computational Engineering*. 14th World Congress on Computational Mechanics (WCCM). 2020. [link]

Ontologies in mechanics

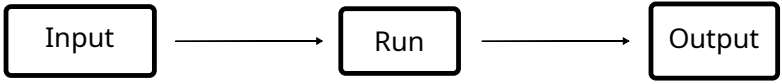
J.-L. Hippolyte, P. Duncan, M. Bevilacqua and M. Chrubasik. *Ontologies for Experimental Mechanics*. British Society for strain measurement, National Physical Laboratory, Hampton Rd, Teddington TW11 0LW, UK. 2022 [link]

Marcin Skulimowski. *An OWL Ontology for Quantum Mechanics*. Faculty of Physics and Applied Informatics, University of Lodz. Pomorska 149/153, 90-236 Lodz, Poland. 2002 [link]

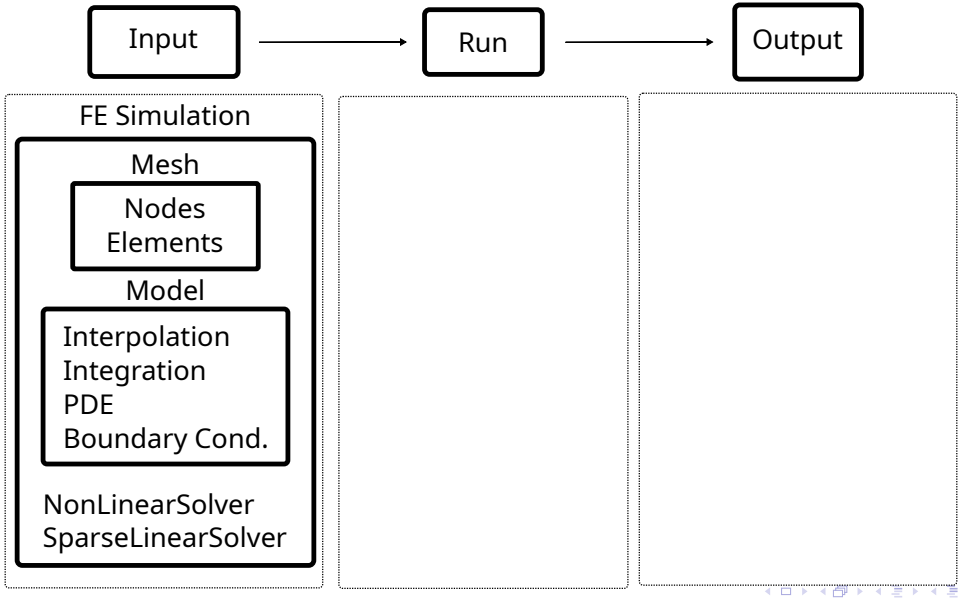
H.A. Preisig, T.F. Hagelien, J.Friis, P. Klein, N. Konchakova. *Ontologies in Computational Engineering*. 14th World Congress on Computational Mechanics (WCCM). 2020. [link]

Need to integrate and mix these initiatives: Forming a committee ?

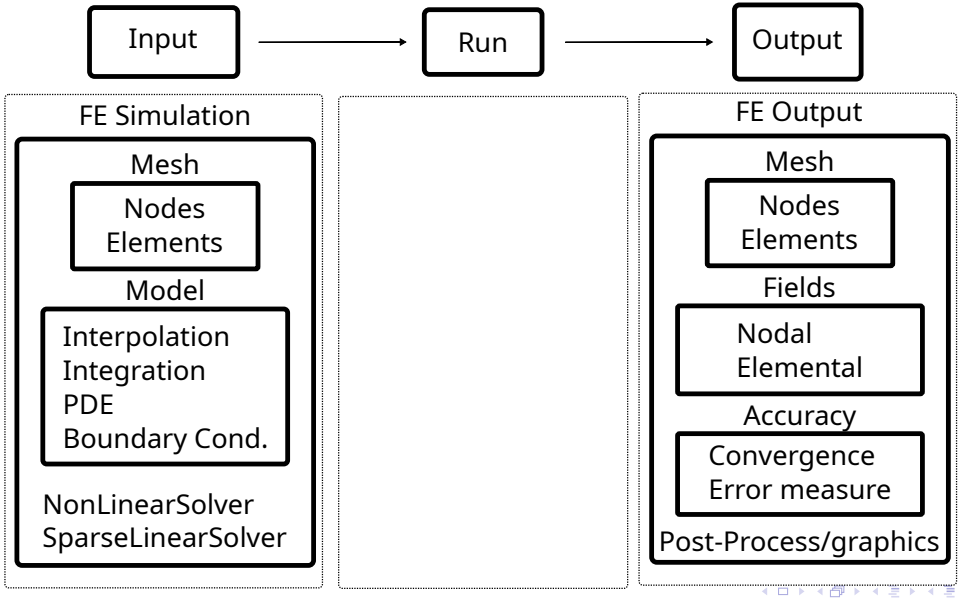
Reproducibility with workflows



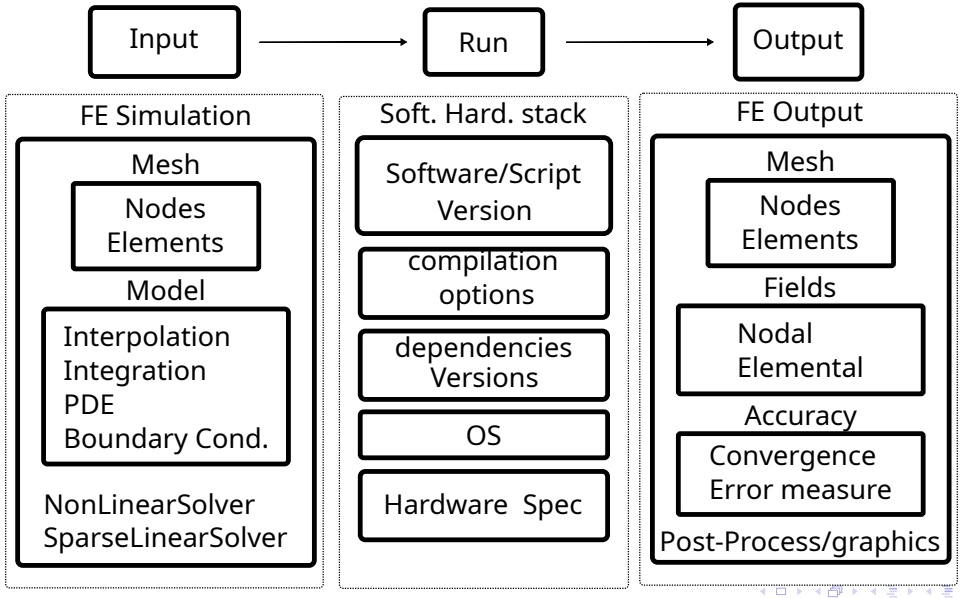
Reproducibility with workflows



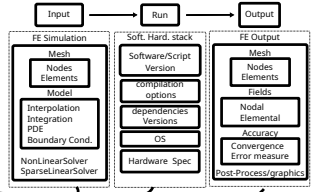
Reproducibility with workflows



Reproducibility with workflows



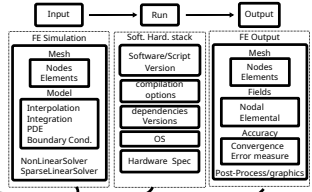
Reproducibility with workflows



Package

- Versions, metadata (yaml, json, ...)
- Results in standard format
- Limit number of files
- Spack/Python lock files
- CMakeCache.txt / build config and log file
- Docker file
- Run logs, execution times, cpu-hours
- How to strictly attach to output results ?

Reproducibility with workflows



Package

- Versions, metadata (yaml, json, ...)
- Results in standard format
- Limit number of files
- Spack/Python lock files
- CMakeCache.txt / build config and log file
- Docker file
- Run logs, execution times, cpu-hours
- How to strictly attach to output results ?

- **Docker** container: Renku, Binder, ...
- **Worflow** management: AiiDA, BlackDynamite
- **Packagers**: ROcrate, reprozip
- **Repository**: WorkflowHub

Acquiring OS context
+ “automatic” DockerFiles?

Conclusion (open access)

Open access problematic

- Public money must turn into public knowledge and goods
- Edition has intrinsic costs, but unreasonable APCs are not acceptable
- Current system leads to explosion of publications and costs

Diamond Open access alternatives

- No fees to authors nor reader => **public good**
- Freedom for researchers: **ethics** and **quality** can be the driving force
- Needs support from institutions (see SwissUniversities and ETH-Board ORD programs
- UNESCO Diamond Open Access Global Alliance is born
- Unclear link with bibliographic metrics
- **JTCAM**: a healthy Diamond open access journal, with innovative dataset handling

Researchers can be game changers, if they want to...

Conclusion (data curation)

Where is Solidipes

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

Conclusion (data curation)

Where is Solidipes

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

Conclusion (data curation)

Where is Solidipes

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

Conclusion (data curation)

Where is Solidipes

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

Conclusion (data curation)

Where is Solidipes

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

Conclusion (data curation)

Where is Solidipes

- **Data curation** with loaders/viewers (mostly for continuum solidmechanics)
- **Flexible** (remote) data storage
- Publication and archive on **Zenodo/Renku + dtool**
- **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
- **Plugins** remain to be contributed (contact us)

You are a Research Software Engineer

- You have a job in research? You write software? → **You are an RSE!**
- We often lack recognition/career progression despite contributions

→ **Join RSE communities !**

- Associations at international, national, instutitional scale
  `https://researchsoftware.org/assoc.html`
- Networking and events (RSE CON 24)
- Acquire skills for managing projects and building careers