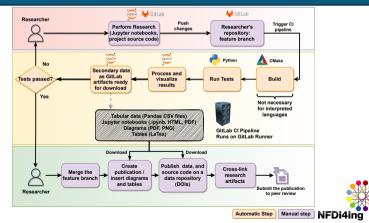
# A Research Software Engineering Workflow for Computational Science and Engineering



#### TECHNISCHE UNIVERSITÄT DARMSTADT

### SFB1194/Z-INF, Base Service S-2 Research Software Development



### Computational Science and Engineering software in university research groups Boundary and initial conditions



TECHNISCHE UNIVERSITÄT DARMSTADT

- Publish or perish <sup>1</sup> prioritizes publications over scientific software.
- Dedicated resources for increasing software quality are usually not available.
- Ph.D. students rotate every 3-5 years, postdocs every 1-2 years.
  - Little or no overlap between successors and predecessors.
- Large-scale software design is not a mandatory part of the CSE curriculum.
  - Different CSE background: (Applied) Mathematics, Mechanical Engineering, Physics, Informatics.

<sup>1</sup>Symbol of a publish-or-perish simplification of the workflow :)

### Computational Science and Engineering software in university research groups The chaos scientific legacy code





engineering research software Betty is a CSE researcher, working with a legacy research code. Why is Betty so (rightfully) angry?

- Betty inherited a research software that is only partially tested.
- Betty inherited a research software that isn't automatically tested.
  - Betty changes one part of the code and gets her model running, only to see 10 other things fail, after days of manually running tests.
- Betty's software has no documentation of the scientific workflow.
  - Betty doesn't know how to use existing scripts to run simulations and analyze (reproduce) results.
- Betty's software has disjoint (diverging) versions that she can't integrate.
- Betty can't even find code versions used to generate results in the publications from her research group.

### Computational Science and Engineering software in university research groups Continuous integration and cross-linking to the rescue



Automated testing (verification and validation), version control, and cross-linking reports, source code and research data increase Findability, Accessibility and Reproducibility (FAIR) and speed up research.

- Continuous Integration (CI) = automatic testing + version control.
- CSE research requires scientific workflows: initialize simulations, run parameter variations, agglomerate data, visualize, and check results.
- Cl can be used to automate and document scientific workflows.
- Cl ensures that the integration of new changes does not break existing functionality.
- Once the changes are integrated, the publication, the source code and the data are published on pre-print and data repositories and cross-linked using git tags and DOIs.

Collecting the components: Automation I



TECHNISCHE UNIVERSITÄT DARMSTADT

while Results are unsatisfactory do Work on algorithms. (Compile the code.) for All studies do Prepare the study. Run the study. Analyze results. Move results to a report. end for Compare old and new results. end while

Collecting the components: Automation I



#### TECHNISCHE UNIVERSITÄT DARMSTADT

### while Results are unsatisfactory do

Work on algorithms. (Compile the code.) for All studies do Prepare the study. Run the study. Analyze results. Move results to a report. end for Compare old and new results.

# while Results are unsatisfactory do Work on algorithms. (Compile the code.) Run initialization scripts (jobs). Run simulation scripts (jobs). (Run postprocessing scripts (jobs)). Visualize results live in Jupyter notebooks. end while

Collecting the components: Automation II

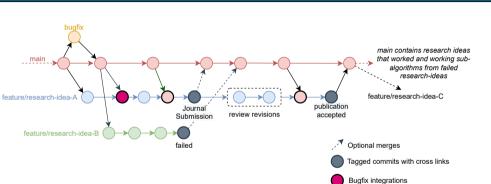




- Manual testing takes a lot of time.
- Manual testing of all previous tests is prone to error - even if V&V scripts do not require metadata.
- Only the researcher knows the details behind the initialization, running and post-processing scripts
- Relevant V&V tests are automated using Continuous Integration (CI).
  - The remote repository starts the so-called Cl test pipeline (a sequence of tests).
  - All tests are automatically run, processed and visualized.

**Collecting the components: Version Control** 





The branching model increases reproducibility.

A Research Software Engineering Workflow for Computational Science and Engineering -M. Schwarzmeier, T. Tolle, J.P. Lehr, I. Pappagianidis, B. Lambie, D. Bothe, C. Bischof, T. Marić Development integrations

### A Research Software Engineering Workflow Collecting the components: Continuous Integration I



- A text (YAML) file is added to a repository, that specifies the tests (jobs) in a CI pipeline.
- When the YAML file is pushed to an upstream git repository (GitLab), GitLab creates a CI pipeline from the YAML file.
- The CI pipeline needs a machine for running tests the GitLab runner.
  - Shared runners on gitlab.com have limited capacity.
  - We can install and register our own GitLab runner.
- A Docker image encapsulates the computing environment.
  - Virtualization/Containerisation increases reproducibility and simplifies testing.
- The Docker image must be publicly accessible for it to be used by a shared runner.

Collecting the components: Continuous Integration II



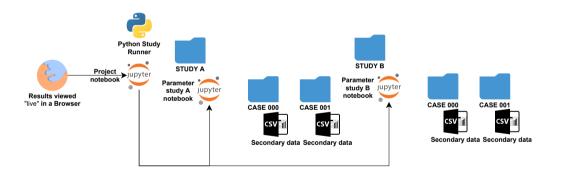
#### TECHNISCHE UNIVERSITÄT DARMSTADT

#### An example CI pipeline

Pipeline Needs Jobs 5 Failed Jobs 1 Tests 0 Building Running Visualization Testing C C C test\_hadamard...  $\mathbf{\nabla}$ build param\_study convert not...  $\mathbf{x}$ 9  $(\mathbf{J})$ ✓ test\_shear2D C

Collecting the components: Parameter studies I

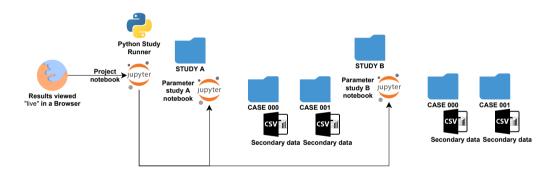




### Organize your simulation studies.

Collecting the components: Parameter studies I





- Associate simulation cases with their metadata.
- {case000 : {N\_CELLS: 32, MODEL : shear2D}}
- Store this information using a standard open-source format (Interoperability in FAIR).



TECHNISCHE UNIVERSITÄT DARMSTADT

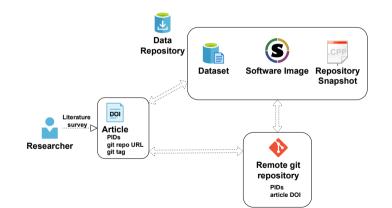
Use Jupyter notebooks<sup>2</sup> and pandas<sup>3</sup> for

- Documentation: geometry, initial and boundary conditions, error norms, comparison data.
- **Data processing**: verification errors (conservation, convergence, stability), validation errors
- Result analysis: interactive and remote, while simulations are running!

<sup>&</sup>lt;sup>2</sup>https://jupyter.org/ <sup>3</sup>https://pandas.pydata.org/

Collecting the components: Cross-linking





**Collecting the components: Others** 



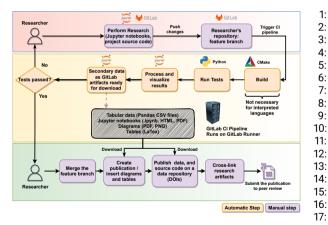
TECHNISCHE UNIVERSITÄT DARMSTADT

We did not cover in this talk:

- Build System and Containerization
- Test Driven Development (TDD)
- Continuous Benchmarking
- Regression Tests
- Parameter space exploration
- Metadata standard for secondary data
- Issue Tracking

# A Research Software Engineering Workflow Assembling the Workflow





- 1: Track changes using version-control.
- 2: while Milestone not reached do 3:
  - for study in studies do ▷ On an HPC cluster.
  - Automate data processing and visualization.
  - Run study.
  - Check results and apply code changes.

#### end for

4:

5:

6:

7:

- 8: if results are improved on the HPC cluster then q٠
  - Push changes to the remote repository.
- 10: if CI pipeline tests pass then
  - Milestone reached
  - Add new tests to the CI pipeline.
  - Merge feature into development branch.
    - Cross-link publication, data, and source code.
  - end if
- 16: end if
- 17. end while

# We run a Knowledge Base

that contains much more information



home - Knowledge Bess - X	
· 0 0	🗘 🚯 hitzyschooleitigt beer attelling de
iFDi4ing	
	GETTING STARTED
Leane	This knowledge base is based on a weinitizer for increasing the waiting of scenario and based for the contract and based for the originated from the work at the SFB 104 at TU
	Darwetadt: Far this pages places see our section in these are
	the presented this knowledge base at the PONDEM'22 #PC demonsh the directly to the value resorting or move to can not floatery.
	The activate here are equipped with logit to help that them as well as they are labeled with the chapters according to the institution of allowmentioned paper, namely constrained paper, and paper, acceled to the chapters of the chapters and the
	conceptry, independent the antibiotic program of a minimum or a full workflaw.
	It is reconversified to read the paper constituty and to try to follow the desting started is the guides-section.
	Consultation hour
	0.00
	Ven any now offering desaultation focus www.ywellenaby.at.93.AM (Inter), GETDEGET, Bring visite (Hall, in case part one case of a prior prior and the prior and the prior of the prior and the prior a
	Contributing
	We are always harpy about proper
	Supported by
	Runding
	The authors would like to thank the Faderal Edwarmmer and the Instant of Government of the Sadera as well as the Just to Science SaMe and support within the framework of the Artitional consortionar will be 1981 THE Antonia the Instant Framework in Sadera Sa
	enclare glassipariore en los or ore non-render el con control el c
	Weblate Chemon
	NFDi4ing
	word

#### knowledge-base.nfdi4ing.de

# Glossary terms display helpful information when you hover over them



#### TECHNISCHE UNIVERSITÄT DARMSTADT

# Opening a Merge Request on GitLab

If you already have a <u>feature-branch</u> that you want to <u>merge</u>, in collaborative projects it is best to open a <u>Merge Request (MR)</u> on <u>where all the collaborators</u> can

revise and discuss the new featue A branch is a development line.

Another great possibility to open a MR is from an existing 'issue', where a branch will be created as well. Please follow **these steps**.

https://www.vecteezy.com/free-vector/mouse-pointer, Darmstadt, 21.01.2022

### We just added interactivity by offering a weekly consultation hour



The consultation hour is every wednesday at 10 AM.

- Information is displayed on the landing page.
- It is a small initial investment,
- but it scales badly.

We are curious, whether this will work out as:

- super-interactive for the users and
- super-direct feedback for us.

Let us know your opinion in the afterwards discussion!



# **Further reading**



 This presentation is available at https://doi.org/10.5281/zenodo.7215818

### The Knowledge Base

is available at https://knowledge-base.nfdi4ing.de

#### Preprint describing the workflow

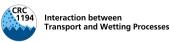
A Research Software Engineering Workflow for Computational Science and Engineering; Marić, Gläser, Lehr et al., 2022, https://doi.org/10.48550/arXiv.2208.07460

#### Slides about the workflow at full length including hands-on "Continuous" Integration of Scientific Software (in Computational Science and Engineering);

Marić et al., 2021, https://zenodo.org/record/5522820.YnTOvnVByXI

# Acknowledgements





Funded by the German Research Foundation (DFG) – Project-ID 265191195 – CRC 1194 : Z-INF



Funded by the German Research Foundation (DFG) – Project-ID 442146713 – NFDI4Ing : Base Service S-2 Research software development