REPRODUCIBILITY IN PRACTICE
TOOLS AND METHODS FOR EVERY STEP

RDA DE 22 conference - Research Data Alliance Germany

24.02.2022 | OLIVER BERTUCH, CENTRAL LIBRARY

JÜLICH Forschungszentrum

AGENDA

- Introduction
- Example workflows and solution ideas
- The missing link: pipelines and publications
- Automating software publications with HERMES

Please write any questions into the chat or hold for later.

JÜLICH Forschungszentrum

ALWAYS REMEMBER

You are not alone!

ReproHack Hub

Host a hackathon doing live peer reviews trying to reproduce your publication

German Reproducibility Network

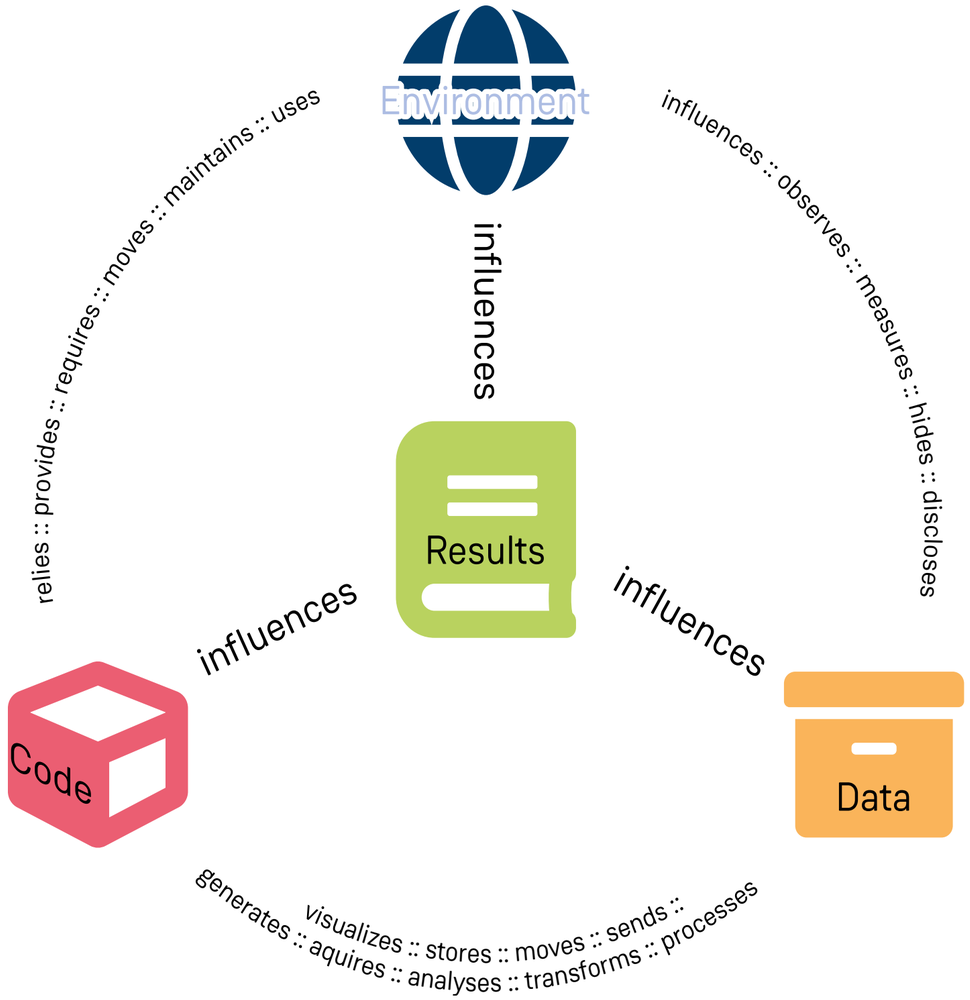
Reach out to experts for reproducibility in Germany

Turing Way Project

The place-to-go for all things reproducibility

CODE, DATA AND ENVIRONMENTS

An overlooked codependency



JÜLICH Forschungszentrum

A SIMPLIFIED SCIENTIFIC WORKFLOW

Let's start small

SPREADSHEETS

Simple but not reproducibility-friendly

Pros:

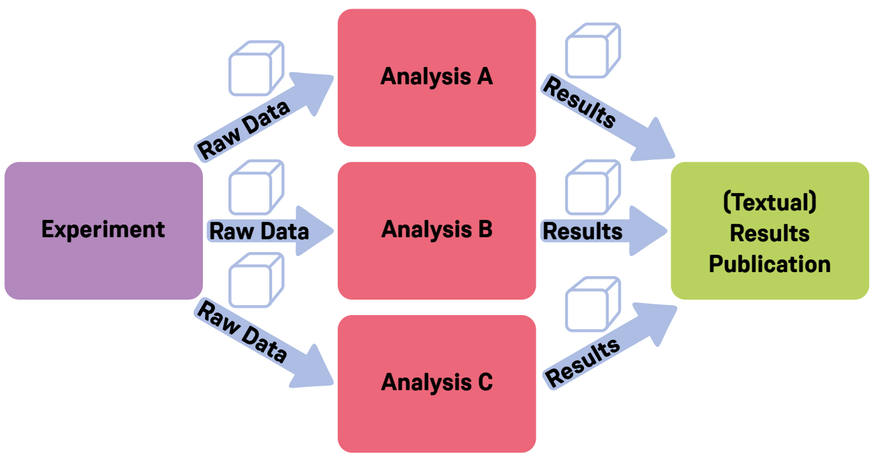
- Interactive data exploring
- No programming skills
- Easy to share
- *Perfect to view and edit tabular data files*

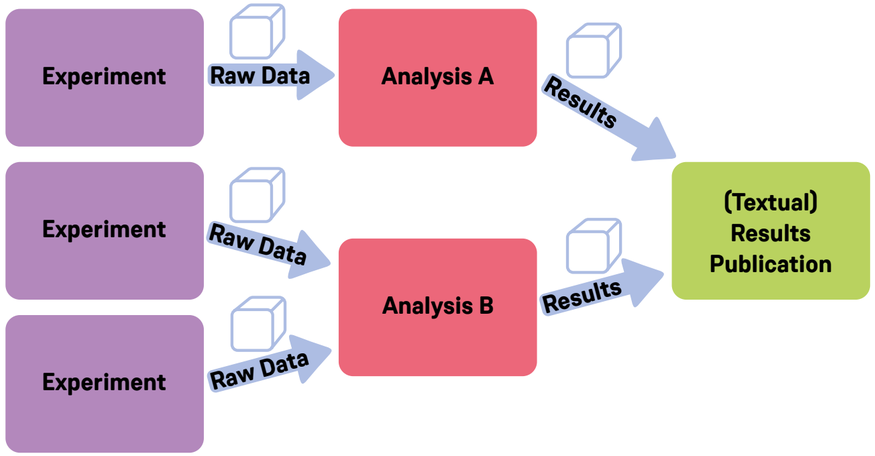
Cons:

- Hard to test
- Hard to debug
- Hard to track changes
- Hard to reproduce
- Hard to preserve & archive
- Hard to extend & program
- Tight coupling of data and code
- Potential legal issues
- Version compatibility
- Interesting bug history (genes renamed!)

ISOLATED SCIENTIFIC WORKFLOWS

Sections of reality, but not your entire research life





RESEARCH IDE I

Proprietary flavors - tasty but costly

Pros:

- Decoupling of data + code
- Still interactive
- At least minor coding skills
- Many use ASCII files for code
- Some provide test frameworks and version control integration
- Share and reuse possible

Cons:

- Usual SE chaos hazard
- Many "walled gardens"
- Huge cost factor
- Extension packages for convenience at extra cost
- License requirement impedes sharing & reuse
- Usage of extensions makes sharing hard
- Troublesome to archive

RESEARCH IDE II

Open Source flavors

Pros:

- **Free Open Source Software**
- Decoupling of data + code
- Still interactive
- At least minor coding skills
- ASCII files for code
- Test frameworks and version control integration possible
- **Sharing is easy**, reuse possible
- **Easy to archive**

Cons:

- Usual SE chaos hazard
- Maybe cumbersome for complex or production grade projects
- Short distance to full-fledged coding ecosystem
- Notebooks = Junk Food? [1], [2], [3], [4], fastai/nbdev to the rescue!

JÜLICH Forschungszentrum

RIAAS

No radio station but "Research IDE as a Service"

Often browser based

- Researcher convenience is key
- Some propietary tools (i. e. MatLab, Stata) provide integrations
- Many local, institutional cloud offerings (near to big datasets!)

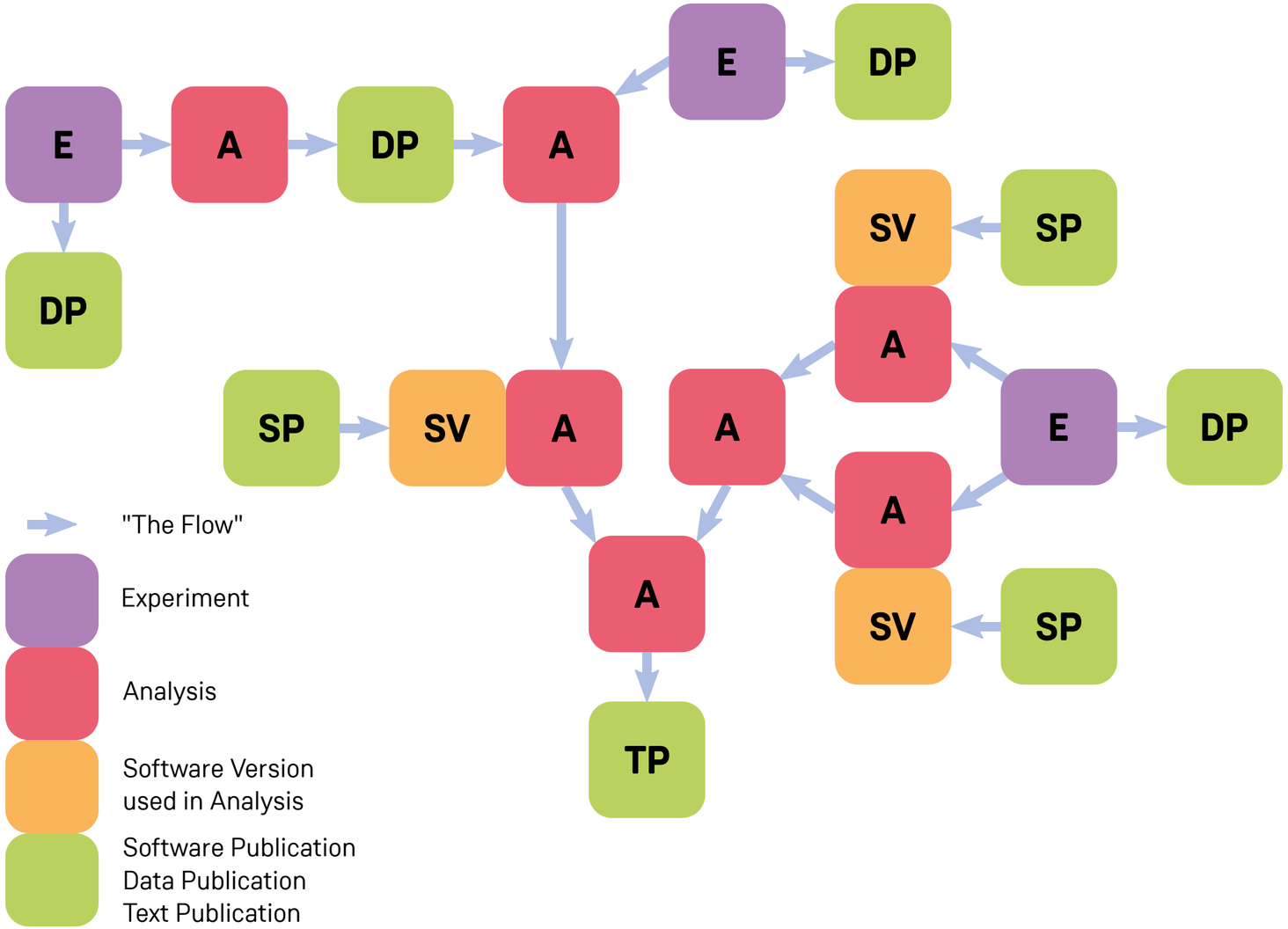
Paid external offerings

- Examples: CodeOcean, WholeTale, MyBinder
- Some provide fire-and-forget archive depositing
- Beware of vendor lock-in effects and legal issues!

JÜLICH Forschungszentrum

THE MISSING LINKS

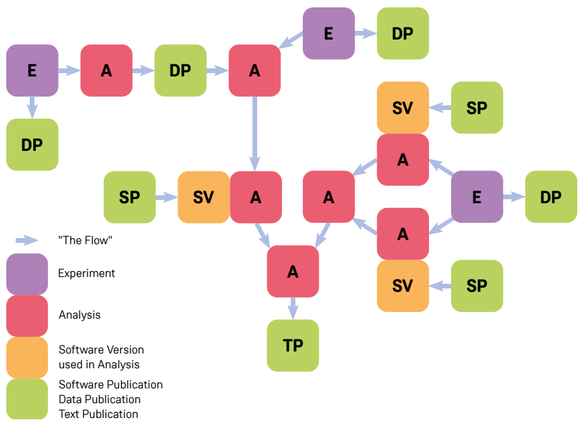
Reality be more like this. Do you keep track of this?



JÜLICH Forschungszentrum

THE MISSING LINKS

Transform reality into ...



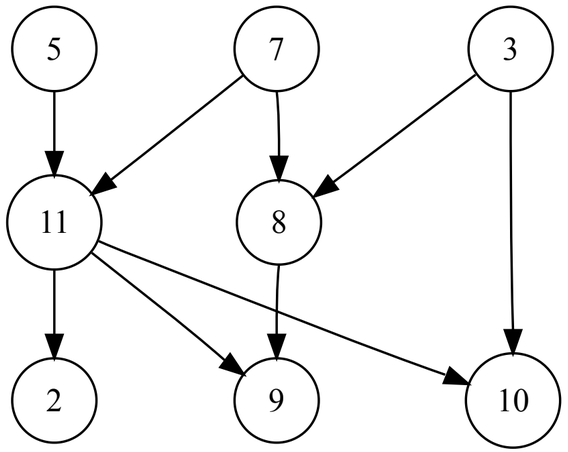
... pipeline workflows ...

Using DSLs*:

- CWL
- NextFlow
- SnakeMake
- or Galaxy, KNIME, Guix, Jupyter, Shell, etc.

*) Domain Specific Language

... and graphs!



CS: "directed graphs"

RDM: "Knowledge Graph", "Data lineage" and "Provenance"

JÜLICH Forschungszentrum

PIPELINES & WORKFLOWS

Pros:

- Loose coupling of data & code
- Reuse existing codes
- Easy to preserve & archive
- Easy to reproduce
- Easy to share & reuse (e. g. workflowhub.eu)
- Self-Documenting, no junk food
- Plays well with tools like Singularity and all things HPC

Cons:

- Verbose
- Needs more & new skills
- Steep Learning Curve
- Ecosystem not yet grown up, not very integrated in RDM infra.

*The landscape of workflow systems for scientific applications is notoriously **convoluted** with hundreds of seemingly equivalent workflow systems, many **isolated** research claims, and a **steep learning curve**.*

Quoted from Da Silva 2021

JÜLICH Forschungszentrum

FUTURE: GOTTA GRAPH 'EM ALL

Just a quick glance

- CWLProv to create provenance from workflows
- RO-Crate bundles in a package:
  - provenance,
  - workflows,
  - resources,
  - people, licenses and more
- SciMesh for interoperable electronic lab notebooks
- Metadata4Ing to capture the whole data generation process

JÜLICH Forschungszentrum

STEPS TO EASE GRAPH CREATION

- Make all resources FAIR and reference by their identifier
- Make the ecosystem more suitable and accessible for science people
- Integrate RDM and RSE infrastructures into workflow tooling

FAIRable resources

- Scholarly Publications ✅
- People ✅
- Licenses ✅
- Research Organisations ✅
- Research Datasets ✅
- Research Software ...? ✅

But - do **you** publish **your** software to make it FAIR4RS?

Bonus: would/do you cite the software you're using?

MAKING SOFTWARE F.A.I.R.

Why?

- Software is an important research output [1], [2], [3]
- Ergo: Reproducibility relies on Research Software Engineers work, too
- Consequence 1: adademic credibility is due to RSEs, too
- Consequence 2: play by scholarly rules - publish software!

What's the catch?

- It's a lot of work.
- It's currently manual work mostly.
- Chicken and egg situation:
  - When publishing is to hard, no one will cite.
  - When no one cites, why would you do any heavy lifting?

JÜLICH Forschungszentrum

INTRODUCING PROJECT HERMES

Idea

Automate it (as much as possible)

with **HE**lmholtz **R**ich **ME**tadata **S**oftware Publication

https://software-metadata.pub

Team and Funding

Made by DLR, FZJ and HZDR

Funded from 07/2021 till 06/2023

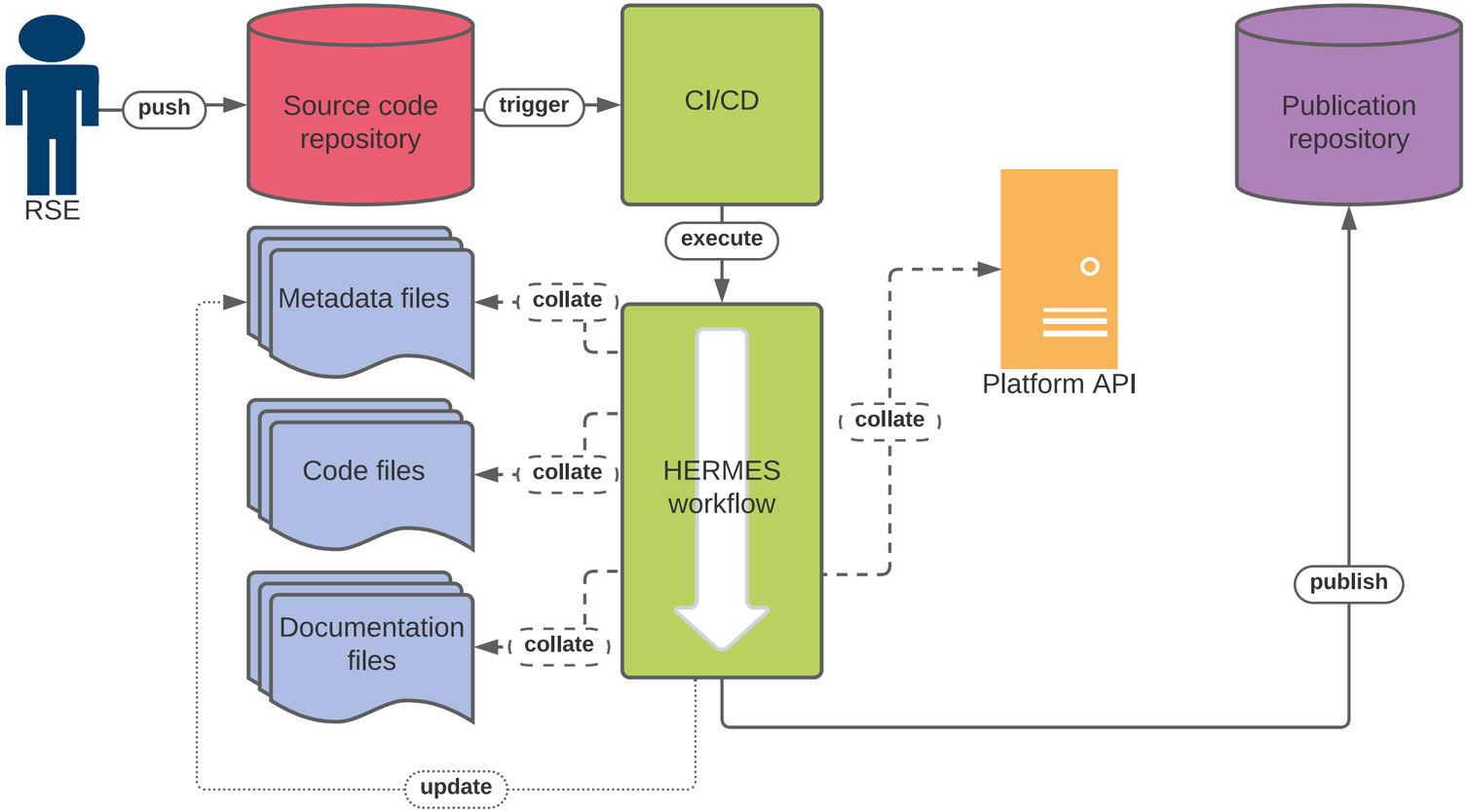
by *Initiative and Networking Fund* of *Helmholtz Association*

within <HMC> | HELMHOLTZ METADATA COLLABORATION project call (ZT-I-PF-3-006)

JÜLICH Forschungszentrum

HERMES CONCEPT I

Visualization of (simplified) use-case



Find details in our concept paper: arXiv:2201.09015

Please leave feedback with PeerPub

HERMES CONCEPT II

Workflow

Chaining four pipelines

1. Automated metadata collation from different sources
2. Process metadata: validate and merge
3. Deposition into publication repositories optionally w/ artifacts
4. Post-processing like updating metadata files

Scope

- Allow mixed-style source repositories and multiple targets
- Avoid dead services, reuse CI/CD and workflow DSLs
- Be open for reuse, extension and customization
- Targeting Dataverse and InvenioRDM in first iteration

THANK YOU FOR YOUR ATTENTION!

$ whoami

Oliver Bertuch
Central Library

$ reachout

o.bertuch@fz-juelich.de
@poi_ki_lo_therm
@poikilotherm

$ ls /workplaces

Research Data Management FZJ_RDM
+
HElmholtz **R**ich **ME**tadata **S**oftware
Publication @ HMC

$ attribution

JÜLICH Forschungszentrum