NFDI4Ing – the National Research Data Infrastructure for Engineering Sciences

Task Area *DORIS*: Research Data Management in High-Performance Measurements and Computation

Vasiliki **Sdralia** | 11/05/2023 | 15:00

TUM

lrz Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften

HLRS High-Performance Computing Center | Stuttgart

RWTH AACHEN UNIVERSITY

NFDI

“**Nationale Forschungsdateninfrastruktur**” (German National Research Data Infrastructure) – NFDI

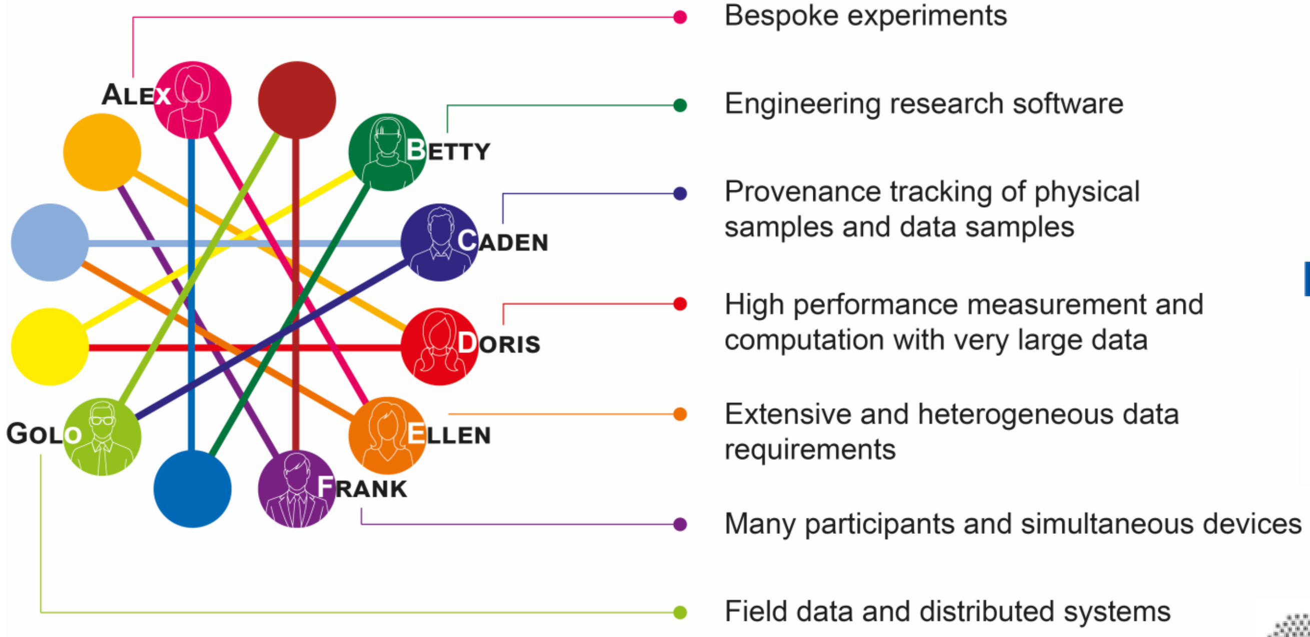
- Registered association funded by the federal government and the federal states (90 Mio. Euro / year)
- Goals:
  - Set **standards** in data management
  - Digital, regional and interconnected **data storage**
  - Enable **innovations and new findings** through available research data

- 29 consortia selected by the German Research Foundation (DFG)
  - from cultural sciences, social sc
    sciences

NFDI4Ing

Consortium for Engineering Sciences – NFDI4Ing

- 14 “steering institutions”
- 30 participant institutions
- 8 engineering archetypes



Bespoke experiments

Engineering research software

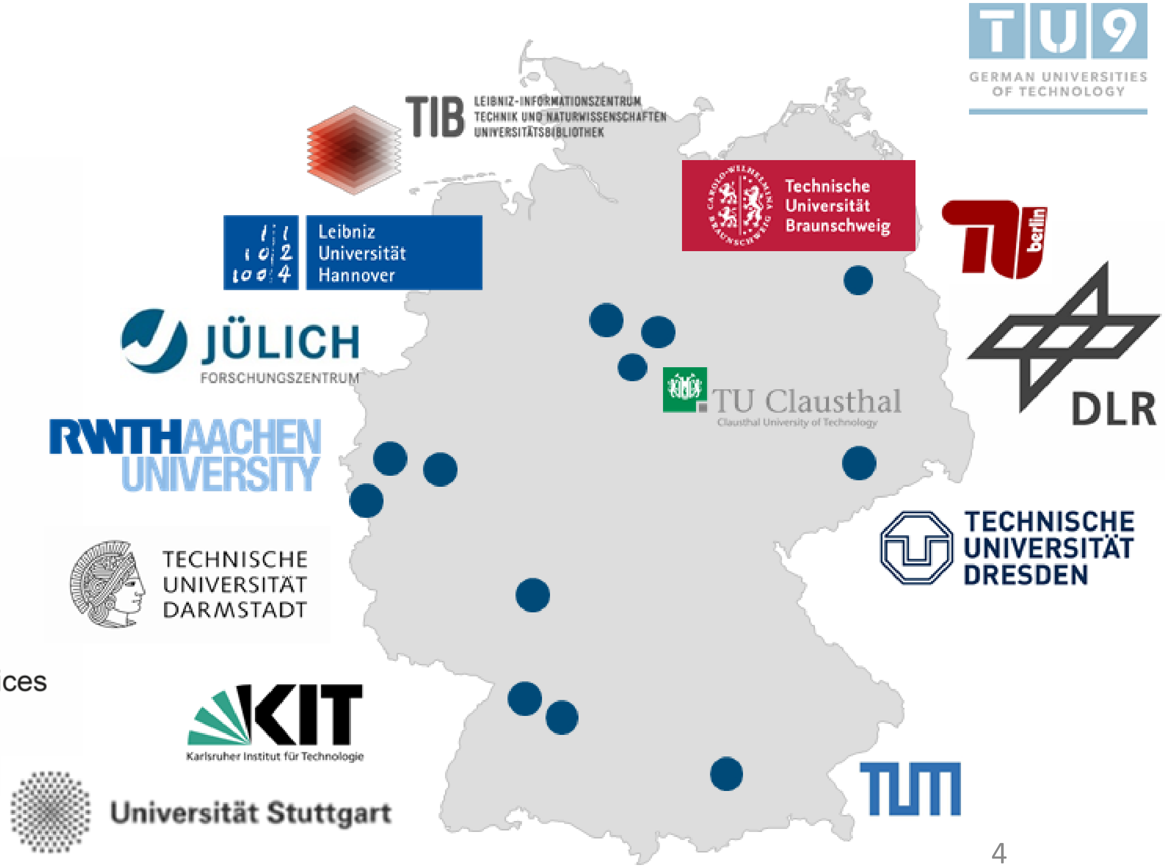
Provenance tracking of physical samples and data samples

High performance measurement and computation with very large data

Extensive and heterogeneous data requirements

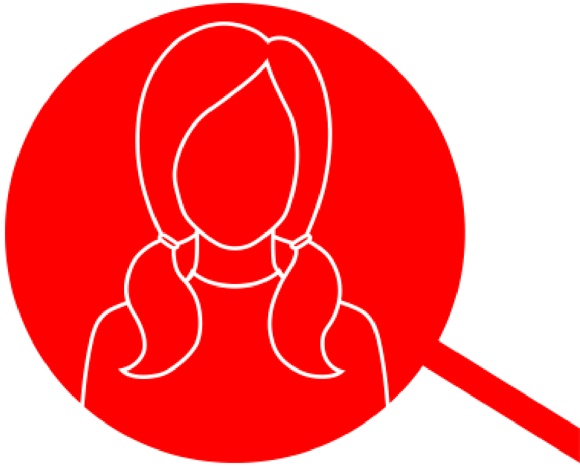
Many participants and simultaneous devices

Field data and distributed systems



TU9
GERMAN UNIVERSITIES OF TECHNOLOGY
TIB
LEIBNIZ-INFORMATIONSZENTRUM TECHNIK UND NATURWISSENSCHAFTEN UNIVERSITÄTSBIBLIOTHEK
Technische Universität Braunschweig
TU berlin
Leibniz Universität Hannover
JÜLICH
FORSCHUNGSZENTRUM
DLR
TU Clausthal
Clausthal University of Technology
RWTH AACHEN UNIVERSITY
TECHNISCHE UNIVERSITÄT DRESDEN
TECHNISCHE UNIVERSITÄT DARMSTADT
KIT
Karlsruher Institut für Technologie
Universität Stuttgart
TUM

Archetype DORIS: HPMC

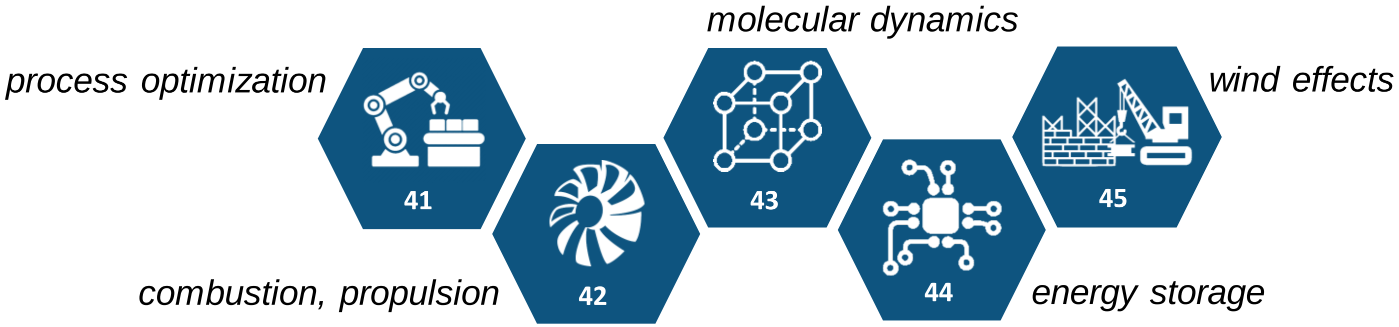


*… I'm an engineer conducting and post-processing high-resolution and **high-performance measurements and computation** (simulation) **with very large data** on HPC systems.*

The data sets I work with are extremely large and as such are largely immobile. This mandates tailored, hand-made software."

My needs are

- ➔ *Enable **exchange** of **huge** high-quality **datasets**.*
- ➔ *Provision of HPC-data to foster **wide-spread usage**.*
- ➔ *Drive NFDI-wide **new methodologies** for data sharing*



DORIS's patron is Christian Stemmer

HPMC Research Data

Characteristics

- Data are created and stored in personalized accounts directly at HPC centres → **no indexing** by repositories or search engines
- **Special hard- & software** required for creating, reading or processing data
- Size: terabyte to petabyte → **data is not mobile**
- “Data” consists of **various components** (code, raw data, processed data, metadata etc.)
- No established **terminology** or metadata scheme
- Little best-practice or showcases for research data management

Implementation of FAIR data principles?

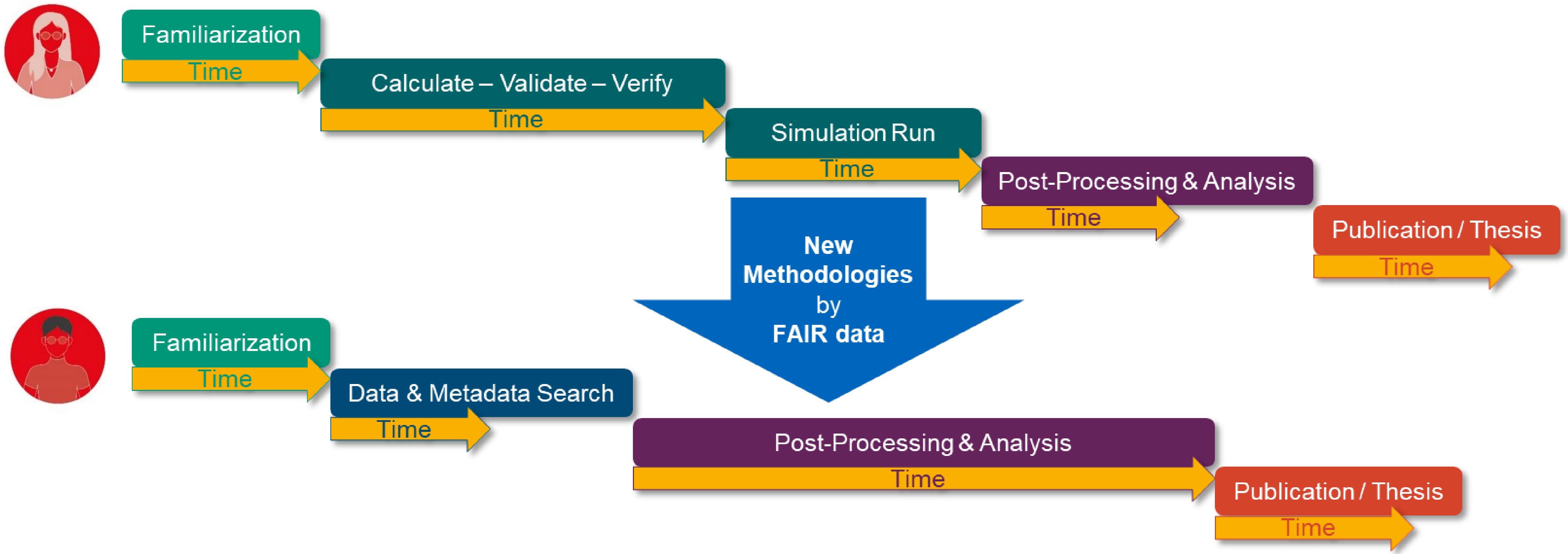
Findable?	Storage in personalized accounts, little metadata
Accessible?	No access for third parties, insufficient transfer tools
Interoperable?	Depending on formats and enriched metadata
Reusable?	Computing time at HPC centres required or virtualization (e.g. container)

HPMC Research Data

Why research data management for HPMC-Data?

- Scientific integrity and fulfilment of (external) compliance (e.g. DFG)
- Secondary research (e.g. energy consumption or temperature in HPC centres)
- **New findings, new methodologies, new workflows, new opportunities by re-using existing data**

HPMC Research Data: New Methodologies



DORIS: Measures and Milestones

- **Accessibility** and **access rights**, data security and sovereignty

- **Support** for third-party users & community-based **training**, provision of post-processing algorithms and modules

- **Metadata** definitions & **terminologies**, support to data-generating groups

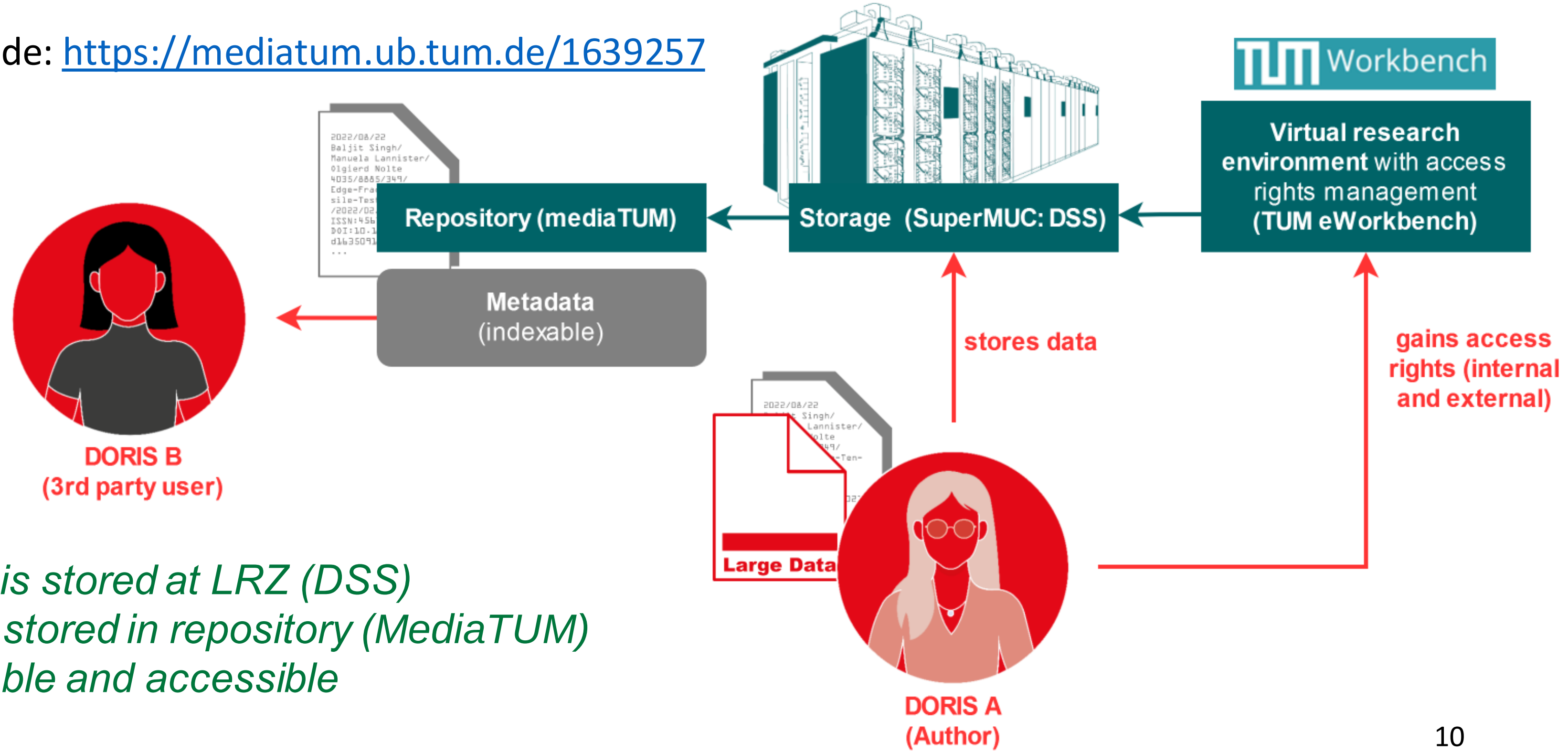
- **Storage & archive** for very large data

- **Reproducibility** on large-scale high-performance systems

DORIS: Activities & Results

Data storage and sharing (TUM only)

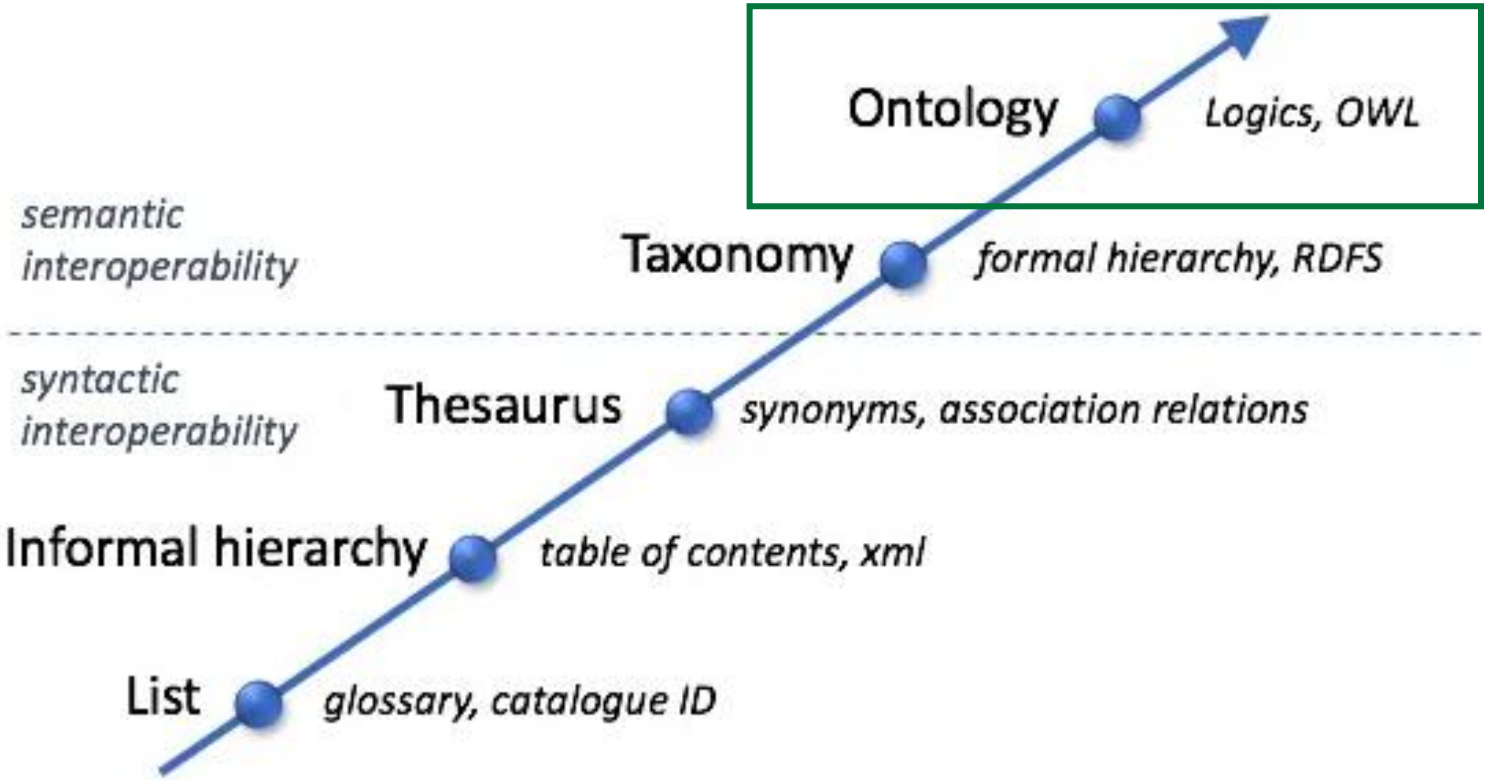
- Store data in DSS (LRZ) / manage data via TUM Workbench
- User guide: https://mediatum.ub.tum.de/1639257



→ (large) data is stored at LRZ (DSS)
→ metadata is stored in repository (MediaTUM)
→ data is findable and accessible

DORIS: Activities & Results

Metadata – Terminologies and automated metadata extraction

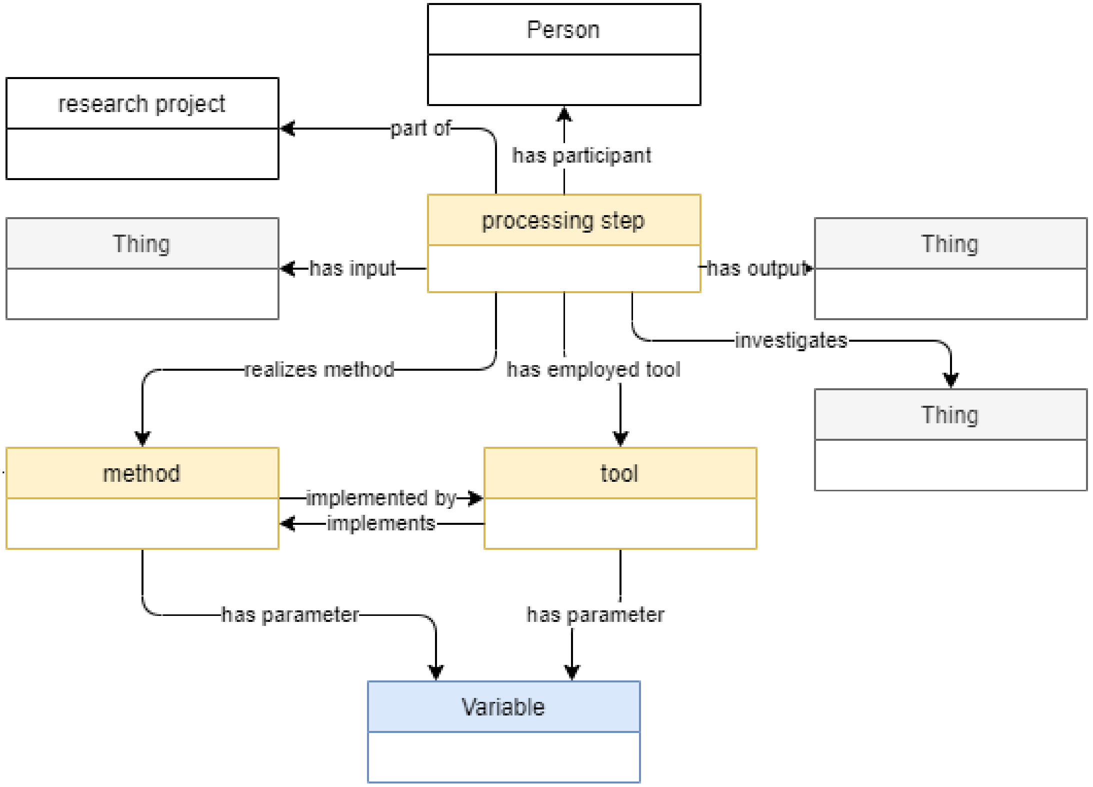


https://i1.wp.com/emmc.info/wp-content/uploads/2019/05/alternatives_2.png?w=600&ssl=1 (04.05.2021)

DORIS: Activities & Results

Metadata4Ing - An ontology for describing the generation of research data

https://git.rwth-aachen.de/nfdi4ing/metadata4ing/metadata4ing



https://git.rwth-aachen.de/nfdi4ing/metadata4ing/metadata4ing (modified)

- **Processing step as the central element**
- Connects in- and output
- Describes
  - The object of investigation
  - What has been done („**method**“)
  - What has been used (“**tool**”)
  - By whom („person“)
- Specifies the **parameters** used

HPMC workflows in Metadata4Ing

HPMC extension / domain-ontology

- Set **classes and properties**
  - **Domain**

    → Flow? Solid state?
  - **Processing Step**

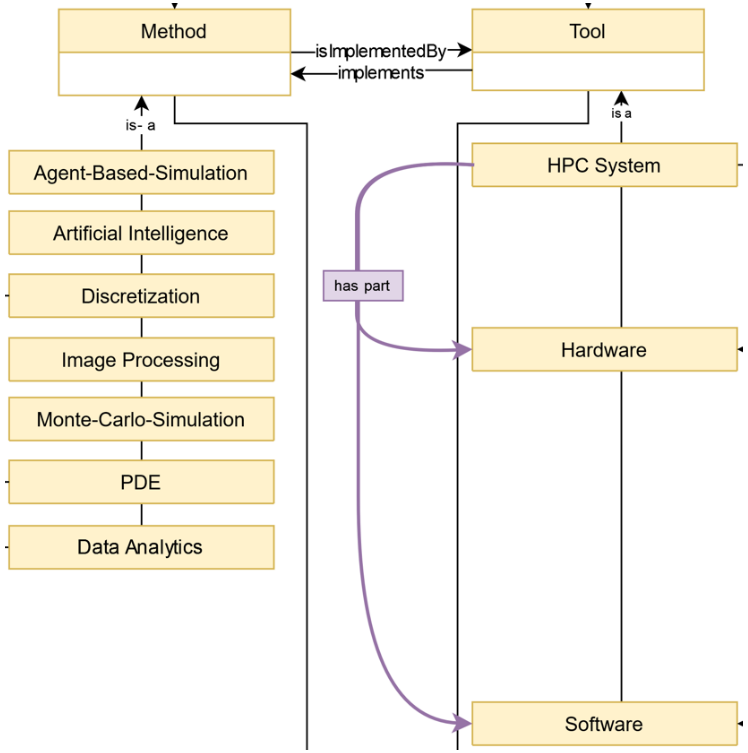
    → Compilation, Pre-Processing, Simulation run, Post-Processing etc.
  - **Tool**

    → HPC system ("*has part*:" hardware & software)
  - **Method**

    → PDE, Monte-Carlo-Simulation, Image processing etc.

  - optional: detailed metadata, e.g. energy consumption, used nodes, temperature in cluster etc.

    → useful for secondary research



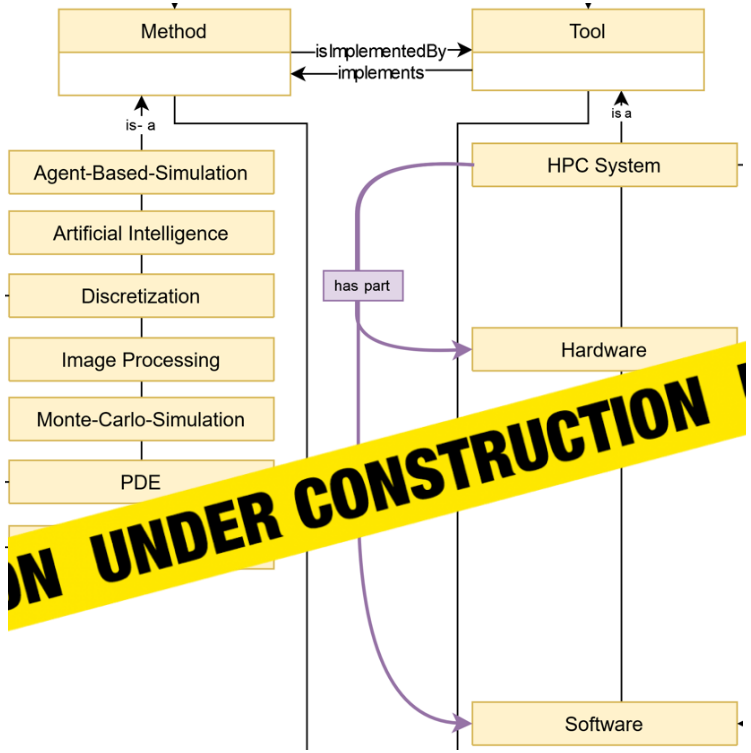
12.05.2023

HPMC workflows in Metadata4Ing

HPMC extension / domain-ontology

- Set **classes and properties**
  - **Domain**
    → Flow? Solid state?
  - **Processing Step**
    → Compilation, Pre-Processing, Simulation run, Post-Processing etc.
  - **Tool**
    → HPC system (“*has part*:” hardware & software)
  - **Method**
    → PDE, Monte-Carlo-Simulation, Image processing etc.

- optional: detailed metadata, e.g. energy consumption, used nodes, temperature in cluster etc.
  → useful for secondary research

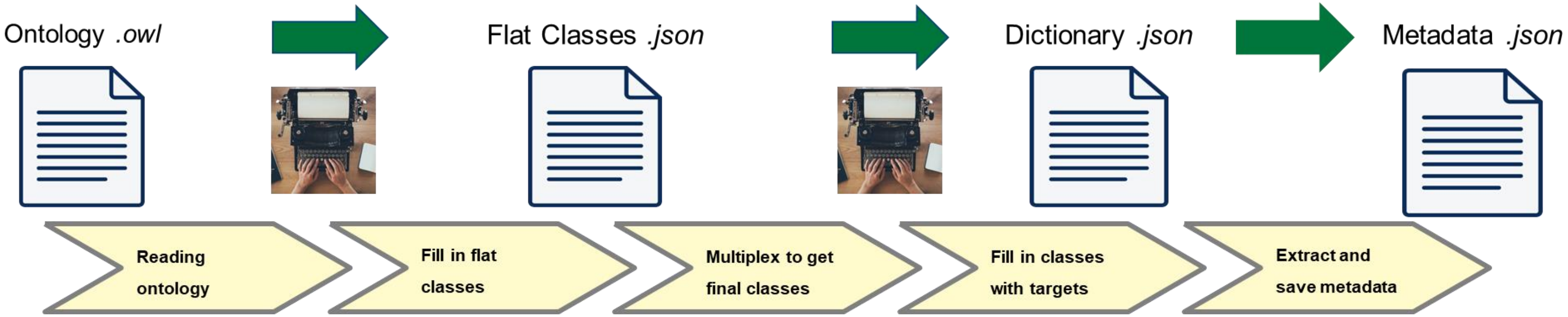


DORIS: Activities & Results

Metadata – Terminologies and automated metadata extraction

gitlab.lrz.de/nfdi4ing/crawler

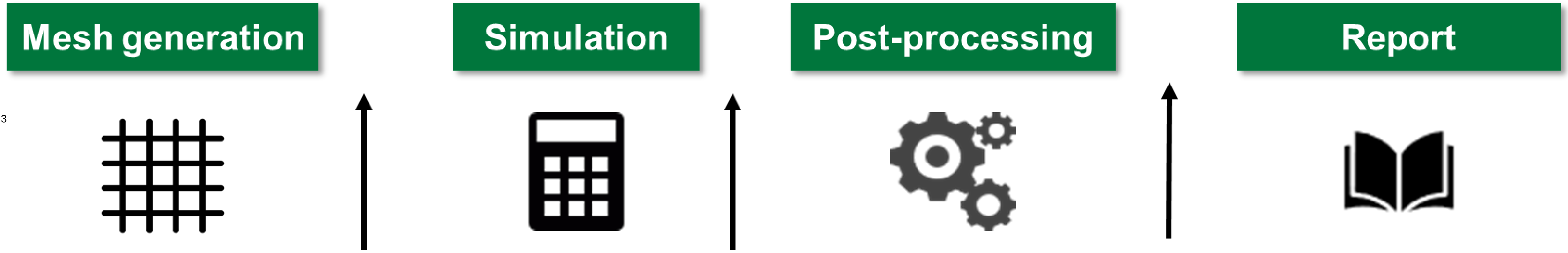
- **HOMER** (**H**PMC tool for **O**ntology-based **M**etadata **E**xtraction and **R**e-use)
- Publication (preprint): https://preprints.inggrid.org/repository/view/12/
- HOMER is a flexible python-based application that, through limited user input, automates metadata extraction starting from any ontology file.
- Metadata can be retrieved from text and HDF5 files, from outputs of console commands or can be directly hardcoded in the configuration file.
- Easy to integrate within (script-based) workflows & employable after any processing step



DORIS: Activities & Results

Metadata – Terminologies and automated metadata extraction

- Depending on the application, the crawler can be used at different steps within the workflow of CFD (or similar) applications
- Wherever data files are created, the crawler can be used to extract relevant metadata



Mesh generation

Simulation

Post-processing

Report

DORIS: Activities & Results

Reproducibility: Containerization of CFD Workflows on HPC systems

- Evaluate the **feasibility of containers/dockers** for reproducibility for HPC systems.
- Develop standards on reproducibility on HPC systems.
- Prepare best-practice guidelines on reproducibility issues for HPMC users.

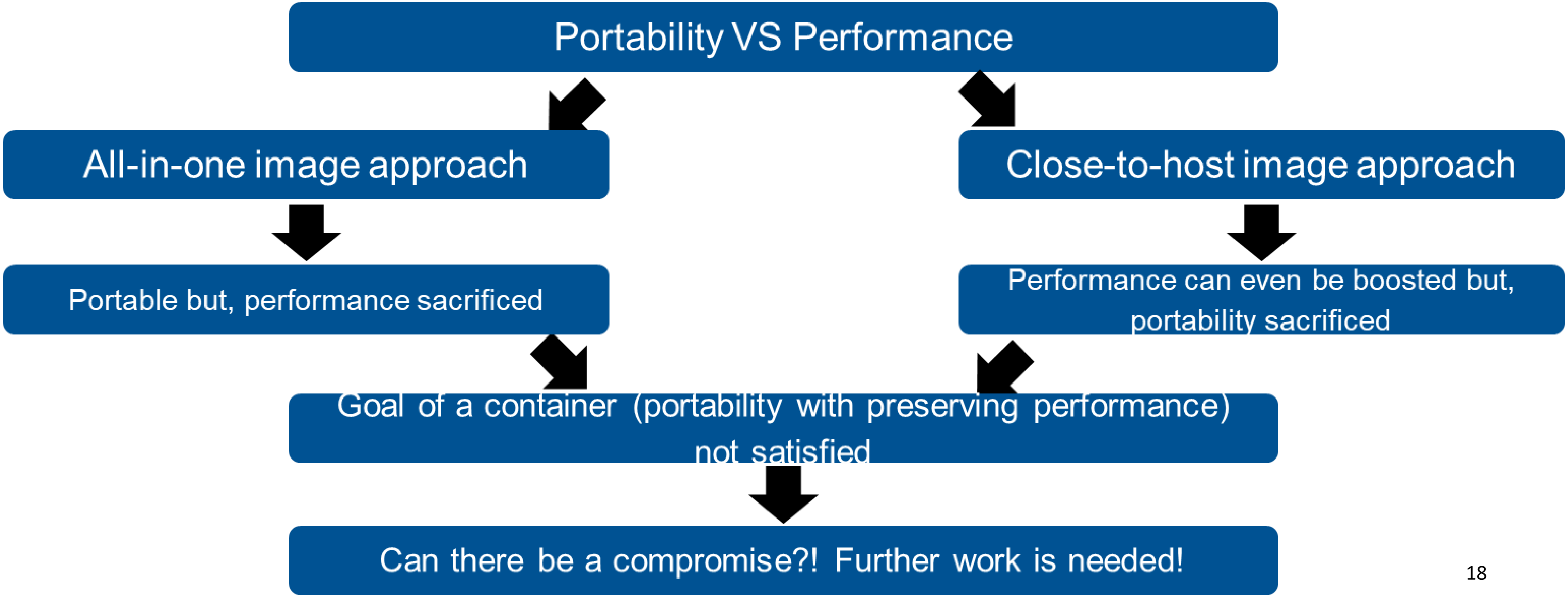
Current Status

- Application of containerization to a **typical CFD problem**
  - ➢ simple multiphase-flow problem investigated with MPI-parallel code ALPACA (Adaptive Level-set Parallel Code Alpaca) to be run at LRZ Container platforms: Docker / Singularity / Charliecloud
- Different approaches have been investigated:

  1. All-in-one image: From libraries to code, everything in the container!
  2. Close-to-host image: Everything in the image must mimic the runtime system!

DORIS: Activities & Results

Reproducibility: Containerization of CFD Workflows on HPC systems



Portability VS Performance

All-in-one image approach

Close-to-host image approach

Portable but, performance sacrificed

Performance can even be boosted but, portability sacrificed

Goal of a container (portability with preserving performance) not satisfied

Can there be a compromise?! Further work is needed!

Dummy Subheader

Reproducibility: Containerization of CFD Workflows on HPC systems

- Containerization of CFD workflows is possible
  - Start off with Docker and Charliecloud
- Choose approach depending on use case

	HLRS		lrz
Vanilla code	✓	✓	✓
All-in-one	?	?	◌
Close-to-host	?	?	◌

DORIS: Activities & Results

Reproducibility: Containerization of CFD Workflows on HPC systems

- Evaluate the feasibility of containers/dockers for reproducibility for HPC systems.
- **Develop standards on reproducibility** on HPC systems.
- **Prepare best-practice guidelines on reproducibility** issues for HPMC users.

→ → **Survey:** https://wiki.tum.de/display/rdm/Survey%3A+Reproducibility+and+Postprocessing+in+HPC

https://www.flaticon.com/de/kostenloses-icon/diskussion_2821271

Information and Links

Downloads

- **Slides**, **publications** etc.: https://zenodo.org/communities/nfdi4ing?page=1&size=20

- **DORIS' Software**: https://gitlab.lrz.de/nfdi4ing

- **Data Management Plan - HPMC-template:** https://zenodo.org/record/5801838#.YjSN0DUxmUk

- **Metadata4Ing Ontology** for Workflows in (Engineering) Science
  - Documentation: https://nfdi4ing.pages.rwth-aachen.de/metadata4ing/metadata4ing/index.html
  - GitLab: https://git.rwth-aachen.de/nfdi4ing/metadata4ing/metadata4ing
  - Publication: https://zenodo.org/record/5957104#.ZBxm5M7MKUk

- **Software Metadata Schema** CodeMeta: https://codemeta.github.io/terms/

- **Handreichung zu rechtlichen Aspekten** des Forschungsdatenmanagements: https://mediatum.ub.tum.de/1690463

Information and Links

Contact

- **Newsletter:** https://lists.tu-darmstadt.de/mailman/listinfo/nfdi4ing_taskarea_doris

- **Workshops:**
  - **Research Data Management for PhD students** (TUM only) on October 18
  - **Research Data Management in HPMC** in April 2024 (tbd)

- **Mail**: info-doris@nfdi4ing.de

- **Web**:
  - https://www.epc.ed.tum.de/en/aer/research-groups/nfdi4ing/ (TUM)
  - https://nfdi4ing.de/archetypes/doris/ (NFDI4Ing)

Outlook

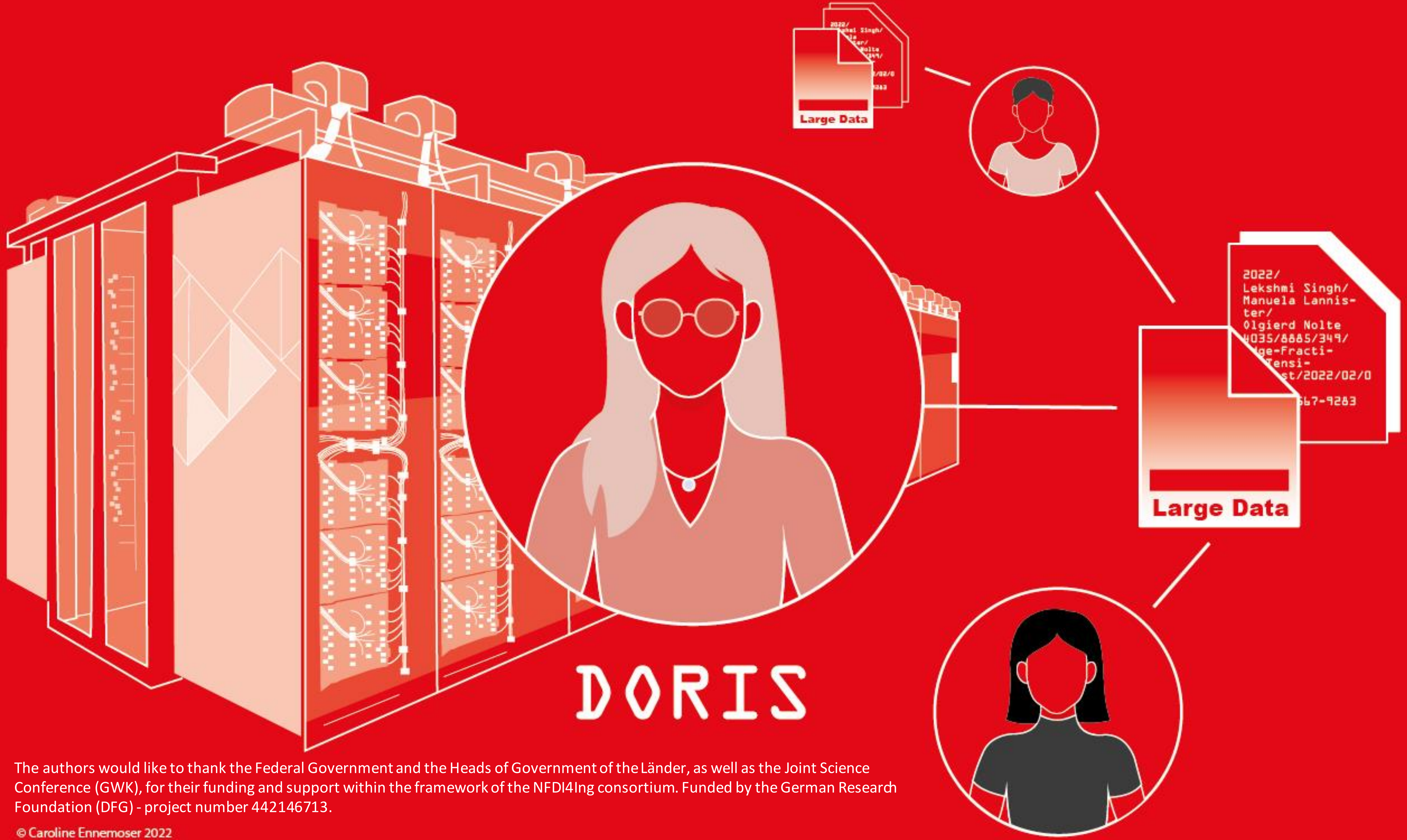
- Test and comparison of **transfer, storage, containerization and post processing** tools
- Community based further development of the **Metadata4Ing (sub-)ontology**
- Further development and functionality expansion of the DORIS **metadata crawler**
- **Publication** in journal
- Provision of metadata through the NFDI4Ing metadata hub
- **Workshop** on best-practices for RDM
- Foster the possibilities of data (re-use) projects at HPC centres or within multicloud projects (NFDI section common infrastructures)
- Installation of granted LRZ **cloud servers** to provide large data and VM images

Further Information

Acknowledgement

The authors and speakers would like to thank the Federal Government and the Heads of Government of the Länder, as well as the Joint Science Conference (GWK), for their funding and support within the framework of the NFDI4Ing consortium. Funded by the German Research Foundation (DFG) - project number 442146713.

The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time on the GCS Supercomputers SuperMUC at Leibniz Supercomputing Centre and JUWELS at Jülich Supercomputing Centre.



The authors would like to thank the Federal Government and the Heads of Government of the Länder, as well as the Joint Science Conference (GWK), for their funding and support within the framework of the NFDI4Ing consortium. Funded by the German Research Foundation (DFG) - project number 442146713.