Chair of Aerodynamics and Fluid Mechanics
TUM School of Engineering and Design
Technical University of Munich

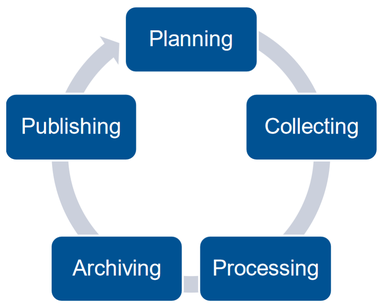
NFDI4Ing - the National Research Data Infrastructure for Engineering Sciences

DORIS – Data from High-Performance Measurement and Computation (HPMC)

Background and Motivation

What are Research Data?

All data and material generated during the **research data life cycle.** There is a great need to manage and to take advantage of these huge amounts of data that are generated every day in the research community.

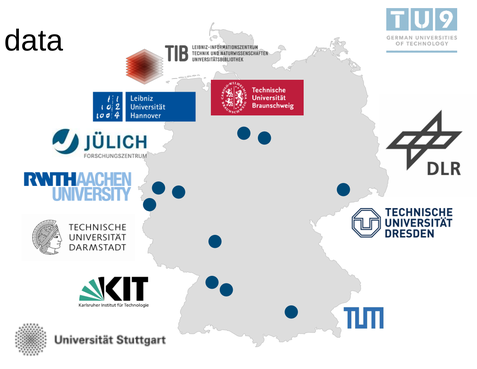


Why Should I Invest in Research Data Management?

- Scientific integrity and compliance with funders' guidelines and compliance (e.g. DFG, TUM, Horizon)
- Increasing political importance (e.g. Federal Data Strategy, NFDI)
- Simplifies re-use by third parties as well as secondary research
- New findings, new methodologies, new workflows by re-using existing data
- Saves time and resources in the long run, minimizes risk of data loss
- Visibility and improved odds for collaborations and funding
- New publishing opportunities (e.g. peer reviewed data publishing)
- Re-use or dissemination of data
- Increase of citations through published research data

NFDI4Ing

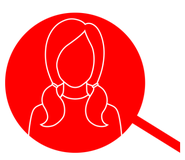
The NFDI4Ing consortium is jointly funded by federal and state government. It's objective is to systematically index, edit, interconnect and make available the valuable stock of data from engineering science.



Research Data Management with Large (HPMC) Data

NFDI4Ing derived seven **method-oriented archetypes** being representative for the majority of the engineering sciences:

ALEX	BETTY	CADEN	DORIS	ELLEN	FRANK	GOLO
bespoke experiments with high variability of setups	engineering research software	provenance tracking of physical samples & data samples	high performance measurement & computation	extensive and heterogeneous data requirements	many participants & simultaneous devices	field data & distributed systems



*Doris is an engineer conducting and post-processing high-resolution and **high-performance measurements and computation** with very large data on HPC systems.*
The data sets can be extremely large and as such largely immobile. This mandates tailored, hand-made software.

Doris is conducting projects at all tier 0 (EU) **Gauss Centre for Supercomputing** facilities. Engineers using these supercomputers mostly come from fluid mechanics, thermal and heat science, materials and construction engineering.

What Are HPMC Research Data?

High Performance Measurement

- Measurement data
- Metadata (hardware, method, processing steps, descriptive data etc.)

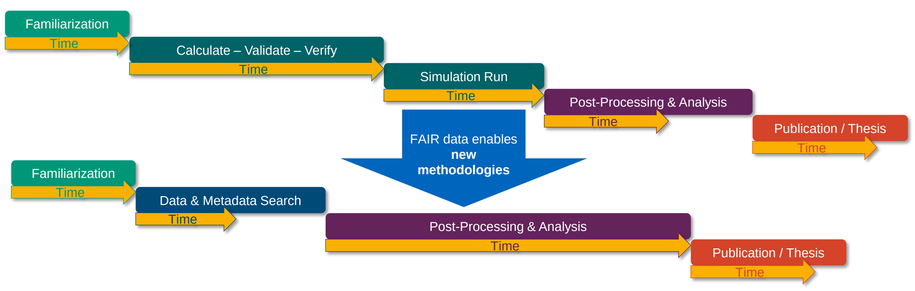
Analysis and processing of measurement data using HPC

High Performance Computing

- Script / code
- Input file, output file, log file
- Raw data
- Processed data
- Metadata
- Data for secondary research (e.g. energy consumption in HPC)

FAIR Data Principles in HPMC

- **F**indable: storage in personalized accounts, little metadata
- **A**ccessible: no access for third parties, insufficient transfer tools
- **I**nteroperable: depending on formats and enriched meta
- **R**eusable: computing time at HPC centres required or virtualization



Research Data Management at TUM

- **Munich Data Science Institute**: The MDSI is TUM's central interface and innovation hub for data science, machine learning and AI.
- **University Library**: The TUM Research Service Centre supports the entire life cycle of research projects and provides advice on all questions relating to research data management.
  → **mediaTUM**: The TUM **repository** supports the publication of digital documents and research data and is indexed by many third-party services.
  → **TUM Workbench**: The web-based platform is a tool for an integrated **project and research data management.** It offers numerous functions, e.g. electronic lab books and **data management plans**. Research data can be uploaded, structured, shared and provided with **metadata**.
- **Leibniz Supercomputing Centre**: LRZ engages in several projects, systems and services to support research data management, e.g.
  → **(TUM) Data Science Storage**: A dropbox-like file system that can be shared amongst the LRZ / TUM ecosystem including external access. TUM has purchased a dedicated DSS building block. Research projects from most TUM faculties are eligible to apply for this storage space.
  → **Globus Online**: A (non-profit) platform as a service for moving, sharing, publishing and discovering data. SuperMUC-NG and DSS can be accessed, data can be moved between GridFTP servers or endpoints.

What Are We Trying to Accomplish?

Metadata Standards and Support to Data-Generating Groups for HPMC

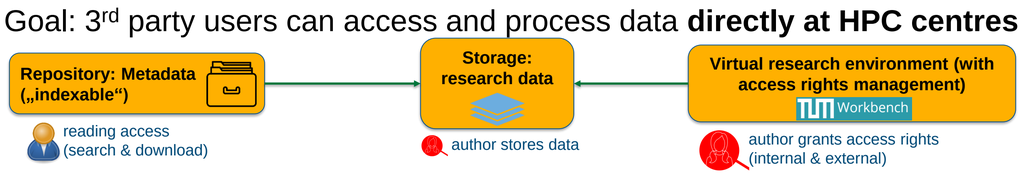
- Define and disseminate **metadata standards** for HPMC environments
- Transposal of metadata standards into an **semantic HMPC-sub-ontology** in the framework of the Metadata4Ing top-level-ontology
- Development and provision of a **metadata crawler** to read out ontologies, create a dictionary and fill in a metadata-file
- Provisioning of retrieved metadata through the MediaTUM repository and mapping with the NFDI4Ing **metadataHub**

Reproducibility and Reusability of HPMC data

- Foster the possibilities of **data (re-use) projects** at HPC centres or within multicloud projects
- Provide **virtual machine images** of HPMC research data per Compute Cloud server (LRZ) and provide reduced datasets (HLRS)
- Evaluate **container-virtualization** for re-use and reproducibility and prepare best-practice guidelines

Storage, Access and Transfer

- **Synchronization** of encryption, encrypted workflows and other integrity and access management mechanisms **within GCS centers**
- **Indexable** storage to front-end **interface** with **access management** for large data
- Goal: 3rd party users can access and process data **directly at HPC centres**



Support to the Community and 3rd Party-Users

- **Workshops** and lectures
- Policies, **guidelines** and templates (e.g. data management plan)

What Do We Have to Offer?

Find out more on our **website** with the following content:

- **Ph.D.-Workshop** on October 11: Research Data Management for Science and Engineering at TUM
- **Lecture**: Introduction to Research Data Management for Engineering Students (ED140003 / summer semester)
- **NFDI4Ing Conference** on October 26 and 27 (online)
- **ing.grid**: a data management **journal** to publish articles, datasets and software
- **Metadata4Ing**: an **ontology** for describing the generation of research data within a scientific activity
- **Data Management Plan**: a generic template with extension for high performance measurement and computing
- Software: a **crawler for automatic metadata generation**

Subscribe to our **newsletter** for events and updates:

Authors: Prof. Dr.-Ing. Christian Stemmer, Giuseppe Chiapparino, Benjamin Farnbacher, Nils Hoppe, Vasiliki Sdralia (Chair of Aerodynamics and Fluid Mechanics / TUM)
Contact: info-doris@nfdi4ing.de