Research Data Management for Data from High Performance Computing (Tier 1)

Best-Practice and Applications

2022/04/06, 13:00

HLRS High-Performance Computing Center | Stuttgart

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

lrz Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften

Participants

Technical University of Munich (TUM)
Prof. Dr.-Ing. Christian Stemmer
Benjamin Farnbacher
Nils Hoppe
Vasiliki Sdralia

High Performance Computing Center Stuttgart (HLRS)
Dr.-Ing. Thomas Bönisch
Nadiia Huskova
Volodymyr Kushnarenko

Jülich Supercomputing Centre (JSC)
Sander Apweiler

Leibniz Supercomputing Centre (LRZ)
Dr. Stephan Hachinger
Stephan Peinkofer

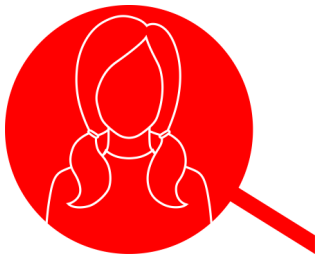
Further Information

Downloads

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

Contact

- **Newsletter**: https://lists.tu-darmstadt.de/mailman/listinfo/nfdi4ing_taskarea_doris
- **Mail**: info-doris@nfdi4ing.de (confirmation of participation needed?)
- **Web**: https://nfdi4ing.de/archetypes/doris/



Agenda

- Introduction
- Hardware Systems, Storage Systems, Transfer Tools (HLRS, JSC, LRZ)
- Research Data Management Basics
- Break
- Research Data Management on HPMC (further slides by HLRS, JSC, LRZ)
- Future Developments and Outlook (InHPC-DE slides by GCS)
- Feedback

NFDI & NFDI4Ing

German National Research Data Infrastructure (NFDI)

- 19 consortia (up to 30)

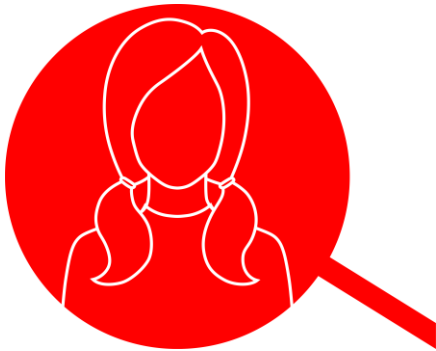
Consortium for Engineering Sciences – NFDI4Ing

- 14 “steering institutions”
- 30 participant institutions
- divided in 8 engineering archetypes
  - **Alex**: bespoke experiments with high variability of setups
  - **Betty**: engineering research software
  - **Caden**: provenance tracking of physical samples & data samples
  - **Doris**: high performance measurement & computation
  - **Ellen**: extensive and heterogeneous data requirements
  - **Frank**: many participants & simultaneous devices
  - **Golo**: field data & distributed systems

Archetype DORIS: HPMC

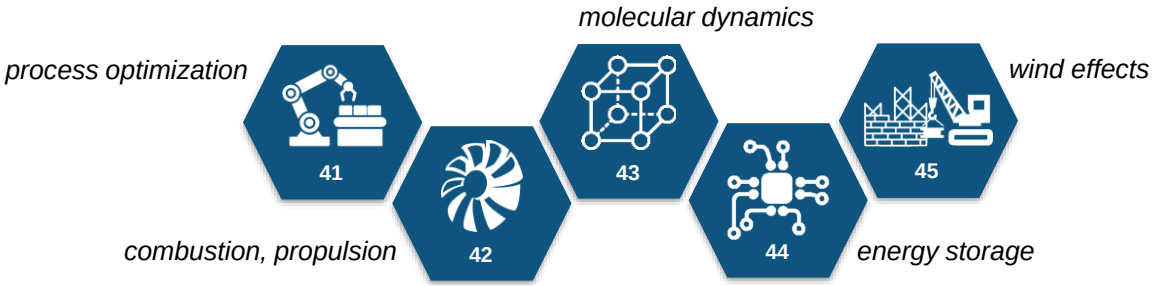
*… I'm an engineer conducting and post-processing high-resolution and **high-performance measurements and computation** (simulation) **with very large data** on HPC systems.*

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*

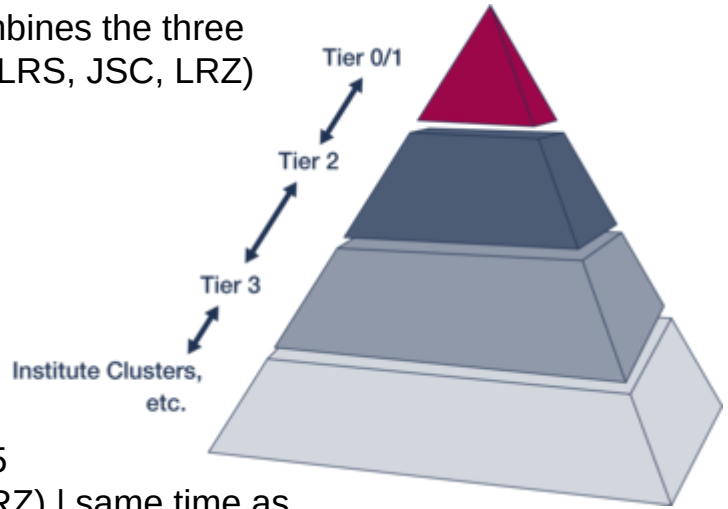


DORIS's patron is Christian Stemmer

High Performance Measurement and Computing (HPMC)

Focus on tier 0 (EU) / tier 1 (DE)

Gauss Centre for Supercomputing (GCS) combines the three national tier 1 / tier 0 supercomputing centres (HLRS, JSC, LRZ)

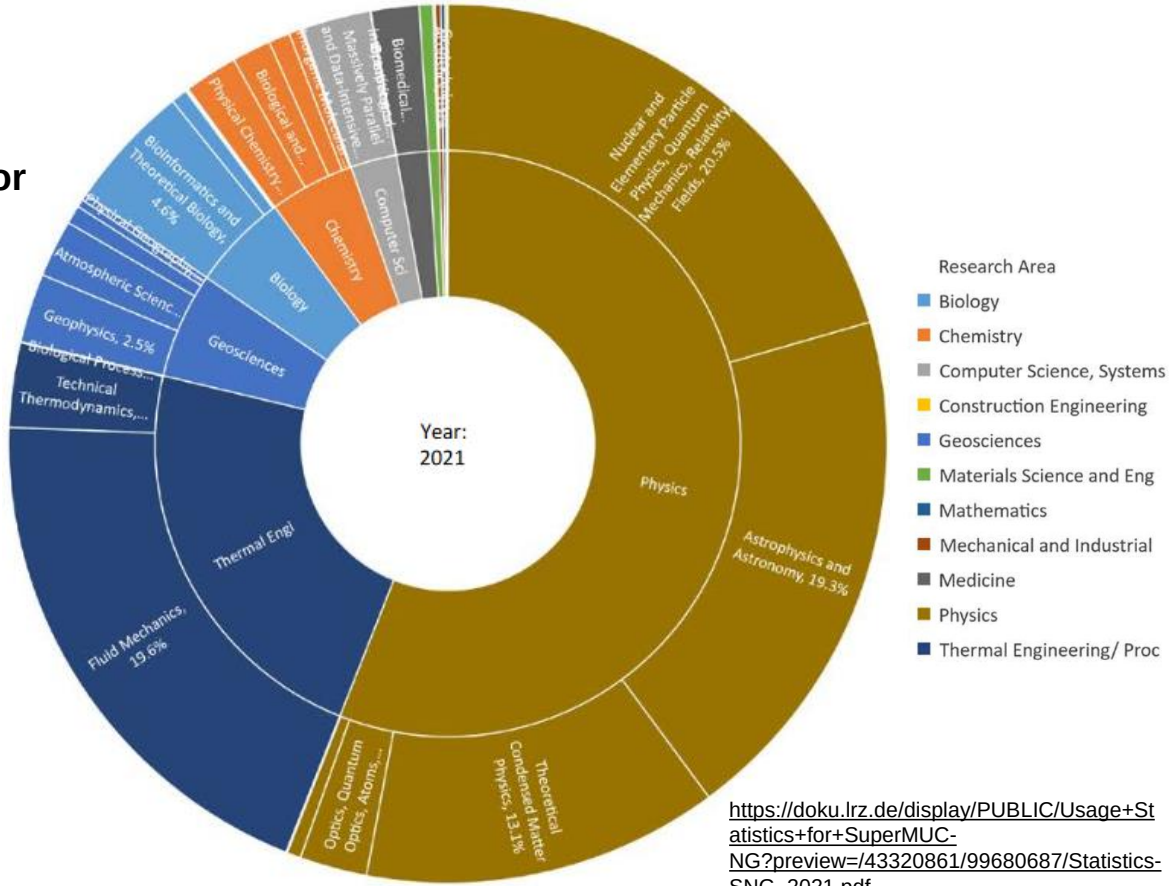


- Project proposal via GCS required
- Test project: ~300k core hours
  - rolling call, simplified application
- Regular project: ≤ 35 mio core hours
- Large scale project: > 35 mio core hours
  - Peer reviewed
  - Large scale calls: July 11 to August 15
  - Regular calls: rolling calls (HLRS & LRZ) | same time as large scale projects (GCS)

- Further information (link) | application (link)

HPMC

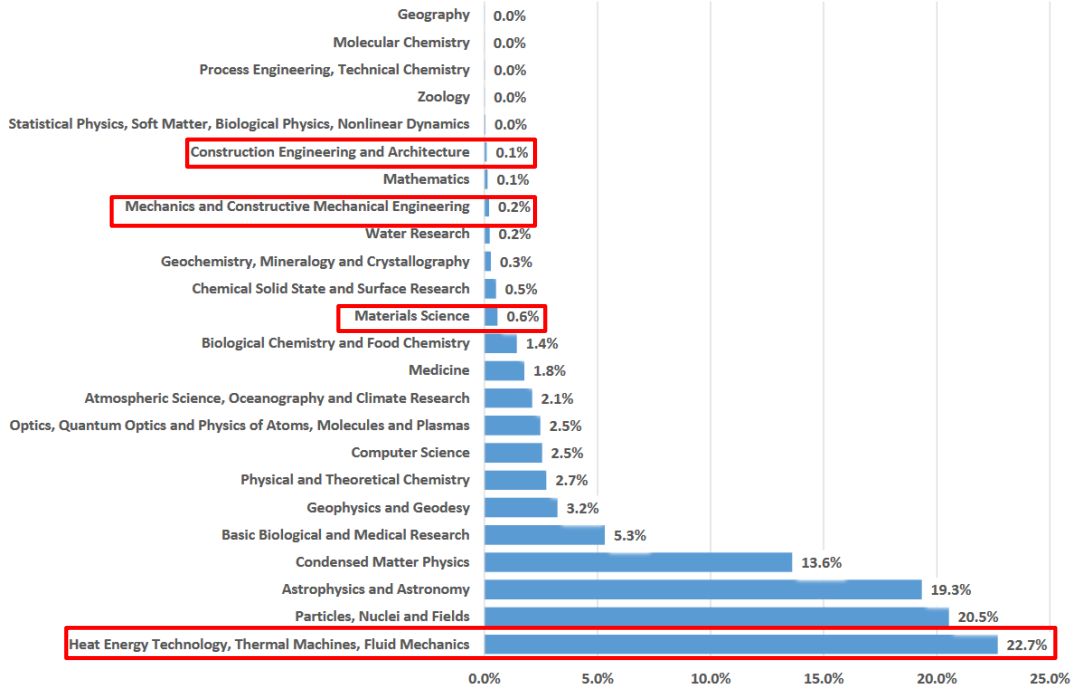
Usage Statistics for SuperMUC-NG



https://doku.lrz.de/display/PUBLIC/Usage+Statistics+for+SuperMUC-NG?preview=/43320861/99680687/Statistics-SNG_2021.pdf

High Performance Measurement and Computing (HPMC)

Usage Statistics for SuperMUC-NG



https://doku.lrz.de/display/PUBLIC/Usage+Statistics+for+SuperMUC-NG?preview=/43320861/99680687/Statistics-SNG_2021.pdf

Poll 1: HPC systems

1. Have you already worked on tier 0 HPC systems? (Einzelne Wahl) *

22/22 (100%) haben geantwortet

Yes	(14/22) 64%
No	(8/22) 36%

2. Do you have plans to work on a tier 0 HPC system in the future? (Einzelne Wahl) *

22/22 (100%) haben geantwortet

Yes	(16/22) 73%
No	(3/22) 14%
Unsure	(3/22) 14%

3. On which (tier 0) HPC centre (Einzelne Wahl) *

22/22 (100%) haben geantwortet

HLRS	(6/22) 27%
JSC	(6/22) 27%
LRZ	(5/22) 23%
Other	(5/22) 23%

4. Are you working or planning to work on tier 2 HPC systems (NHR etc.)? (Einzelne Wahl) *

22/22 (100%) haben geantwortet

Yes	(1/22) 5%
No	(10/22) 45%

HPMC Research Data

What are HPMC Research Data

Research data are data that are created during a research process or are the result of it

High Performance Measurement

- Measurement data
- Metadata (hardware, method 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 (software, hardware, method etc.)
- Data for secondary research (e.g. energy consumption or temperature in HPC)

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, input file, raw 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 meta 😐

Reusable: computing time at HPC centres required or virtualization (e.g. container) 😐

HPMC Research Data

Why research data management for HPMC-Data?

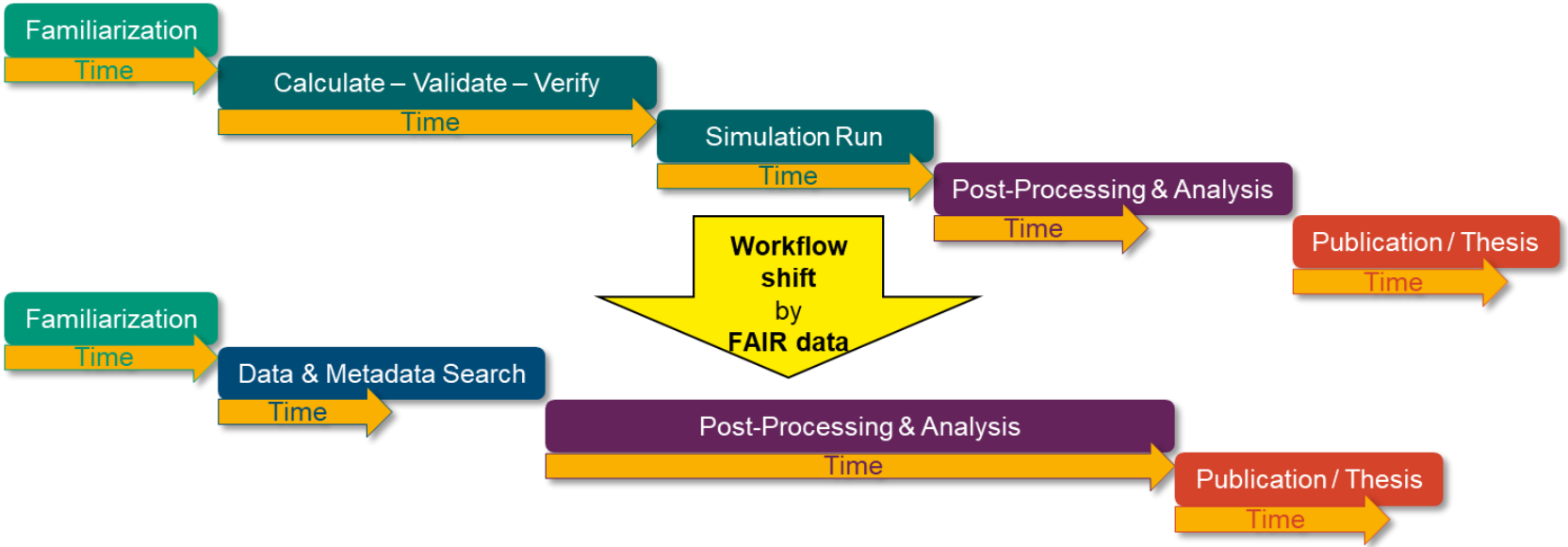
- Scientific integrity and fulfilment of (external) compliance (e.g. DFG)
- Secondary research (e.g. energy consumption or temperature in HPC centres)
- New findings, new methodologies, new workflows, new opportunities by re-using existing data

HPMC Research Data

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



Workflow shift as alternative / not as replacement

Poll 2: HPMC research data

1. Are you providing research data to other researchers? (Mehrfachauswahl) *

23/23 (100%) haben geantwortet

Yes: intern	(16/23) 70%
Yes: extern	(7/23) 30%
No	(5/23) 22%

2. Are you providing HPMC research data to other researchers? (Mehrfachauswahl) *

23/23 (100%) haben geantwortet

Yes: intern	(6/23) 26%
Yes: extern	(0/23) 0%
No	(17/23) 74%

3. Have you already used others HPMC data? (Einzelne Wahl) *

23/23 (100%) haben geantwortet

Yes	(5/23) 22%
No	(18/23) 78%

4. What is your main motivation for FAIR research data management? (Einzelne Wahl) *

23/23 (100%) haben geantwortet

Scientific integrity and compliance	(9/23) 39%
Secondary research	(4/23) 17%
New findings, new methodologies, new workflows	(10/23) 43%

RDM: Benefits

Extrinsic factors

- Compliance with Good Scientific Practice Principles (e.g. DFG Code of Conduct)
- Compliance with internal guidelines

- Required for access to certain funding streams (e.g. ERC Horizon Europe, from 2024 BMBW/BMBI)
- Increasing political importance (e.g. Federal Data Strategy, NFDI)

- Simplifies re-use by a third party
- Enables secondary research, new findings / methodologies based on FAIR data

- RDM also applies for data from industry projects and proprietary data
  - ➢ E.g. access management or deletion deadlines can be controlled by RDM

RDM: Benefits

Intrinsic factors

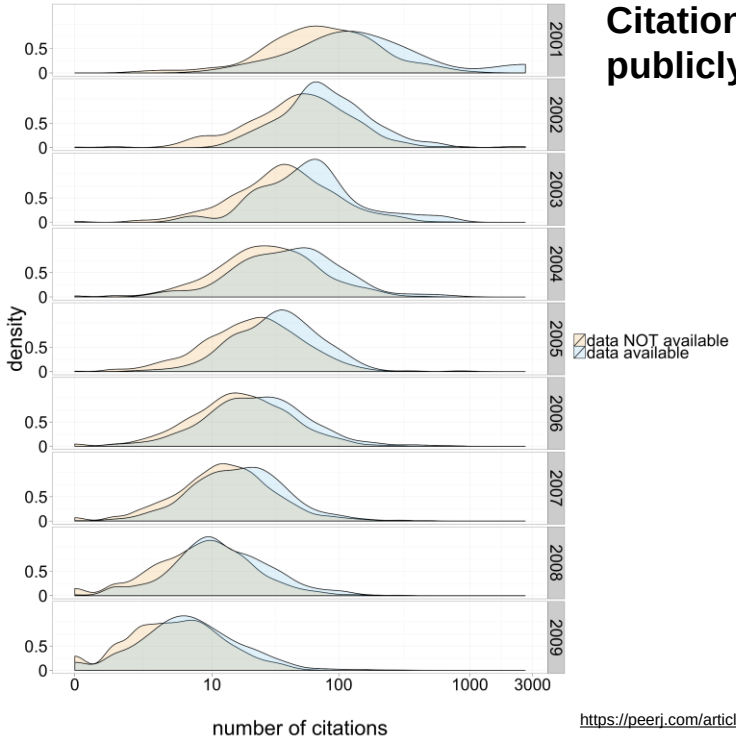
- Simplifies re-use by yourself or your group
- Scientific reputation and transparency
- Visibility and improved odds for collaborations and funding
- New publishing opportunities (e.g. peer reviewed data publishing or data based PhD)
  - Journal “Data in Brief” (Elsevier)
  - Tbd: Journal “ing.grid” (NFDI4Ing)

- Improved project management through RDM
- Minimizes risk of data loss

- Re-use or dissemination of proper data
- „Standing on the shoulder of giants“
  - New findings through (meta-)data analysis
  - Verification and validation of proper models by external data

- Increase of citations by published research data
  (Publications: 2013, 2016, 2020)

RDM: Benefits



Citation density for papers with and without publicly available data (Piwowar & Vision 2013)

https://peerj.com/articles/175/ (04.05.2021)

RDM: Benefits

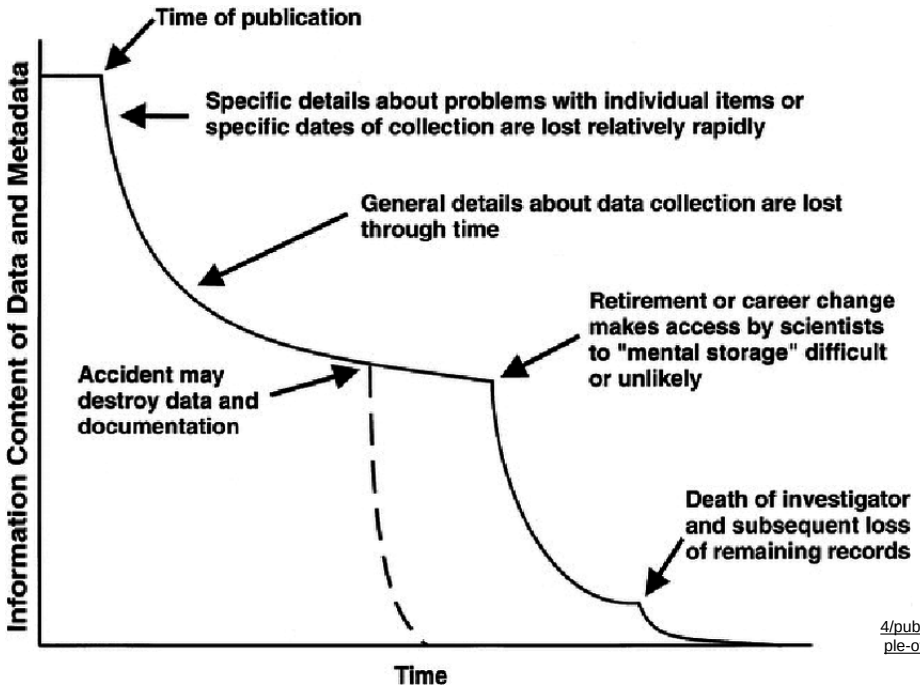
Sharing data increases citations (Drachen et al. 2016)

Journal papers published in 2010	Astro-physical Journal (ApJ)	Astronomy and Astrophysics (A&A)	Astronomical Journal (AJ)
All			
# of papers	2501	1918	388
# of citations	74,663	40,829	9465
Mean citations/paper	29.9	21.3	24.4
Datalink papers			
# of papers	794	875	174
# of citations	27,936	22,308	4754
Mean citations/paper	35.2	25.5	27.3
No datalink papers			
# of papers	1707	1043	214
# of citations	46,727	18,521	4711
Mean citations/paper	27.4	17.8	22.0

https://www.liberquarterly.eu/article/10.18352/lq.10149/ (04.05.2021)

RDM: Benefits

The loss of information about data over time (Michener et al. 1997)



https://www.researchgate.net/profile/Adam-Wilson-4/publication/255571027/figure/fig4/AS:297949465726983@1448048096877/Example-of-the-normal-degradation-in-information-content-associated-with-data-and.png (04.05.2021)

RDM: Basics

Policies and Guidelines

- Regulations by funding agencies, research organizations, disciplines, journals, publishing houses and universities
- German Research Foundation (DFG): Guidelines on Handling Research Data (2015)
- German Research Foundation (DFG): Guidelines on Safeguarding Good Research Practice – Code of Conduct (2019)
- European Commission: Horizon (2020)
- Summary of German funding organizations: https://www.forschungsdaten.info/funder-guidelines/

RDM: Basics

Data Management Plan (DMP)

- Describes data generation, storage, access, backup, sharing etc.
- Is a **prerequisite** for access to most funding streams
- Is a living document and should be updated

Funder	Submission	Contents	Update required?
HORIZON 2020	deliverable in month 6	Horizon 2020 Programme Guidelines	month 18, final version at the end of the project
DFG Deutsche Forschungsgemeinschaft	not specified	specific guidelines not yet available	yes, regularly
Bundesministerium für Bildung und Forschung	in the proposal	Guidelines on the Handling of Research Data	No
VolkswagenStiftung	with proposal if applicable	depends on project and funding programme	depends on project and funding programme
Horizon Europe	with proposal	Contents of the Science Europe Practical Guide	no

RDM: Basics

Data Management Plan (DMP) for HPMC: Tools & Templates

- **HPMC-template:** https://zenodo.org/record/5801838#.YjSN0DUxmUk
  - Basic template (~ 1h required)
  - Feedback?
- RDMO: https://rdmorganiser.github.io/en
  - Web application developed in a DFG project,
  - discipline-specific DMP templates available,
  - Includes (data) project management applications
- List of tools / DMP-Toolguide: https://zenodo.org/record/4632308#.YkV4ONNBw2w
- LIBER DMP CATALOGUE: https://zenodo.org/communities/liber-dmp-cat?page=1&size=20
  - Collection by Europe's research library community (TUM, MPI etc.)
- Short DMP checklist by the DDC (min. requirement):
  - www.dcc.ac.uk/sites/default/files/documents/resource/DMP/DMP_Checklist_2013.pdf

FAIR Data Principles in HPMC

FAIR Principles are the framework for sustainable research data management

- **F**indable
- **A**ccessible
- **I**nteroperable
- **R**eusable

Wilkinson et al.: The FAIR Guiding Principles for scientific data management and stewardship; *Scientific Data* volume 3, Article number: 160018 (2016); https://doi.org/10.1038/sdata.2016.18

FAIR ≠ OPEN (not all FAIR data have to be open and not all open data are automatically FAIR)

EUDAT **FAIRness checklist**: https://zenodo.org/record/1065991#.YkV60tNBw2w

FAIR: Findability

The F in FAIR (Findable) - Conditions

- Data are published in a suitable **repository**
- Data are equipped with persistent **identifiers** (e.g. DOI)
- Metadata are enriched with readable **metadata**
  - Descriptive, administrative, structural
  - **Metadata are crucial for findability in repositories and databases and to enable search-engine indexing**

FAIR: Findability

Persistent Identifiers DOI (Digital Object Identifier)

- Most common for data: **DOI** (Digital Object Identifier)
- Permanent (persistent), digital link, consisting of numbers or strings of alpha-numerical signs, e.g.: 10.5281/zenodo.6375058
- DOI resolver: https://dx.doi.org/
- Makes published data **citable**
- Published data with DOI gets “cold”
  - No more changes allowed
  - Exception: Version DOI / Concept DOI (e.g. Zenodo http://help.zenodo.org/#versioning)
- Most common identifier for people: **OrcID**
- GitLab, GitHub etc. / software repositories: citation file (e.g. CITATION.cff) with Version DOI

GitLab

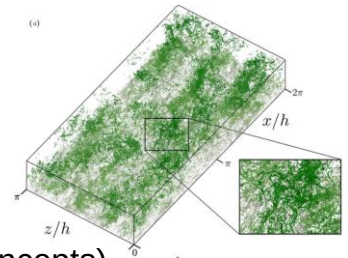
FAIR: Findability

Repositories

- Database systems to store and publish digital objects
  - Ideally offer access control and transparent user rights
  - Ensure visibility of data (via use of PID and meta data)
- Repository finder: https://www.re3data.org

State of the Art HPMC

- Findable HPMC data in repositories do exist
  - Mostly extracted / processed data
  - Replication data / input files, job files, scripts etc.
  - Little “real” raw data
- Raw data is often published on personal / group websites
  - E.g. https://torroja.dmt.upm.es/turbdata/ (UPM)
  - Findable? DOI? Metadata? (repository is older than most RDM concepts)



FAIR: Findability

Repositories

Best-Practice Examples

- **The Dataverse Project**
  - https://dataverse.org/ (about)
  - https://dataverse.harvard.edu/ (search / publish)

- Open source software (Harvard)
- 78 Installations (mainly connected)
- 66 results for HPC
- Commonly replication data / rarely raw data
- Mostly “small” data (KB <> GB)

FAIR: Findability

Repositories

Best-Practice Examples

- **DaRUS (University of Stuttgart)** https://darus.uni-stuttgart.de/
  - Based on Dataverse
  - Max. ~100 GB, association to Uni Stuttgart project required
  - Access management: “hot” data and citable publication

Metadata Blocks and link to controlled vocabulary (has to be linked manually)

- 3 Results for “HPC” (replication data)
- No explicit indexing of HPC / HPMC etc. / authors have to tag keywords
- Taxonomy due to research domain and not the tool / method

FAIR: Findability

Repositories

Best-Practice Examples

- B2FIND / EUDAT: http://b2find.eudat.eu/
  - discovery service harvested from EUDAT data centres and community repositories
  - 117 datasets for HPC (keyword)
  - Members (e.g.): FZ Jülich, Max Planck Computing Facility, KIT

FAIR: Findability

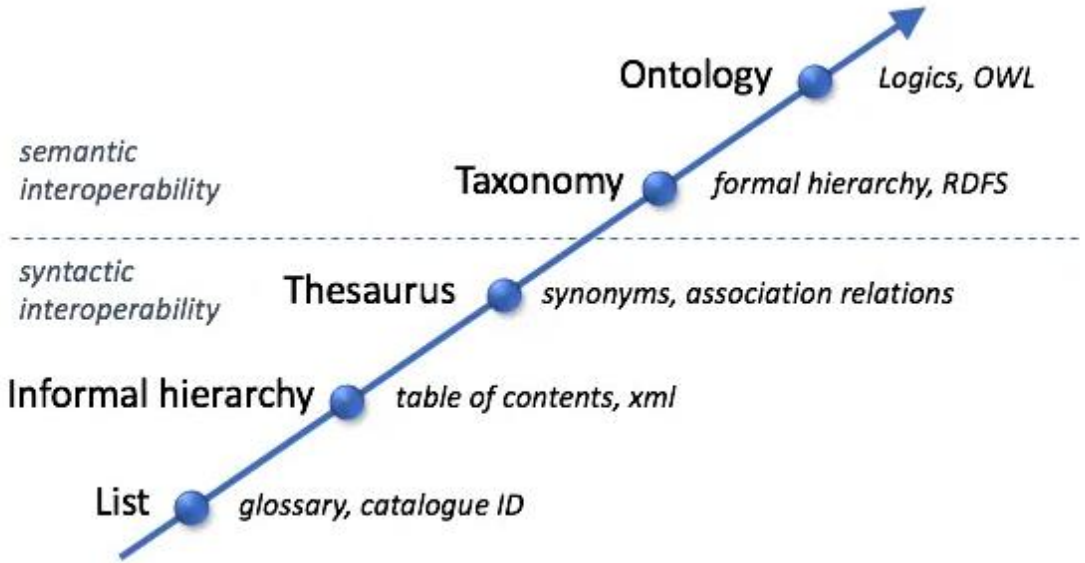
Repositories

Best-Practice Examples

Repository	Host	Data Range	Feature
https://zenodo.org/	CERN (EU)	50 GB per file	generic repository, DOI, GitHub integration
http://turbulence.pha.jhu.edu/	Johns Hopkins University	multi-Terabyte	(discipline) turbulence database, publishing only for associates
https://mediatum.ub.tum.de/	Technical University of Munich	multi-Terabyte	institutional repository, publishing only for TUM members, connection to LRZ storage (DSS)
https://coscine.rwth-aachen.de/	RWTH Aachen University	100 GB per project	generic data management tool (pilot)

FAIR: Findability

Metadata



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

FAIR: Findability

Metadata

Best Practice

- Usage of standardized schemes in human and machine readable formats

Scheme	URL	Purpose / Domain
NASA Thesaurus	https://www.sti.nasa.gov/docs/thesaurus/thesaurus-vol-1.pdf	Aerospace
PhySH	https://physh.org/browse	Physics
DataCite	https://schema.datacite.org/meta/kernel-4.4/doc/DataCite-MetadataKernel_v4.4.pdf	Publication and citation of research data / research outputs
CodeMeta	https://codemeta.github.io/terms/	Software
DCAT	https://www.w3.org/TR/vocab-dcat-3/	Description of data sets
EngMeta	https://darus.uni-stuttgart.de/file.xhtml?persistentId=doi:10.18419/darus-500/3&version=1.0	Metadata scheme for engineering science
Metadata4Ing	https://nfdi4ing.pages.rwth-aachen.de/metadata4ing/metadata4ing/index.html	Ontology for engineering science

FAIR: Findability

Metadata

Best Practice

- No established metadata scheme for HPMC
  - ➢ NFDI4Ing measure: development of a HPMC sub-ontology
- NFDI4Ing Terminology Service (terminology search): https://terminology.nfdi4ing.de/ts4ing/ontologies
- Ontobee ontology server: https://www.ontobee.org/
- Dublin Core metadata design: https://www.dublincore.org/resources/userguide/

Poll 3: Metadata

1. Are you applying standardized metadata schemes within your research? Which schemes? (Lange Antwort)

8/8 (100%) Beantwortet

- MODA/OSMO; other VIMMP ontologies; EMMO; PIMS-II; other EMMO-compliant ontologies; CHADA; m4i
- No
- Wir entwickeln Ontologien
- EngMeta is known.
- no
- yes, as long as it is possible to do so
- No, because in my domain there is no standardized metadata scheme
- DataCite

2. How would you search for HPMC data (buzzwords? terminology?)? (Lange Antwort)

8/8 (100%) Beantwortet

- SPARQL queries
- ??
- terminology
- Using keywords and terminology
- Buzzwords, similar to a generic internet search
- HPC, cluster, simulation, modelling, ...
- terminology
- by keywords

3. Which metadata is the most important to catagorize HPMC data? (Domain metadata? HPC system metadata? Process / workflow metadata? etc.) (Lange Antwort)

8/8 (100%) Beantwortet

- Domain metadata and bibliographic
- process / workflow
- HPC-center name, ...
- domain metadata
- relation to another publication/paper
- Domain and workflow data

FAIR: Accessibility

The A in FAIR (Accessible) - Conditions

- Infrastructure for **long term storage** and **data transfer**
- Access models
  - Data sovereignty by **access rights management**
  - Internal vs. external access / community vs. open access
  - Data security, data privacy and copyright policies
- Low software and hardware thresholds
- Open standard protocols for authorized users

FAIR: Accessibility

The A in FAIR (Accessible)

Problems

- Storage in personalized accounts at computing centres (esp. raw data)
- (Institutional) isolated applications
- Data usually has to be transferred / can't be processed directly at computing centre
- Access rights management and transfer possibilities depends on individuals / individual configurations

Best-Practice

→ Usage of repositories / data containers with access rights management

- E.g.: Data Science Storage (LRZ)

→ Usage of transfer tools with rights management

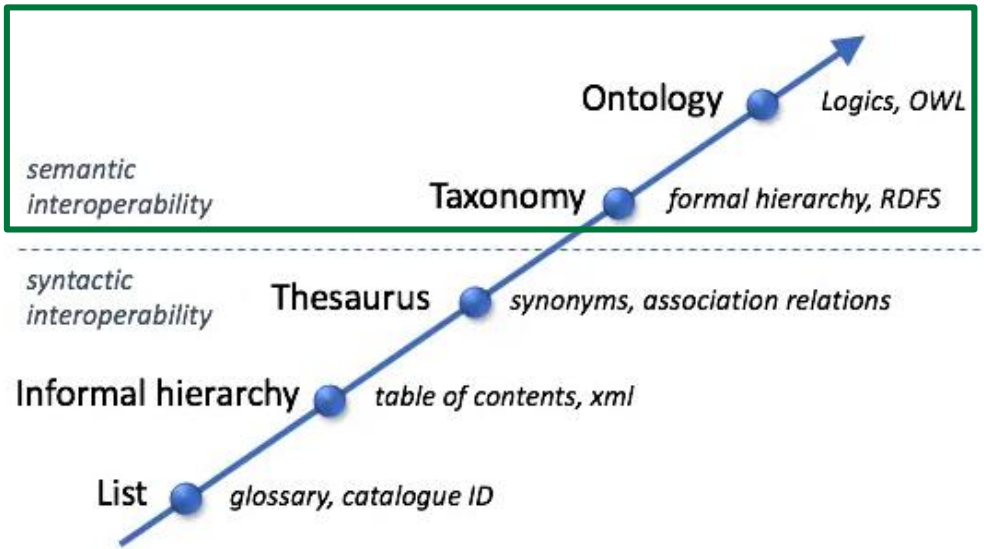
- E.g.: Globus Online (LRZ), GridFTP (HLRS, DKRZ)

→ Usage of data management tools with rights management

- E.g.: Coscine (RWTH, pilot phase): Login with DFN-AAI or ORCID, integrative, max. 100 GB / project

FAIR: Interoperability

The I in FAIR (Interoperable)



→ **Metadata scheme**
→ Perspective: machine learning

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

FAIR: Interoperability

The I in FAIR (Interoperable)

Best-Practice

- Optimize technical compatibility
  - Naming conventions (no special characters, capitals etc.)
  - Standardized, unencrypted non proprietary data formats
- Self created data formats need to be accompanied by
  - documentation
  - software
  - original computing environment

FAIR: Interoperability

The I in FAIR (Interoperable)

Text	PDF/A *no format:* TXT *editable:* ODT, RTF, HTML *formulas:* LaTeX (TEX)
Table	CSV, TSV **HDF5 (*numerical data)***
Visualizations	*Raster:* PNG, TIFF *Vector:* SVG, EPS
Multimedia	*Container:* MKV, WebM, OGG *Video-Codec:* AV1, VP9 *Audio-Codec:* FLAC, WAV, Vorbis, Opus
Data base	SIARD, Dump, XML
structured data	**XML, JSON, YAML (metadata)**

https://www.forschungsdaten.info/themen/bewahren-und-nachnutzen/formate-erhalten/

FAIR: Reusability

The R in FAIR (Reusable)

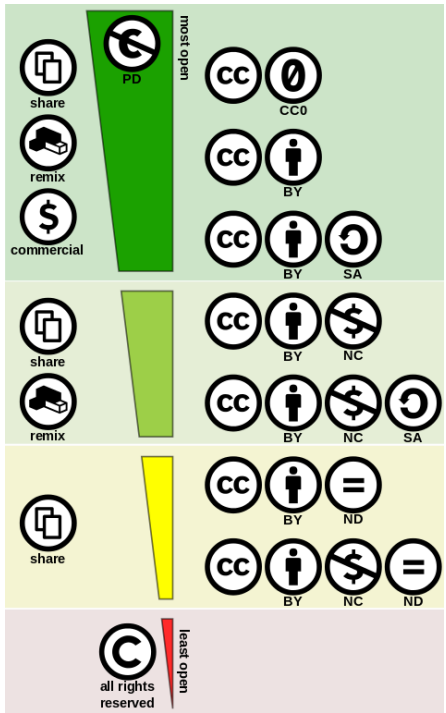
- **F**, **A** & **I** are required for **R**
- **HPMC specialties**
  - No easy and direct reuse at computing centres
  - Successful proposal for **computing time** at HPC centre required
  - **Transfer** to other computing centres is complex and time consuming
  - Vision: Data projects for direct reuse or processing of third party data

HPMC data reuse exists: https://iopscience.iop.org/issue/1742-6596/1522/1 (UPM)

- **Container virtualization** on HPC systems

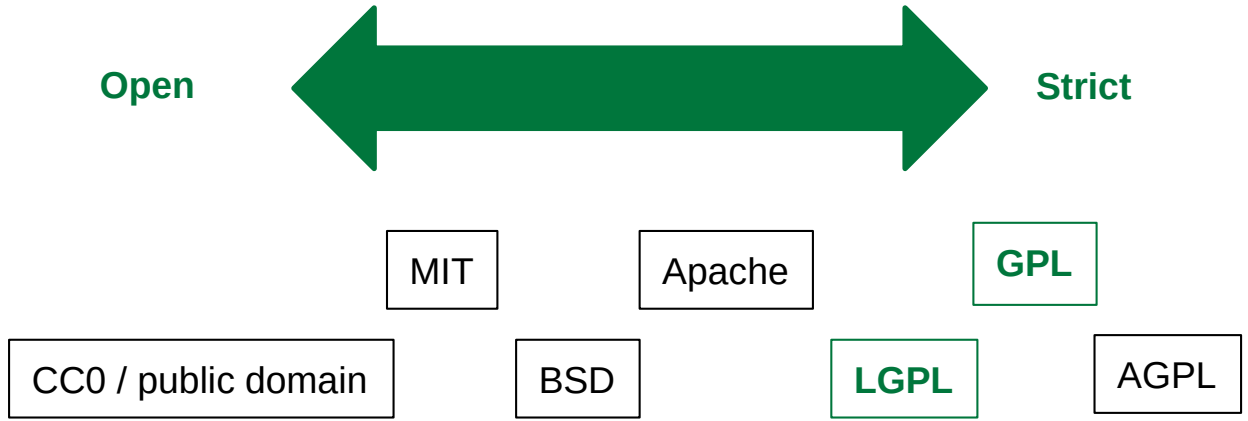
FAIR: Reusability

Licenses – Creative Commons



FAIR: Reusability

Licenses – Software



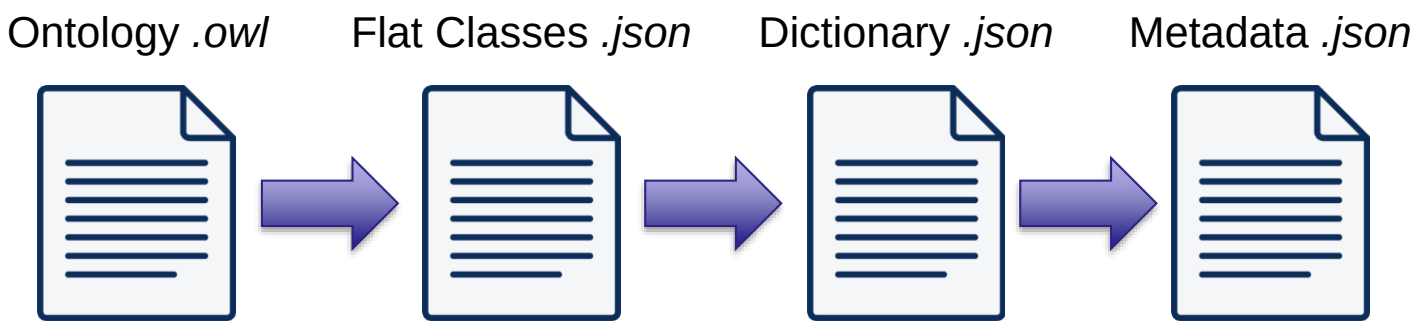
Weitere Informationen:
https://choosealicense.com
https://www.ifross.org/lizenz-center

Automated Metadata Generation

Metadaten-Crawler

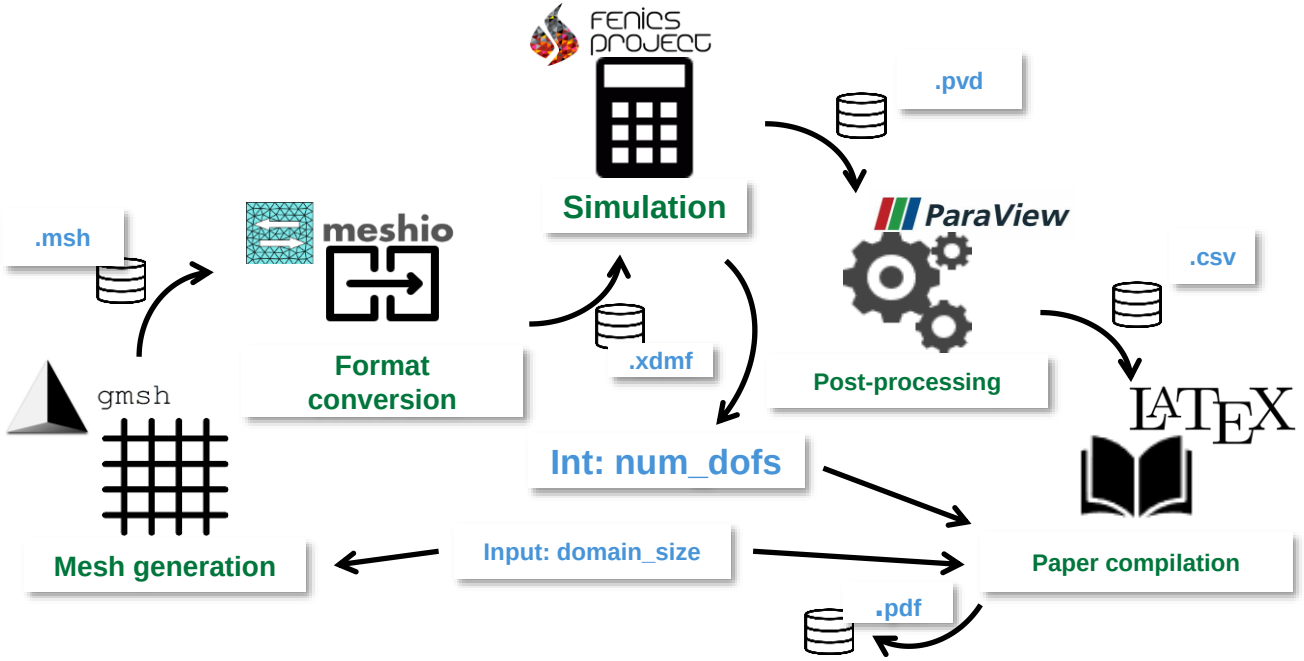
GitLab

`gitlab.lrz.de/nfdi4ing/crawler`



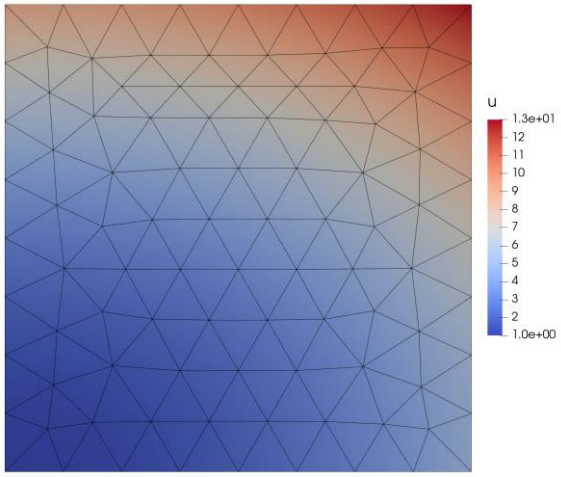
Automated Metadata Generation

An exemplary workflow



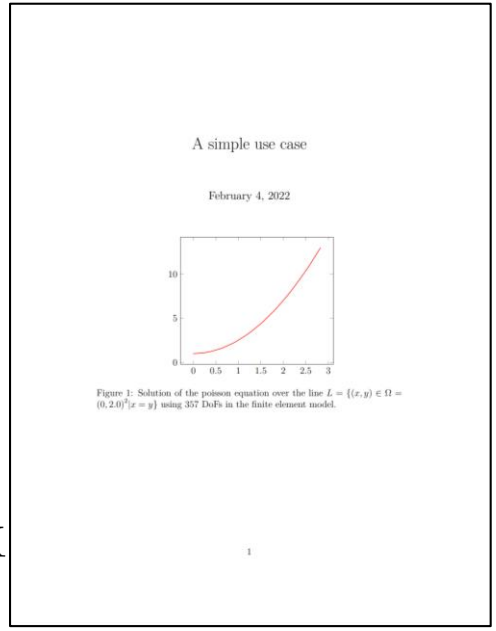
Automated Metadata Generation

An exemplary workflow



FENICS PROJECT

LATEX

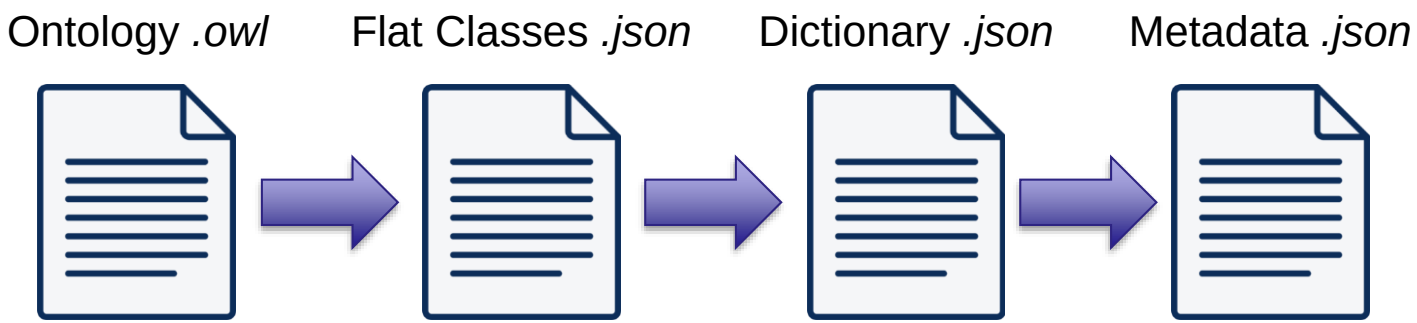


Automated Metadata Generation

Metadaten-Crawler

GitLab

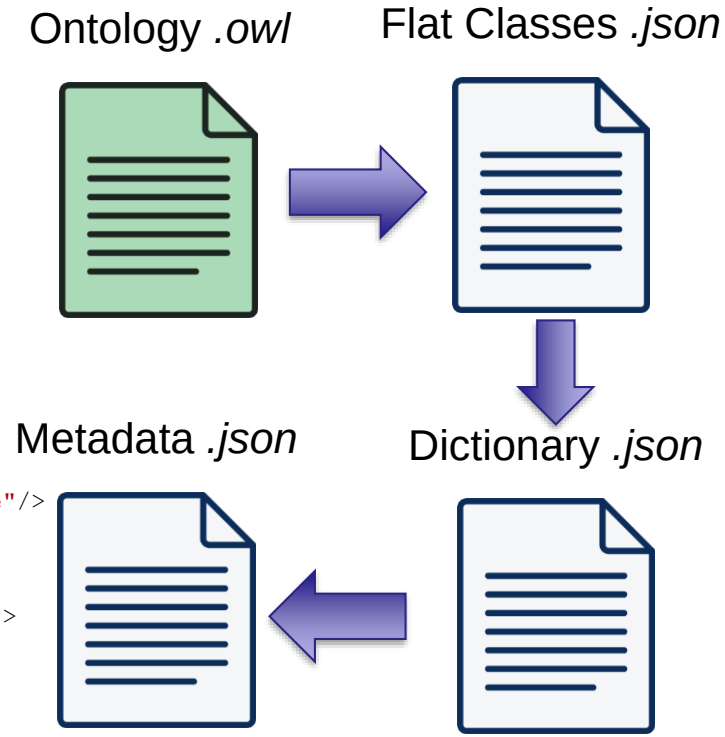
`gitlab.lrz.de/nfdi4ing/crawler`



Automated Metadata Generation

Metadaten-Crawler

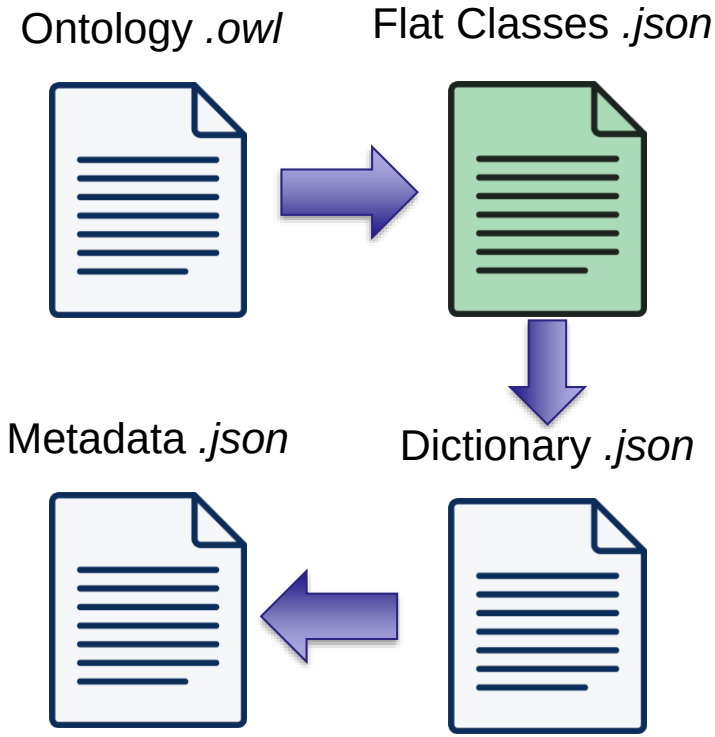
```
…
<Declaration>
   <Class IRI="ProcessingStep"/>
</Declaration>
<ObjectPropertyDomain>
   <ObjectProperty IRI="hasEmployedTool"/>
   <Class IRI="ProcessingStep"/>
</ObjectPropertyDomain>
<ObjectPropertyDomain>
   <ObjectProperty IRI="investigates"/>
   <Class IRI="ProcessingStep"/>
</ObjectPropertyDomain>
<DataPropertyDomain>
   <DataProperty abbreviatedIRI="schema:startTime"/>
   <Class IRI="ProcessingStep"/>
</DataPropertyDomain>
<DataPropertyDomain>
   <DataProperty abbreviatedIRI="schema:endTime"/>
   <Class IRI="ProcessingStep"/>
</DataPropertyDomain>
…
```



Automated Metadata Generation

Metadaten-Crawler

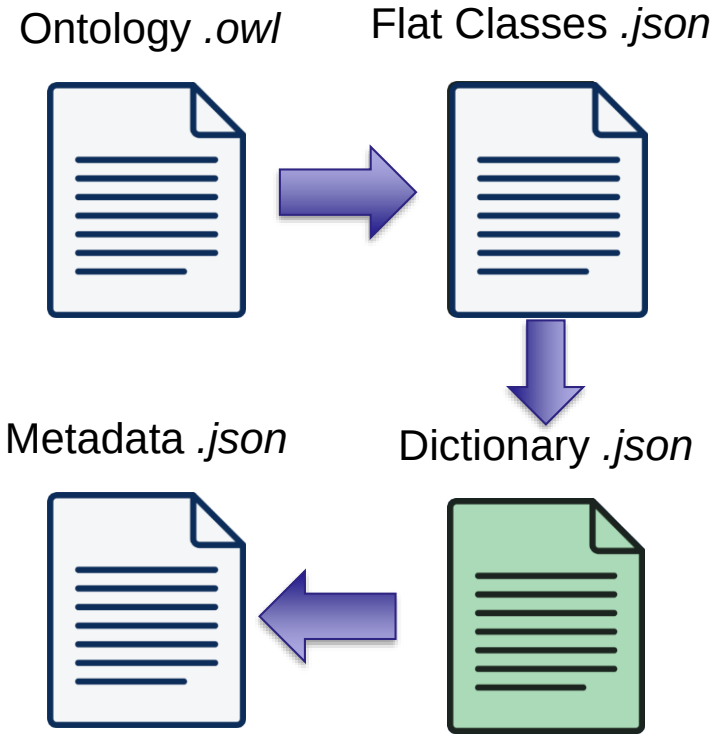
```
…
"ProcessingStep": {
  "__count__": 3,
   "__restrictions__": [],
   "hasEmployedTool": [2,5,3],
   "investigates": [2,1,1],
   "startTime": [1,1,1],
   "endTime": [1,1,1],
},
…
```



Automated Metadata Generation

Metadaten-Crawler

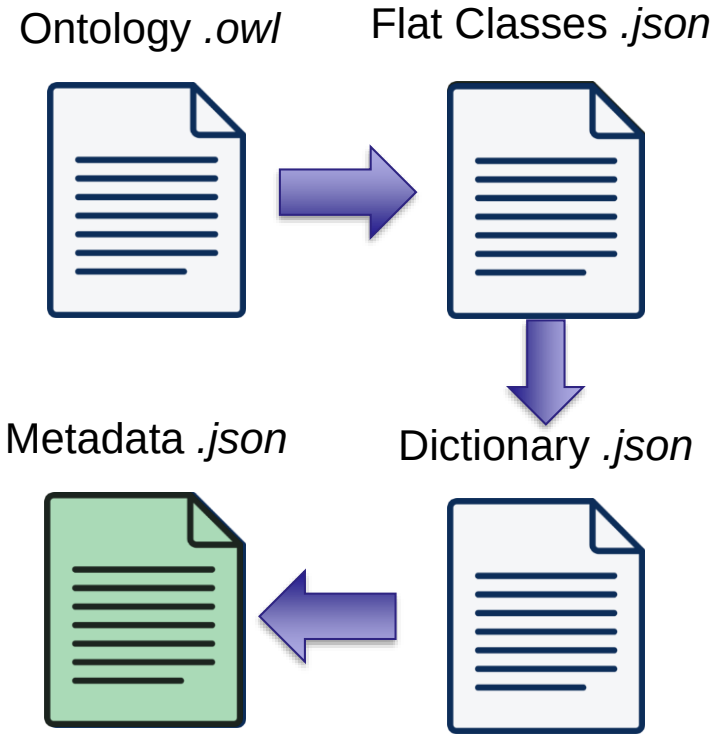
```
"ProcessingStep_1": {
    "hasEmployedTool_1": {
            "type": "regex",
        "path": "mesh_generator.log",
        "pattern": "^(.*)",
    },
    "hasEmployedTool_2": {
            "type": "os",
        "path": "",
        "pattern": "paraview --version",
    },
    "endTime": { … },
    "startTime": { … },
    "investigates_1": { … },
    "investigates_1": { … }
},
"ProcessingStep_2": {
    …
}
```



Automated Metadata Generation

Metadaten-Crawler

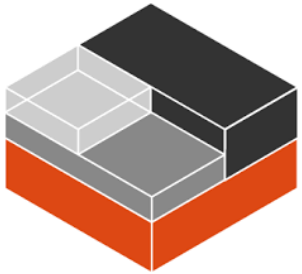
```
"ProcessingStep_1": {
   "hasEmployedTool_1":
      "gmsh V. 1.3",
   "hasEmployedTool_2":
      "ParaView V.5",
   "endTime":
      "2022/02/31 25:19 GMT",
   "startTime":
      "2022/02/31 25:12 GMT",
   "investigates_1":
      "Mesh Quality",
   "investigates_1":
      "Cell Sizes"
},
…
```

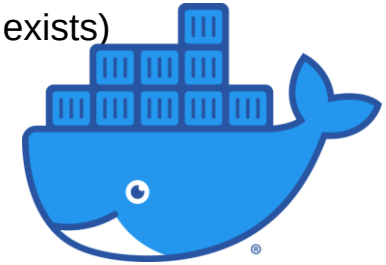


Virtualization: Container

Status Quo: General

- Dominator: Docker (default: Privileged user. Rootless version exists)
  - Alternatives: Podman, LXC/LXD, …





- New forms of Containers for HPC
  - Singularity/Apptainer
  - Charlielcoud
  - …

Virtualization: Container

Status Quo: Supercomputing Centers

	HLRS		lrz
Container	Singularity	Singularity	Charliecloud
Scheduler	PBS	Slurm	Slurm
Build on Ressource	No	(Yes)	No
Modules	Lmod	Easybuild	Spack

Virtualization: Container

Status Quo: Supercomputing Centers

	HLRS		lrz
Vanilla	✓	✓	✓

- Work directly on target machine.
  - ‘Trivial’
  - Outdated on hard- and/or software updates

Virtualization: Container

Status Quo: Supercomputing Centers

	HLRS		lrz
Bring your own container	✗	✗	?

- Create container ‘at home’, transfer and use on target machine
  - Usual container-type workflow
  - Problem: Drivers/Fabrics (e.g. Libfabric at HLRS, MOFED Mellanox on JSC), Performance for unmatching environments in/outside of container (scheduler, ABI-compatibility)
  - Vision: Compatible (performant) container base layer provided by compute centers. Better driver/fabrics support

Virtualization: Container

Status Quo: Supercomputing Centers

	HLRS	JSC	LRZ
Mirror	?	?	?

- ‘Copy’ Host environment into container
  - Manual re-creation worked on LRZ, not tried ad JSC, HLRS
    - Very time consuming
  - Possibilities to achieve this in an automated fashion?
    - In theory yes, but …

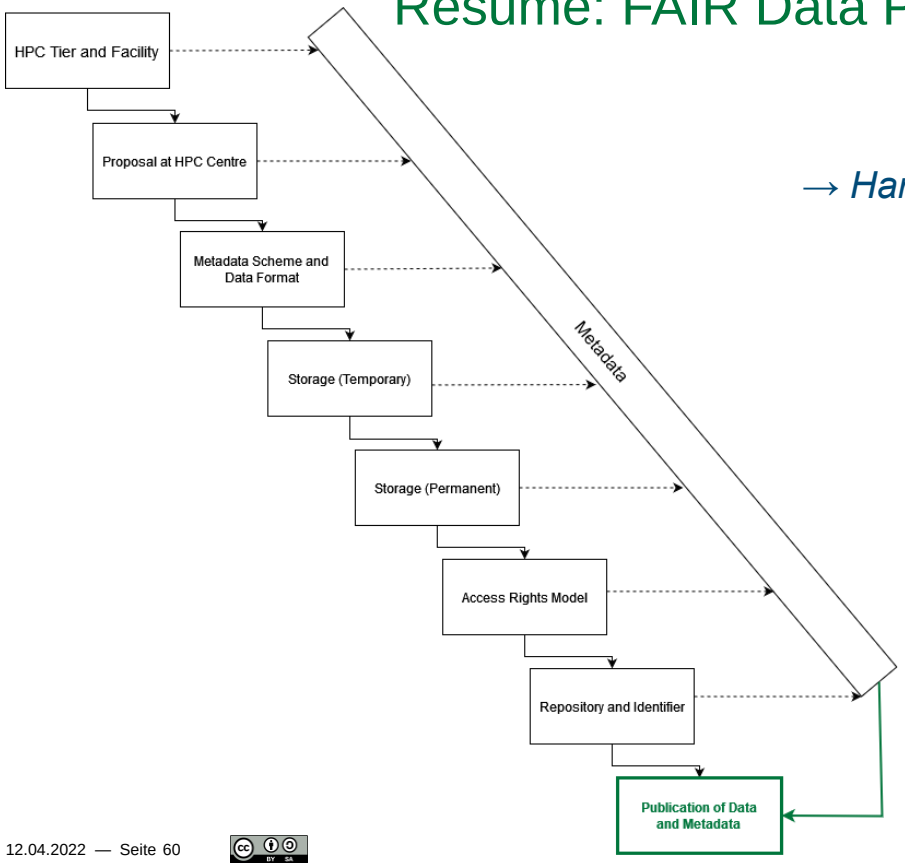
- Vision: Obsolete approach if “Bring your own” vision fulfilled.

Virtualization: Container

Status Quo: Supercomputing Centers

	HLRS		lrz
Vanilla	✓	✓	✓
Bring your own	✗	✗	?
Mirror host system	?	?	?

Resume: FAIR Data Principles in HPMC



→ *Hardly (large) FAIR HPMC raw data >>* ***You can be first !***

Resume and Debate

Let’s discuss

What are the major challenges for a FAIR research data management? Have you been doing a FAIR research data management? Why not?

https://miro.com/app/board/uXjVOD10rQs=/?invite_link_id=266042838313

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

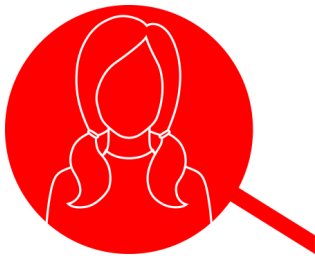
Further Information

Downloads

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

Contact

- **Newsletter**: https://lists.tu-darmstadt.de/mailman/listinfo/nfdi4ing_taskarea_doris
- **Mail**: info-doris@nfdi4ing.de (confirmation of participation needed?)
- **Web**: https://nfdi4ing.de/archetypes/doris/



Poll 4 / Evaluation & Feedback

- Anonymous –

https://evasys.zv.tum.de/evasys/public/online/index/index?online_php=&p=W29A3&ONLINEID=603704261514811903652133944932455083568 17