HPC Research Data MGMT at LRZ & beyond (InHPC-DE)

Principles – Data Storage – Data Transfer – FAIR RDM

Alexander Wellmann | Work by: LRZ FDM (RDM) Team & DSI Group

About me and about LRZ

- Chemist (B.Sc./M.Sc. At TUM) by training
- since 2022 at LRZ: Team Member Research Data Management
- What is LRZ about?

IT Service Provider for Munich Universities

120,000
Students

27.000
Employees

1,900
Professors

IT Solutions for Science

HPC
High-Performance Computing

V2C
Virtual Reality and Visualisation

QC
Quantumcomputing

AI & Big data
Competence Centre

Munich Scientific Network (MWN)

200
Gbit/s internet connection

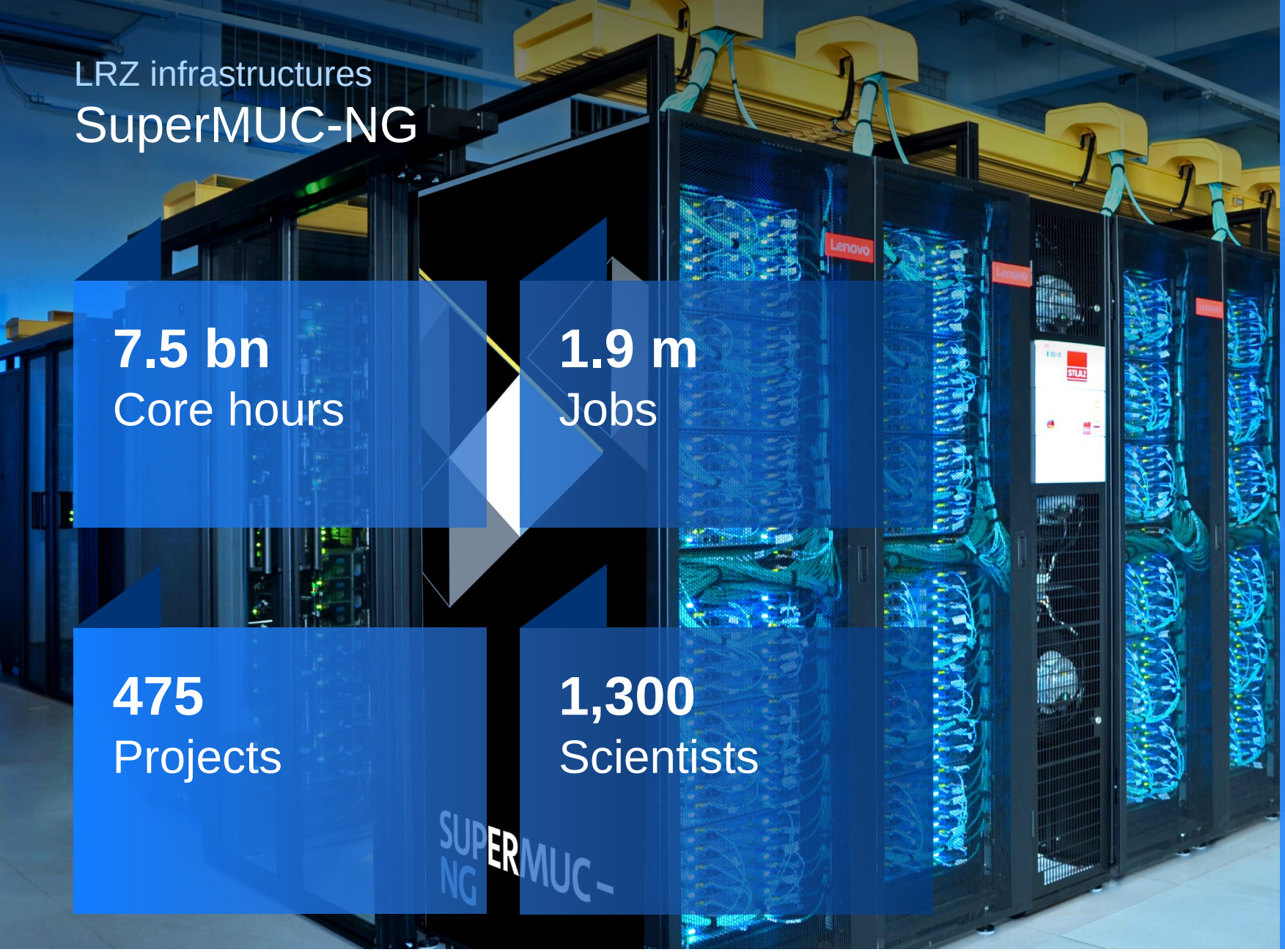
~6,000
WLAN access points

LRZ as national supercomputing centre

German supercomputing infrastructure

LRZ infrastructures

SuperMUC-NG



7.5 bn
Core hours

1.9 m
Jobs

475
Projects

1,300
Scientists

Lenovo Intel (2019)

311,040 Cores
Intel Xeon Skylake

26.9 PFlop/s Peak

19.5 PFlop/s Linpack*

719 TB Main Memory

70 PB Disk

Statistics since the start of the official user operations in August 2019 until end of 2022

SuperMUC-NG (Phase 2)

240
direct warm-water cooled compute nodes (Intel® Sapphire Rapids + Intel® Ponte Vecchio)

SD650-I v3
Lenovo platform

1 Petabyte
DAOS storage system featuring Intel® Xeon® Scalable processors of the 3rd generation as well as Intel® Optane™ Persistent Memory

Typical example of Supercomputing

Weather Simulation and Forecasts – e.g. for flood prediction for Genova

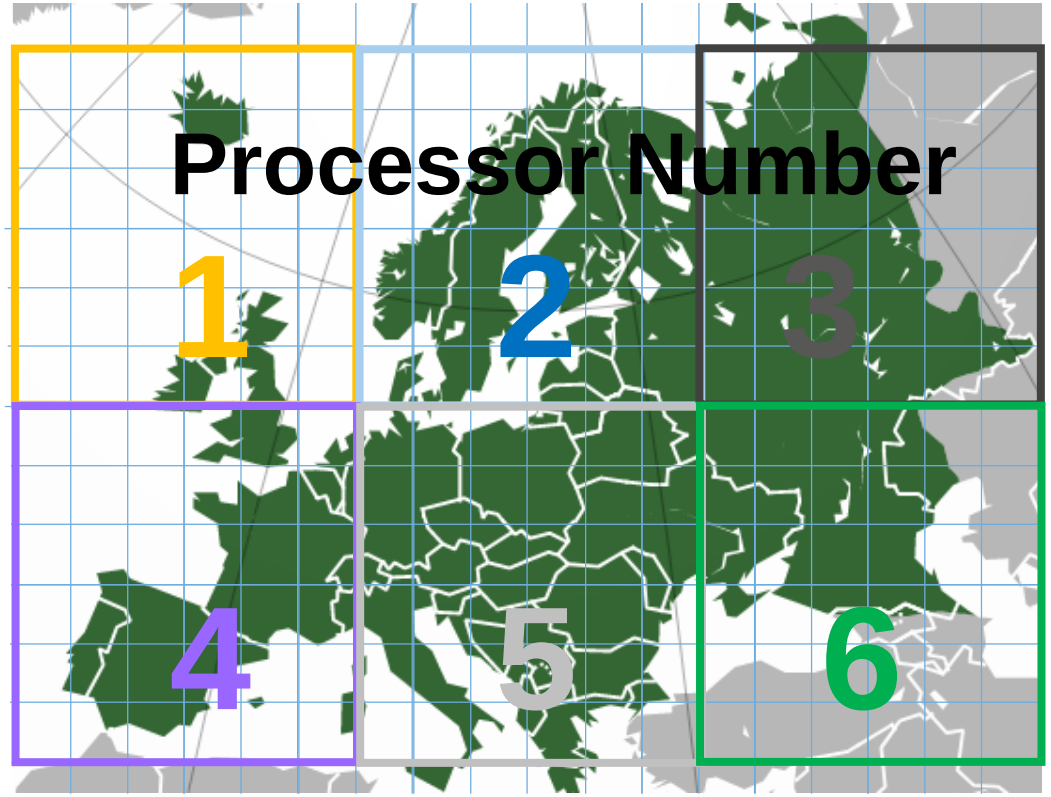
Group A. Parodi / CIMA, Savona

Figure 1: Streets of Genoa turned into rivers - Flash Flood October 2014 (frame from "ALLUVIONE A GENOVA - LE STRADE DIVENTANO FIUMI", Paolo Provenzale / THE STORM, Youtube – License: CC-BY 3.0).

Embarassingly parallel vs. Strongly-coupled tasks

Strongly-coupled tasks:

Prime example: Domain Decomposition in Weather Simulation



Fluid dynamics:

- Domain Decomposition!
  Each processor does some part of the area
- However, when air flows
  from one to another processor (border cells),
  they got to communicate about it
- Thus, a good network
  is needed

Source: Own modification of https://commons.wikimedia.org/wiki/File:Locator_map_of_Europe_with_borders.svg, Rob984 by modification of earlier work – License: CC BY-SA 4.0 (original work and this derivative)

„Flavours of huge data set production“

- **HPC** (High-Performance Computing, „traditional Supercomputing“):
  tasks involving different „computers“ within a cluster **communicating** with one another
- **HTC** (High-Throughput Computing):
  many independent („embarassingly parallel“) tasks
- If you have **parallel data analytics**, it is called **HPDA**
- **AI & ML** in- and outputs are often huge as well
- Besides computing, you can also have **high-performance measurements**,
  or data collection activities etc. producing „Big Data“

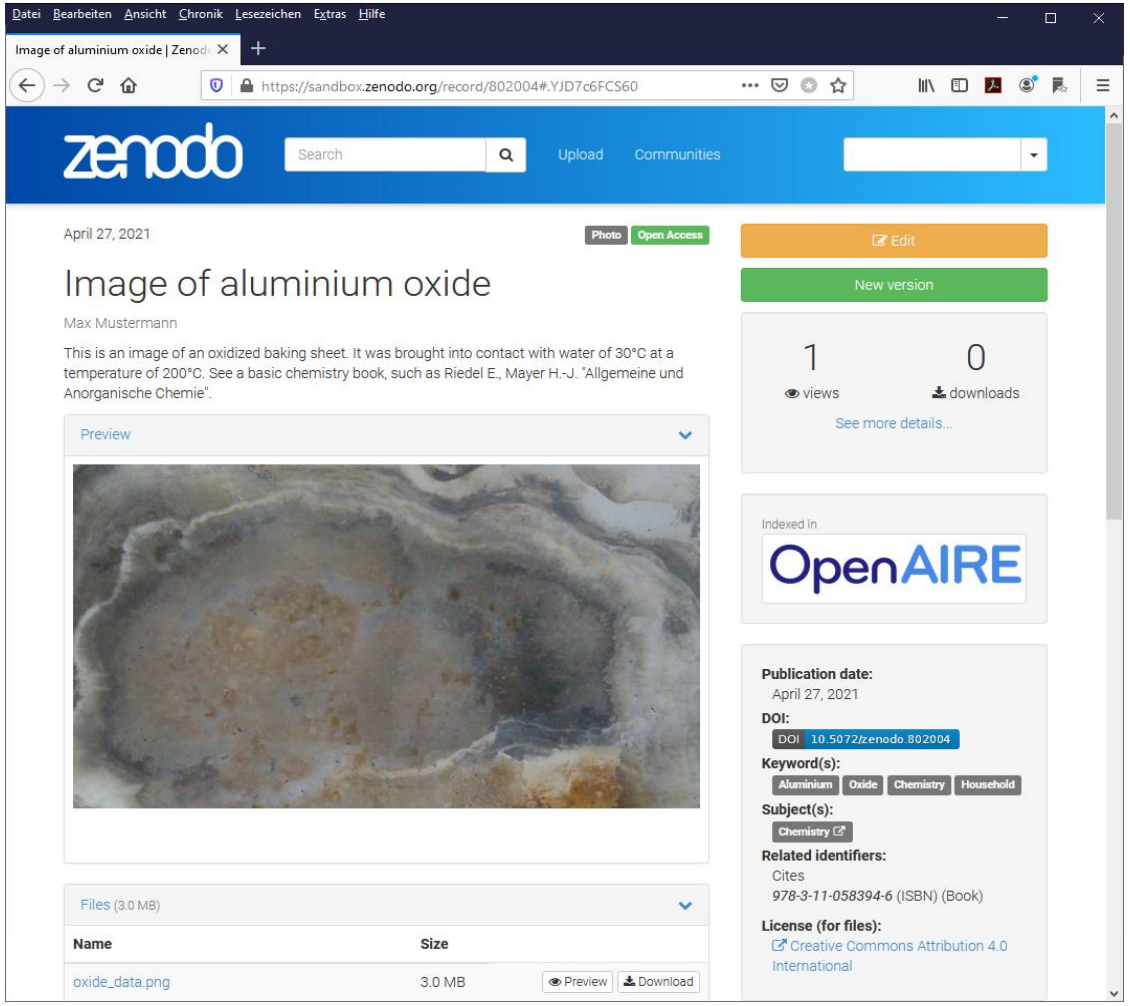
FAIR HPC Data

Why bothering? And why not to use a research data repository?

- You‘d like to make your data „citable“, getting a DOI for it?
- Your funding agency forces you to „publish“ your data?
- You want others to find your data, via data or web search engines?
- Your boss told you to deposit a description („metadata“) with your data?
- Here‘s the solution: a repository!

... but wait … maybe 50 GB of storage is not enough for your HPC data!

HPC centres have to find ways to publish data directly from centres (without repository).



Source: own screenshot and own data product

Storage Systems

Storage facilities at LRZ for different purposes

(simplified overview; official: https://doku.lrz.de/pages/viewpage.action?pageId=17694895)

Large-volume data | live data
total volume 100s of PBs

- **Cluster file systems** (of SuperMUC-NG/LC)
  $HOME, $WORK, $SCRATCH
- **Data Science Storage (DSS)**
  - NFS (Network File System) export
  - access from all LRZ, config via web & API/CLI
  - configurable via Website and API/CLI

Large-volume data | backup/archive

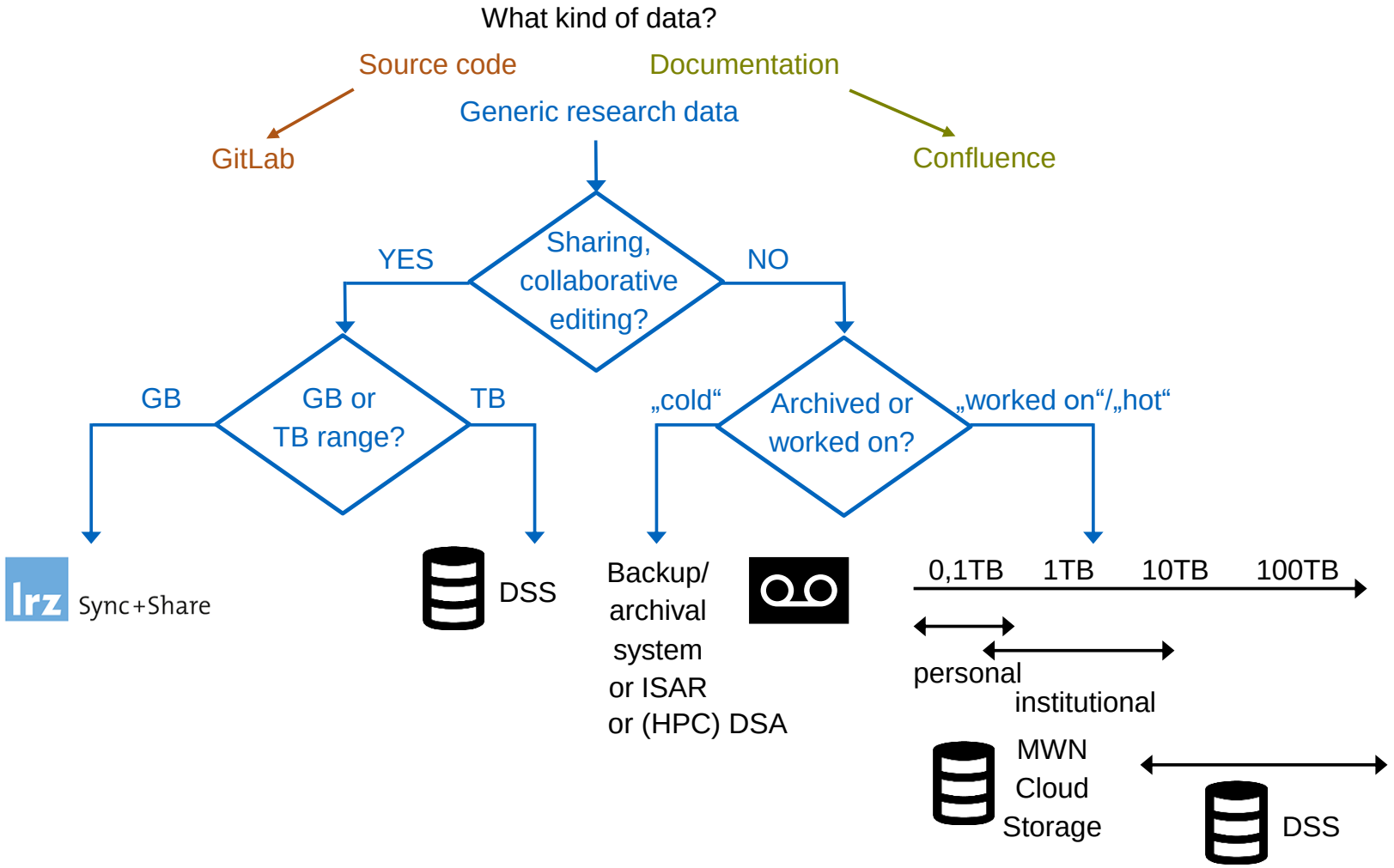
- **Tape Archive** (IBM Spectrum Protect)
- **Data Science Archive (DSA)**
  - Tapes + disk cache (appears to user as File system, automatic Cache → Disk)
  - Reactivation stages files to disk cache

Small-/medium-volume data

- **“Cloud Storage” Personal/Institutional**
  CIFS, NFS (up to 10TB per client)
  - Network Drives + WebDisk
  - ISAR option (Integrated Simple Archive)
- **LRZ Sync & Share**
  with access via app or web GUI
  “Dropbox”-like (<50GB per client)
- **LRZ GitLab**
  Git Repository Platform
  (some GB per client)

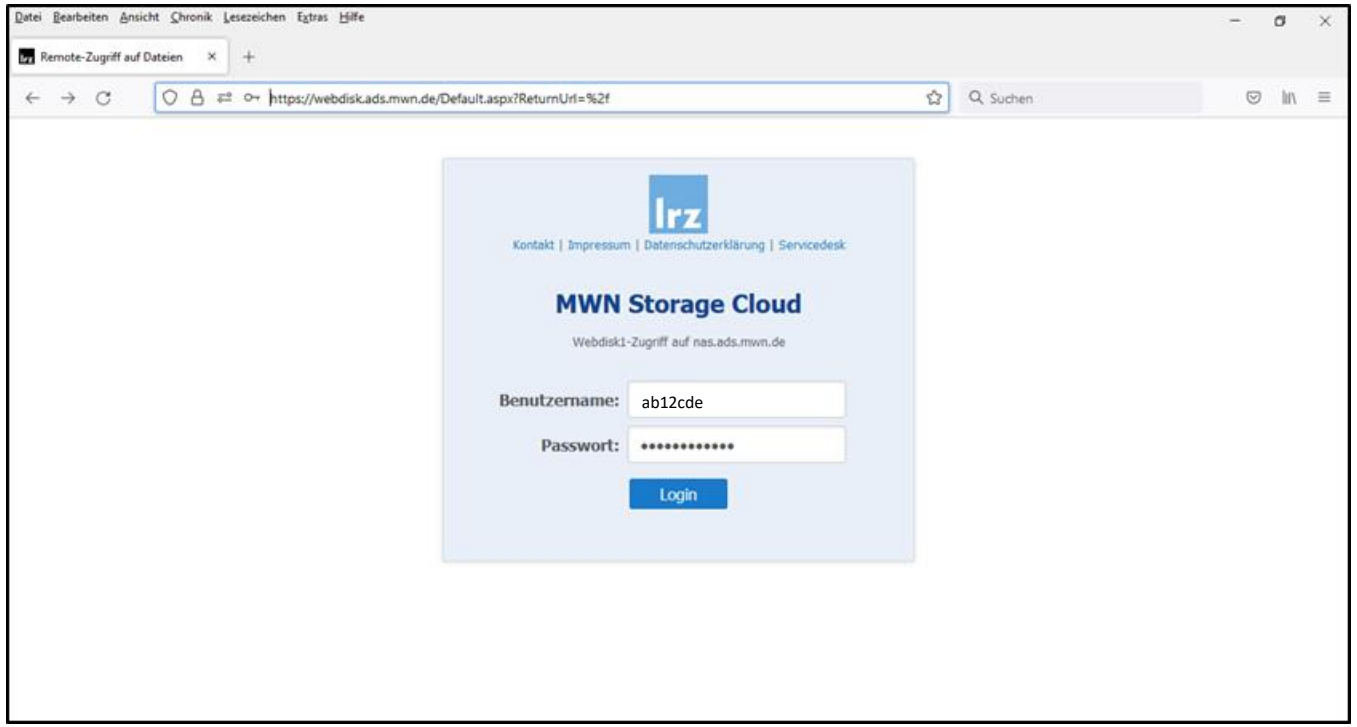
Data Storage Decision Support

(simplified overview; official: https://doku.lrz.de/pages/viewpage.action?pageId=17694895)



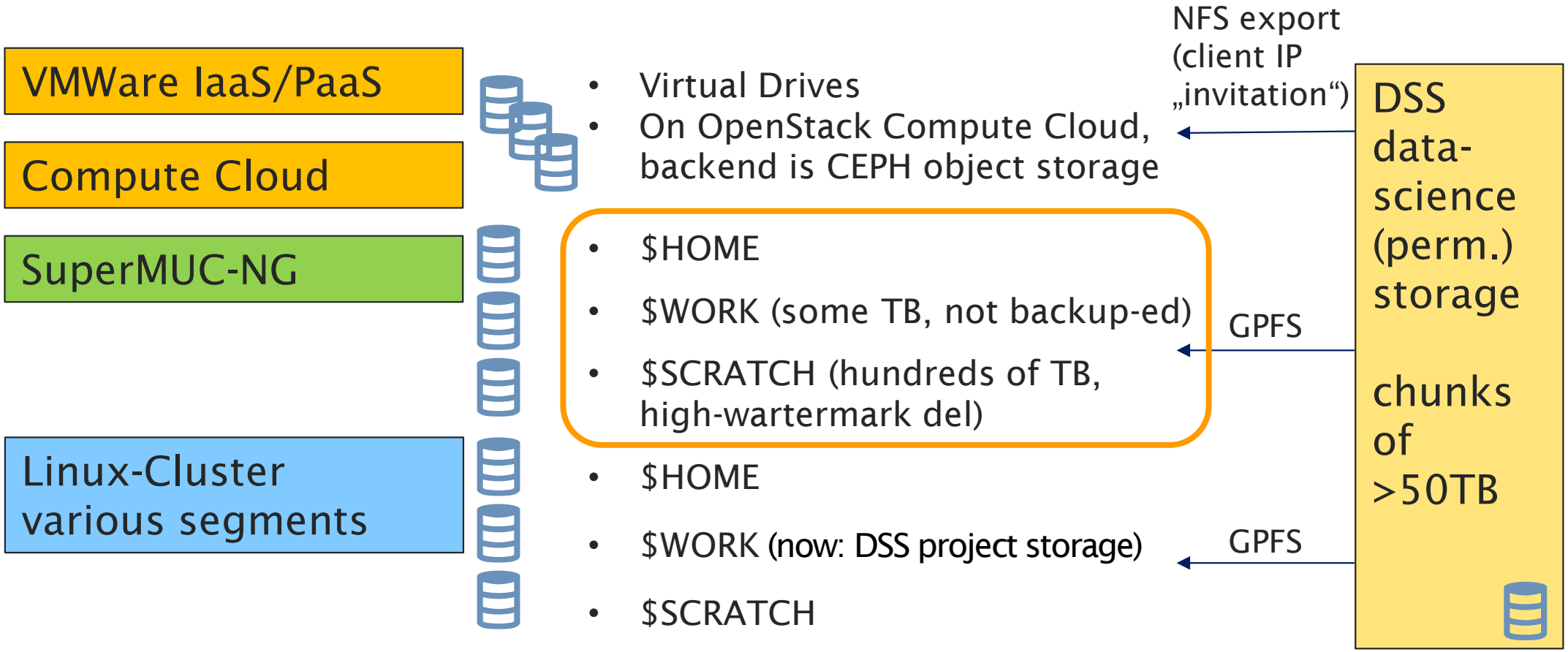
„Universal data storage“: MWN Cloud Storage

- **a.k.a. Personal/Institutional Cloud Storage, NAS, Online Storage**
- Redundant disk storage & webdisk.
- 400GB by default for TUM/LMU members (from student to professor)
- *plus institute partitions*
- Web access: webdisk.ads.mwn.de
- CIFS access (Windows Network FS)



Source: own screenshot

Supercomputing File Systems: Organisation



(LRZ) Supercomputing File Systems: Rationale

- **But Why? Why are there 3 file systems plus one DSS?**
  - $HOME is often backed-up, small-volume
  - $WORK keeps data, is larger, but not backed up
  - $SCRATCH auto-deletes files after ~1 week, if too full
    … but it is HUGE and you can use it without asking
  - DSS is highly configurable and can be accessed from outside HPC sytems
- **Some concept like this exists at most HPC centres**
- At LRZ, $WORK, $SCRATCH and DSS **are able to accomodate large datasets, extremely big write rates** and parallel writes from many computers (nodes) to one file. $HOME, even if fast, is **not** for this – it has little space!

Which systems for which data?

Excercise: Which systems for which data?

DSS and DSA

LRZ Data Science Storage (DSS)

- Scalable IBM Spectrum Scale (ex GPFS) Disk Storage up to multi-Pbyte capacity

  Ideas:
- A system most LRZ systems can connect to with NFS (i.e. mounting in LINUX/MacOS)
  in particular LRZ's High-Performance and Cloud Computing facilities
- Attached (native mount) to LRZ Linux Cluster (all nodes) and SuperMUC-NG ("login nodes^{+}")
- High Performance (easily >500 MB/s transfer rates)
- Institutes can buy own building blocks of ~1PB;
  for medium demand, parts of $\geq$ 20TB can be allocated within a shared LRZ block
- Access concept:
  - shares of storage allocated on request & managed by *data curator*, who creates
  - containers within these shares (like uppermost-level directories), for which the curator can name *container managers*, who can "invite" IP addresses for NFS mount and
  - users who create directories and files within the containers
- Access Management: invitiation + ACLs

LRZ Data Science Archive (DSA)

Source: LRZ (doku.lrz.de Museum of phased-out HPC Systems)

Tape Archival as in the good old Baarer Straße Days?

Not anymore ☺ – LRZ DSA makes tape archival as easy as writing to disk.

LRZ DSA Usage vs. Classical IBM Spectrum Protect Tape

https://doku.lrz.de/display/PUBLIC/DSA+documentation+additions+for+users
https://doku.lrz.de/display/PUBLIC/Backup+und+Archivierung

DSA (dssweb + API/CLI/GLOBUS)

DSA Data Lifecycle

- A few hours after you've copied the file over to DSA: Archival in 2 data centres
- Approximately 24h* after the file has been created in DSA: immutable, not deleteable for 10 years
- High-watermark deletion of cache copy:
  need to re-activate file for reading (otherwise you get *Permission Denied*)
  → Use API/Command-Line Interface to activate files (make them accessible)
  or transfer them with GLOBUS online (will trigger automatic activation)
- After 10 years, you can delete files
- Storing many small files is an anti-pattern!

Tape (datweb + manual config)

Usage from
Virtual Machine

- Follow best-practice guide
- Make config files & include-exclude list – mind special syntax
- After first access, set passwordaccess generate to have automated access (token)

- manually or automatically schedule `dsmc incremental` or `dsmc archive`

Handle your backup/archival so as to be **actually safe!**

- Test regulary if your backup setup is working: Restore data from the backups to check if you can read them.
- Carry out integrity checks to ensure that data has not been corrupted. Use checksum tools.
- Store several copies on different storage devices in multiple locations (3-2-1 Backup Rule).

Data Transfer & Sharing

When http download is not enough… uftp, GridFTP and GLOBUS

https://doku.lrz.de/display/PUBLIC/Data+Transfer+Options+on+SuperMUC-NG

Principles of transfer of huge datasets

- Transfers 100s of TBs – over night, if needed between *“endpoints”*
- Asynchronous (unattended) file transfers
- Resume interrupted transfers
- Parallel data streams

Solutions in InHPC-DE – i.e. between Stuttgart/Jülich/Garching

- uftp (cf. Jülich Supercomputing Centre) – *recommended, e.g. with asymmetric key auth*
- GLOBUS (and GLOBUS online, see following slides) – *handy, with certs or auto-generated certs*
- GridFTP (the classical technique underlying GLOBUS – *only with X.509 certs*), e.g.:
  `globus-url-copy -vb -p 6 gsiftp://datagw.supermuc.lrz.de/PATH/FILE/AT/LRZ gsiftp://judacsrv.fz-juelich.de/PATH/AT/JSC`

GLOBUS online

GLOBUS portal (https://app.globus.org)

- Convenient Web Portal – “file explorer” GUI
- Transfers queued and executed asynchronously
- Login with university credentials (delegation)

What is Globus Sharing?

- Extension to transfer service
- “*Shared endpoint*” data accessible to all Globus users (Link to shared data can be made)
- External user access to shared endpoint: Globus will translate these accesses to the local system as if they were carried out by the local user who created the Shared Endpoint.

FAIR RDM in HPC

(cf. also NFDI4Ing-DORIS & InHPC-DE)

How to make data on LRZ systems FAIR?

Systems/Portals
(TUM library)

Workbench

mediaTUM
Universitätsbibliothek
Technische Universität München

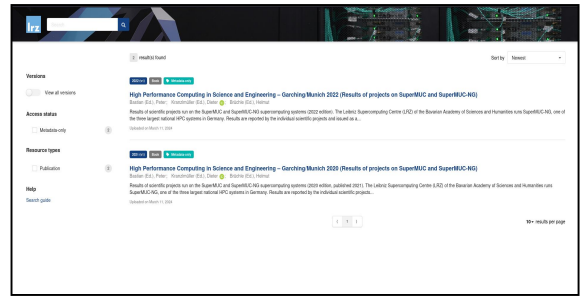
Metadata

Data

User Input

Automatic enrichment
(e.g. from simulation framework or user admin system)

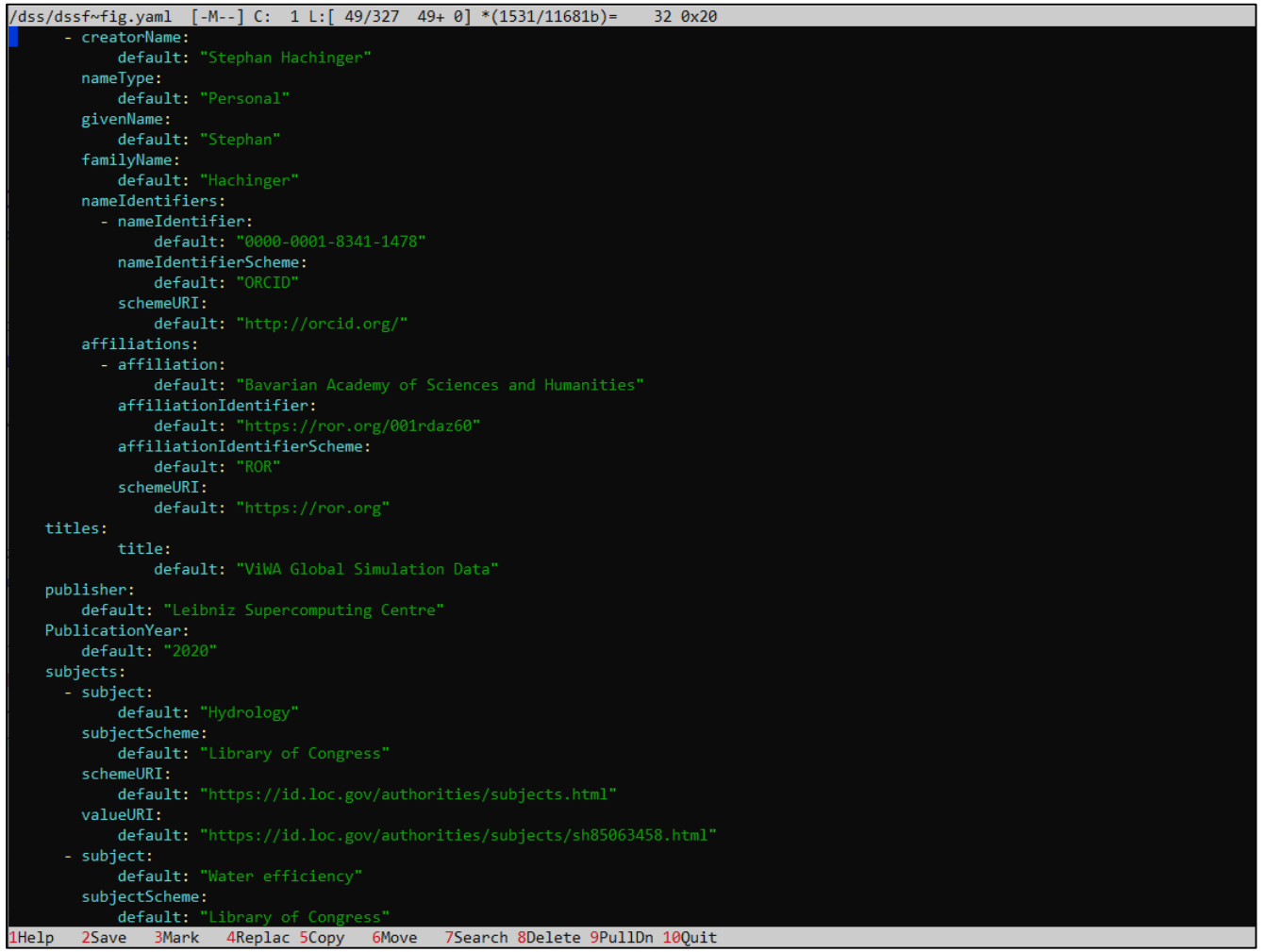
NFDi4ing



LRZ-LTDS Metadata Store and Portal (non-TUM) – work in progress

Metadata standard & „crawling“: Getting LRZ data FAIR

- DataCite metadata
  - universal
  - minimal (for DOI) but extensible
- express your interest plus
  deposit a .metadata.yaml file
  in all directories you want published –
  system looks for it



Source: own screenshot, CC-0

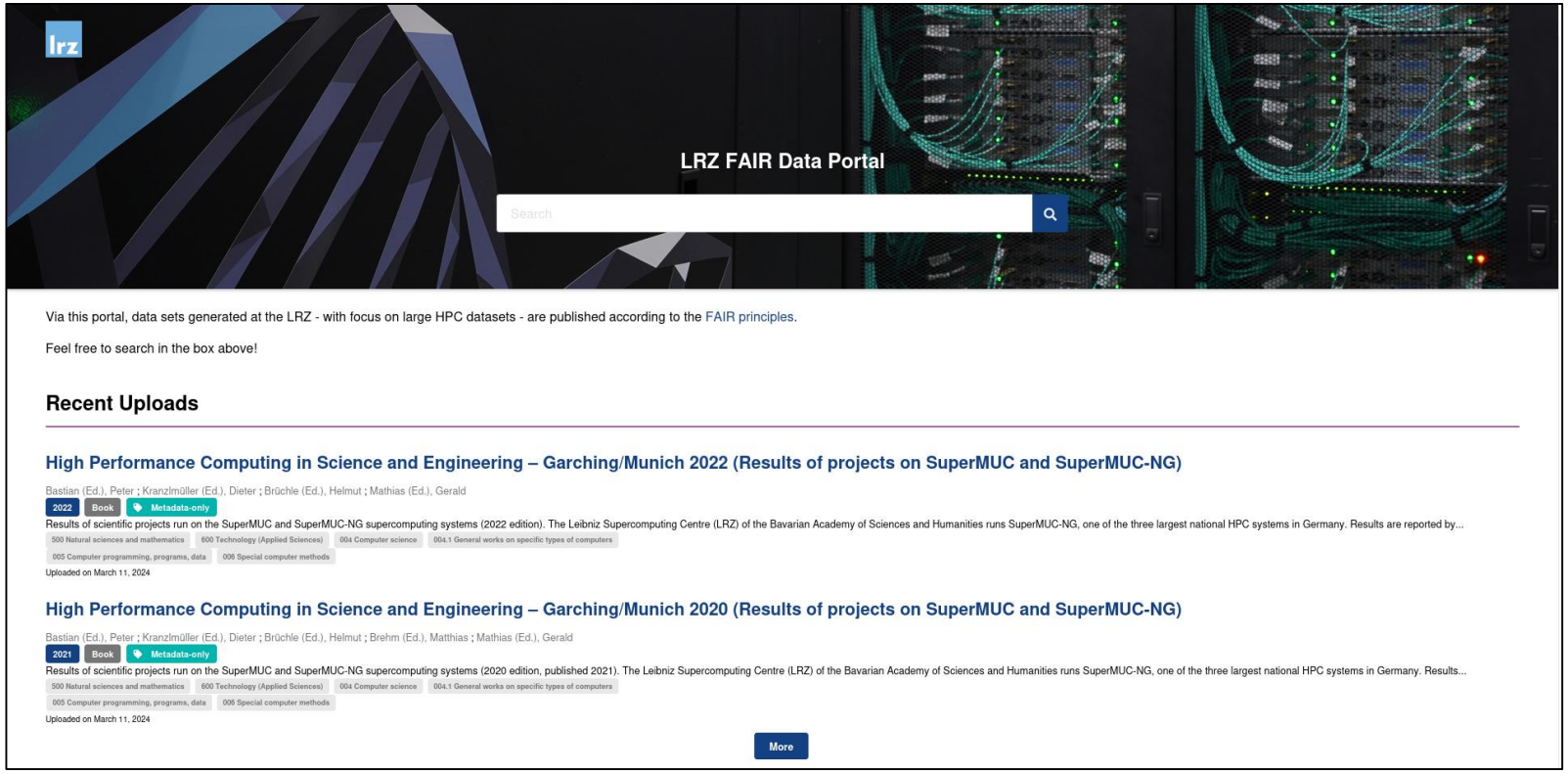
WebUI – Metadata Input Form

Source: own screenshot, CC-0

- Input your metadata in a web-form
- Mandatory and optional fields are marked as such
- Output: a compatible metadata .yaml file to store alongside your data

Work in progress: Soon available to the public!

WebUI – Portal: Customized InvenioRDM Demonstrator



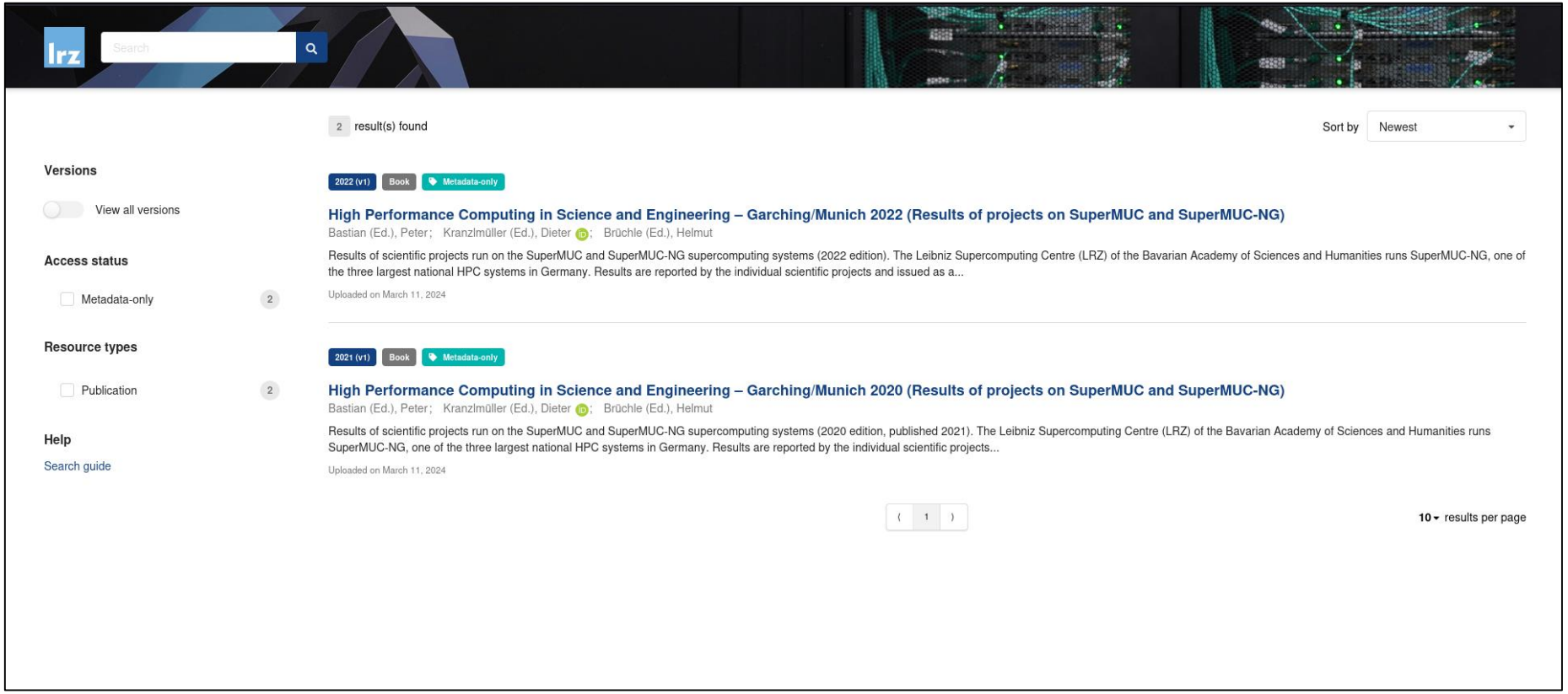
Source: own screenshot, CC-0

- “friendly user phase”: Some selected datasets can already be uploaded
- Automatic larger-scale production service will be available soon!

First demonstrator **already online and searchable by everyone!**

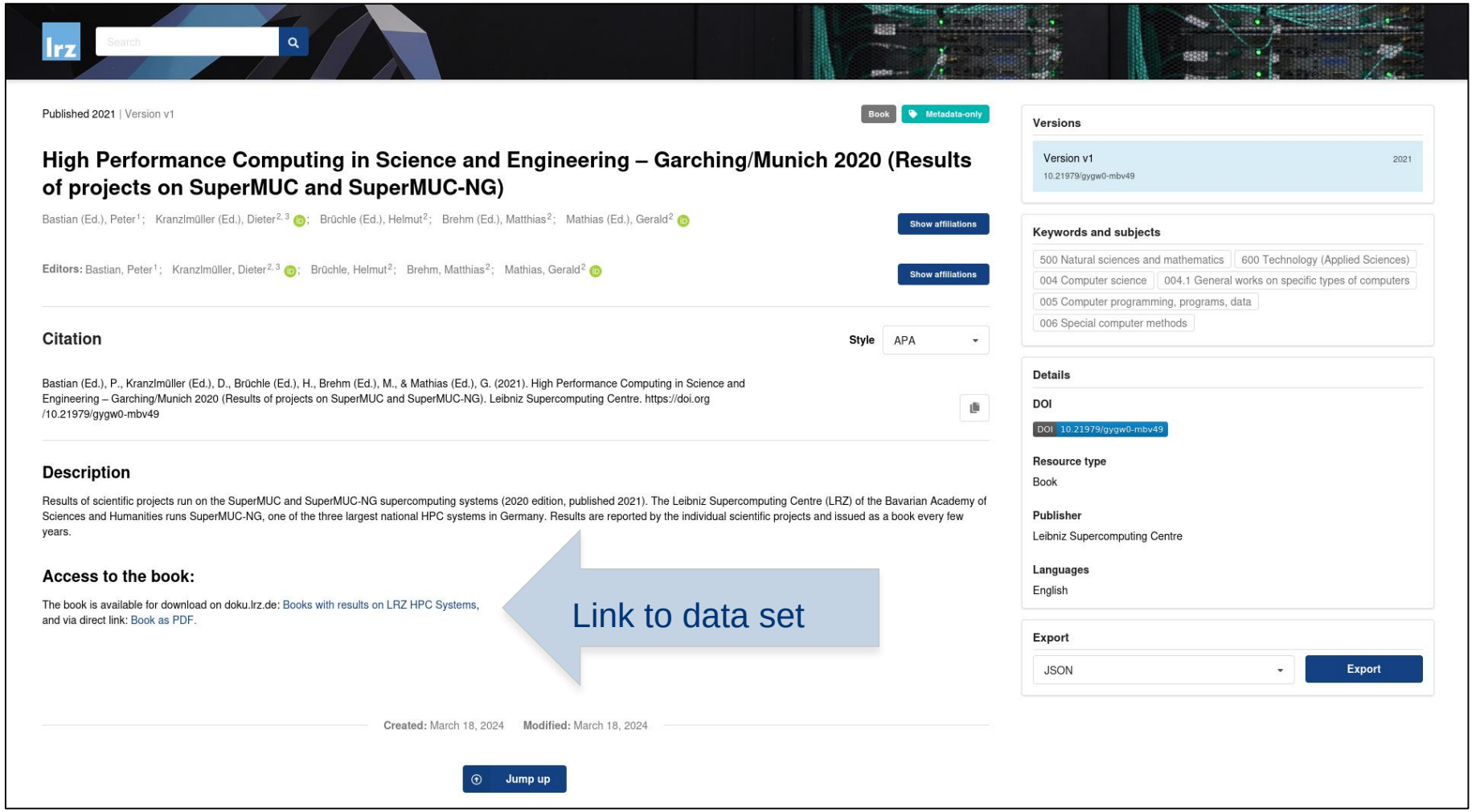
https://rdm.lab.lrz.de/

WebUI – Results Page



Source: own screenshot, CC-0

WebUI – Details Page (DOI Landing Page)



Source: own screenshot, CC-0

Outlook and … thanks for your attention from LRZ FDM/RDM Team

LRZ-RDM-Team:
rdm@lists.lrz.de

Website:
https://www.lrz.de/forschung/projekte/forschung-daten/

- Timescale for LTDS: next yrs automatization for larger-scale production service
- LTDS focused on LRZ HPC customers (those not covered by e.g. TUM-UB/LMU-UB)
- Important collaborations with DSI/CSI, InHPC-DE, University Libraries, NFDI4Ing,…
- Lead FDM Team
  - Dr. Stephan Hachinger (admin)
  - Johannes Munke (tech)
- FDM Team (2023/04)
  - Mukund Biradar
  - Mohamad Hayek
  - Huseyn Gurbanov
  - Viktoria Pauw
  - Alexander Wellmann

Leibniz-Rechenzentrum, Boltzmannstr. 1, 85748 Garching bei München