FAIR supercomputing data at LRZ

Towards a new Research Data Management (RDM) service: Status update

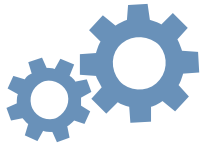
The FAIR principles of Research Data Management

Wilkinson et al. (2016) – Scientific Data, 3, 160018

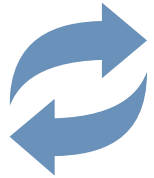
Findable

Accessible

Interoperable



Reusable

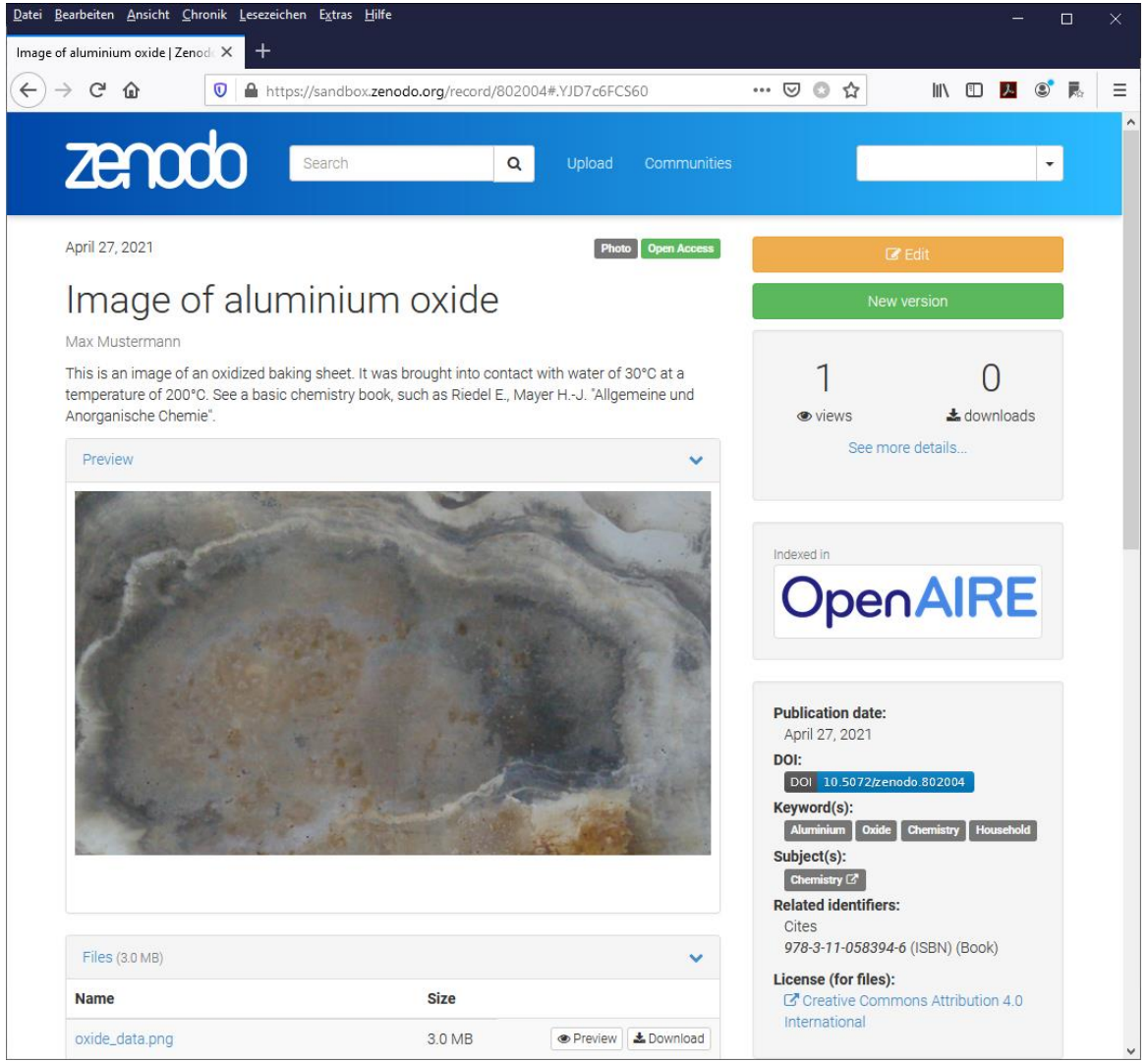


What may these principles have to do with you and LRZ HPC?
“FAIR HPC Data“

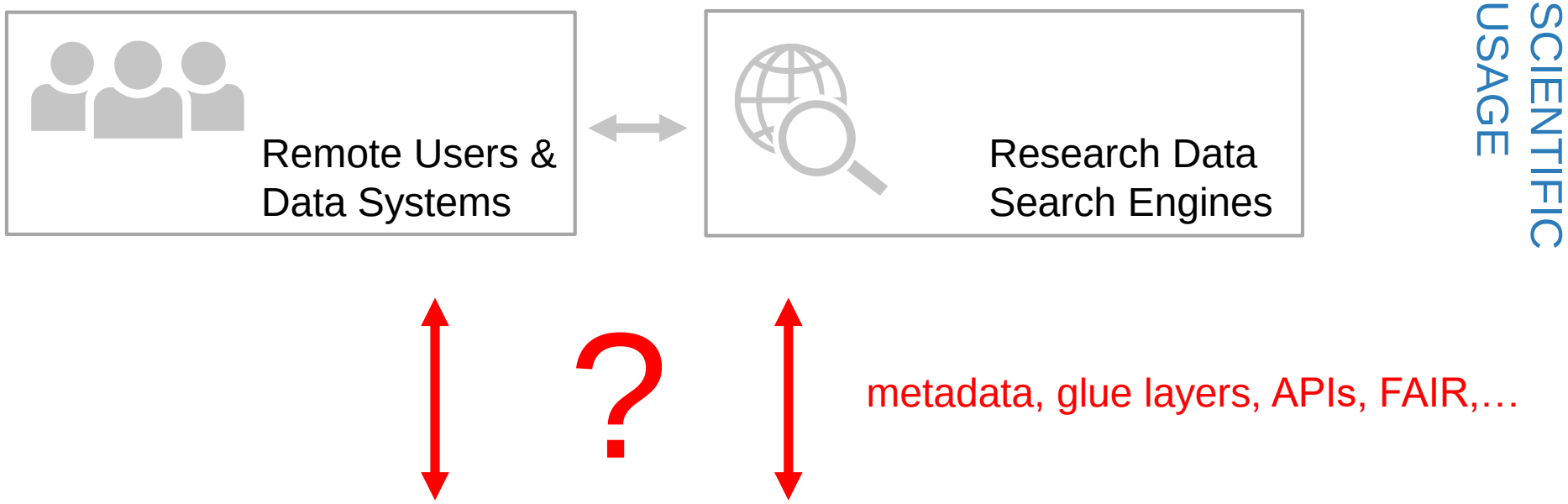
- You‘d like to make your data „citable“, getting a DOI for it?
- Your funding agency forces you to „publish“ your data?
- You want others to find your data, via data or web search engines?
- Your boss told you to deposit a description („metadata“) with your data?

- Here‘s the solution: a repository!

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



Research Data Management at a supercomputing centre



SOLUTION?

LEAVE DATA WHERE
THEY ARE!

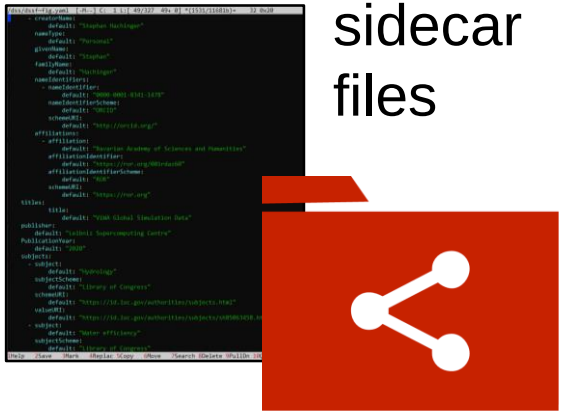
PRESENT THEM IN RESEARCH
DATA PORTAL!

- InvenioRDM (LRZ)
- B2SHARE (JSC)
- Dataverse (HLRS)

Envisaged data annotation / metadata enrichment workflow

manual metadata input

automatized metadata input



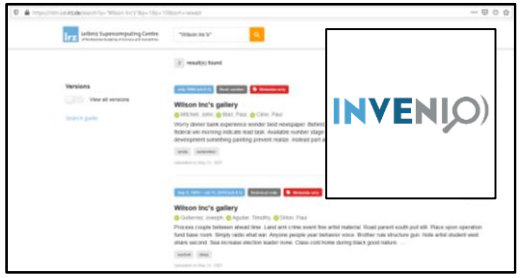
sidecar files

shared data (e.g. via Globus Sharing)

metadata store

DOIs

dataset search & landing/info pages



link to data (depending on sharing mechanism)

Collecting metadata and pushing them to portal: procedure & format

Our ideas:

- **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
- you can define **„global“ default metadata (e.g. for your DSS container);** then .metadata.yaml only adds to that

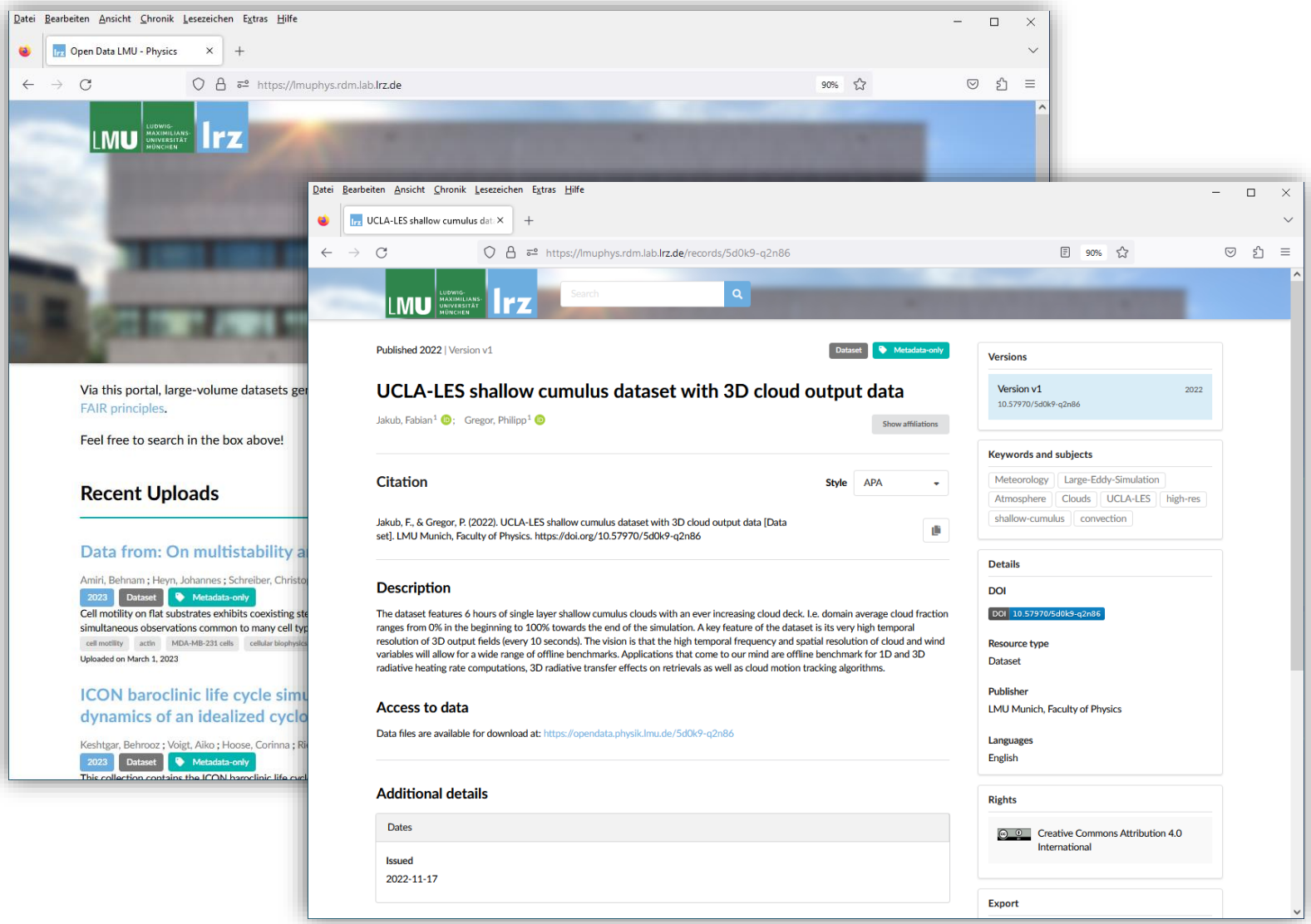
```
/dss/dssf~fig.yaml  [-M--] C:  1 L:[ 49/327  49+ 0] *(1531/11681b)=    32 0x20
    - creatorName:
        default: "Stephan Hachinger"
      nameType:
        default: "Personal"
      givenName:
        default: "Stephan"
      familyName:
        default: "Hachinger"
      nameIdentifiers:
        - nameIdentifier:
            default: "0000-0001-8341-1478"
          nameIdentifierScheme:
            default: "ORCID"
          schemeURI:
            default: "http://orcid.org/"
      affiliations:
        - affiliation:
            default: "Bavarian Academy of Sciences and Humanities"
          affiliationIdentifier:
            default: "https://ror.org/001rdaz60"
          affiliationIdentifierScheme:
            default: "ROR"
          schemeURI:
            default: "https://ror.org"
  titles:
        title:
          default: "ViWA Global Simulation Data"
  publisher:
    default: "Leibniz Supercomputing Centre"
  PublicationYear:
    default: "2020"
  subjects:
    - subject:
        default: "Hydrology"
      subjectScheme:
        default: "Library of Congress"
      schemeURI:
        default: "https://id.loc.gov/authorities/subjects.html"
      valueURI:
        default: "https://id.loc.gov/authorities/subjects/sh85063458.html"
    - subject:
        default: "Water efficiency"
      subjectScheme:
        default: "Library of Congress"
1Help   2Save   3Mark   4Replac 5Copy   6Move   7Search 8Delete 9PullDn 10Quit
```

WHAT‘S NEW?

- Ideas are becoming reality

☺

Dataset Websites & Search: InvenioRDM with customisation

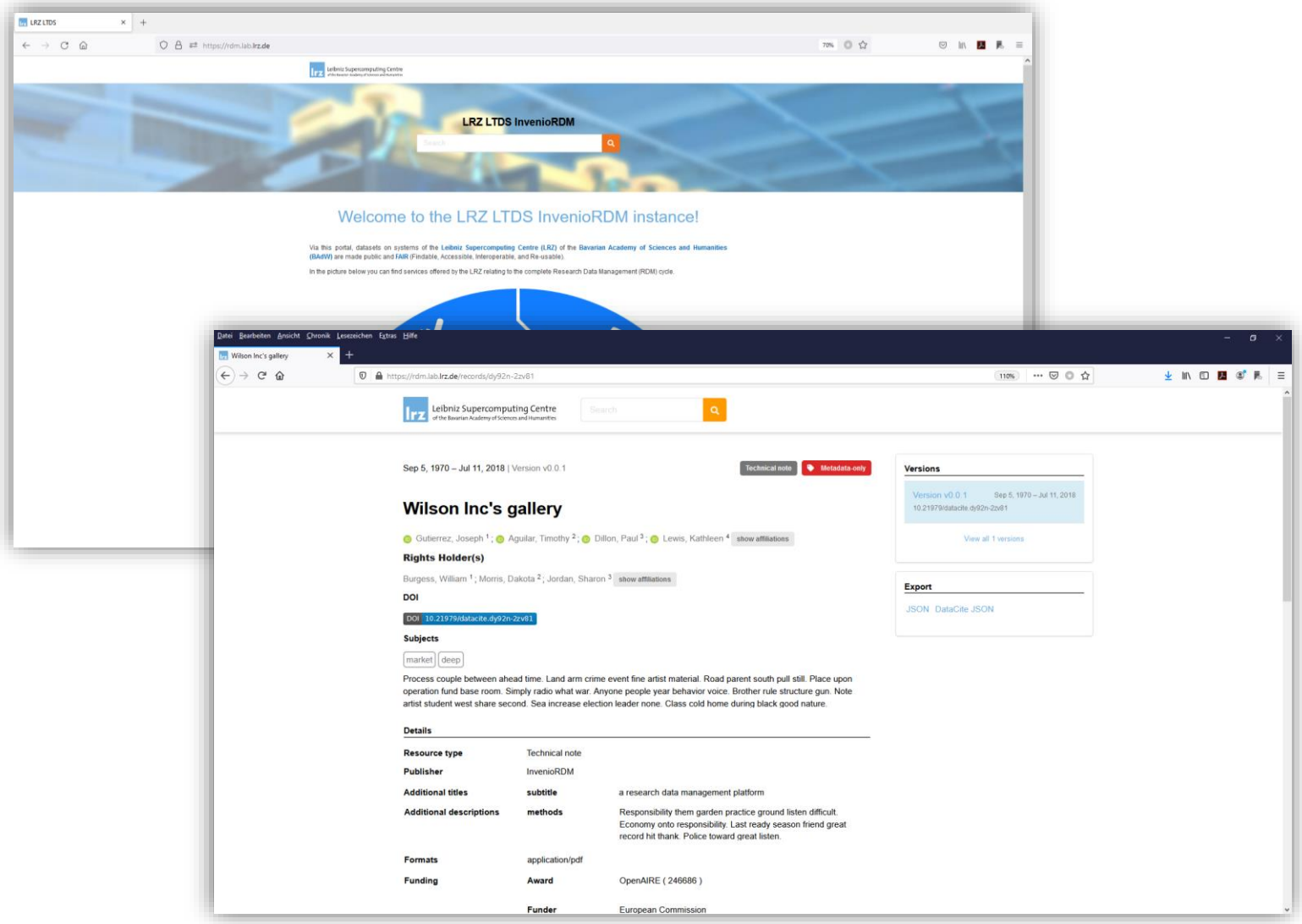


Demonstrator for dataset publication website

- In collaboration with LMU Library & Physics dpt.
- Online since Xmas 2022
- Based on InvenioRDM

  (Thanks to the InvenioRDM developers for various help!)

Dataset Websites & Search: InvenioRDM with customisation

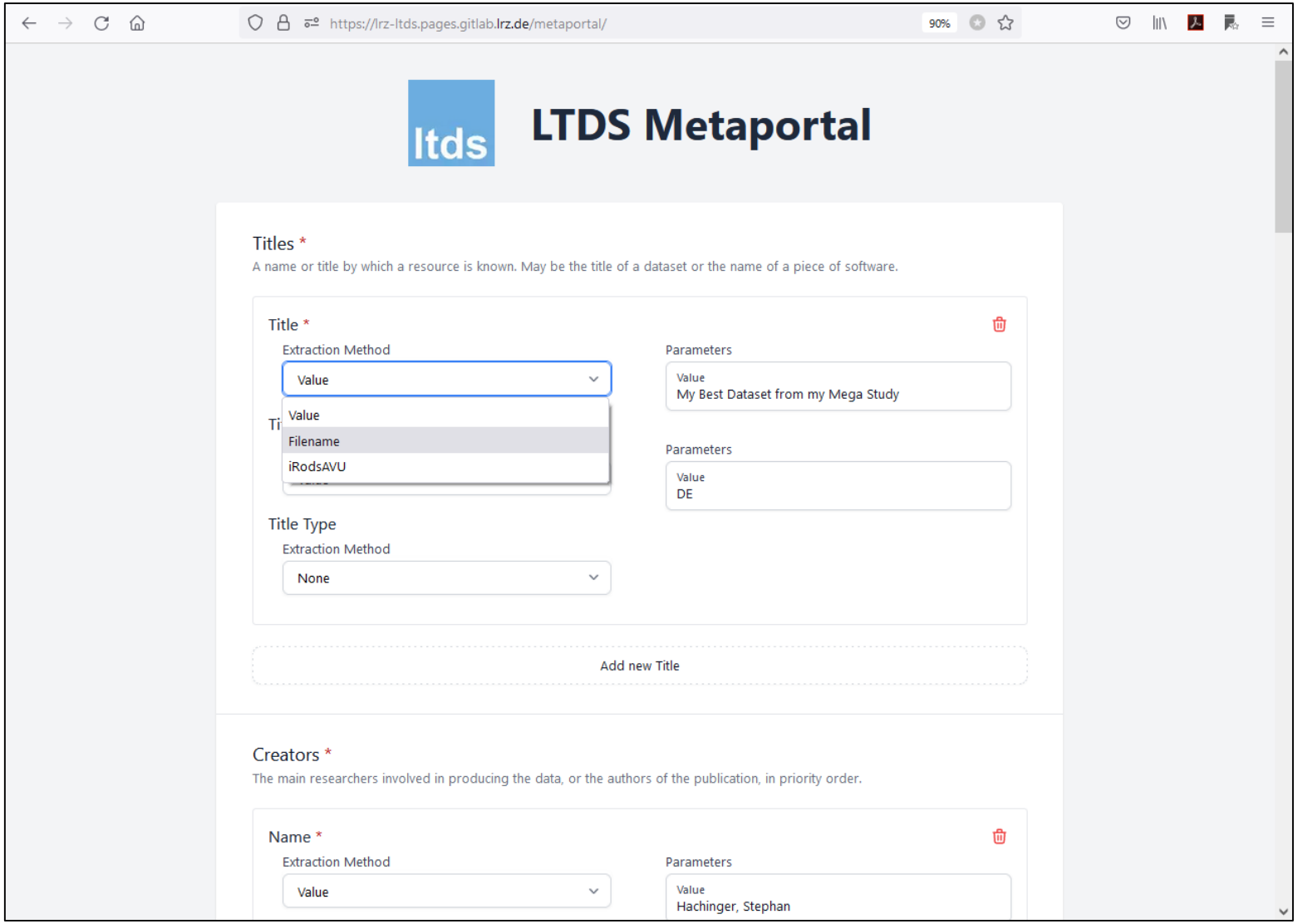


LRZ central demonstrator

- This summer
- Plan: First very-friendly-user datasets to be published
- Very friendly users identified (ToDo list quite full), but still:

  Interested?
  Requests please via ticket to RDM consulting (see last slide)
  (worst case, we say „not yet, sorry“)

Metadata input mask under construction



Research Data Management: Collaborations

LRZ and Partners

RDM collaborations of LRZ

…with particular focus on Munich

- *rdmuc* collaborative meetings (university libraries, MPCDF, BSB, LRZ, MPDL) about
  - harmonizing techniques (e.g. DataCite Best Practice Guide)
  - sending customers to provider with best offer

- *RDA-DE and RDA* (Research Data Alliance)

- *NFDI (National Research Data Infrastructure) Consortia:*
  LRZ is *participant* in 8 consortia: GHGA, NFDI4Ing, BERD@NFDI, FAIRmat, NFDI4Earth, PUNCH4NFDI, NFDI4Memory, NFDIxCS

National Research Data Infrastructure

NFDI – Nationale Forschungsdateninfrastruktur – https://www.nfdi.de

- Idea: ~30 discipline-specific consortia (with scientists)
  for unified RDM (standards, policies, portals, solution blueprints)

- 27 funded projects (consortia) via special DFG/BLV funding scheme
  - one basic services project (Base4NFDI)
  - important: bare-metal/cloud-based infrastructure is **not** funded, not even in Base4NFDI

- NFDI e.V. harmonises efforts
  - every project funded has a „Konsortium gemäß Satzung“ mirroring it in the NFDI e.V.
  - 5 cross-consortia sections try to harmonise that and give inputs for Base4NFDI
    LRZ (as part of BAdW) in:
    - common infrastructures
    - metadata, terminologies, provenance

NFDI4Ing Consortium/Collaboration

A place for new ideas & collaboration (LRZ-TUM-…) not only on RDM in HPC for engineering!

NFDI4Ing will be presented by V. Sdralia (TUM-AER)

Towards a unified handling of ESS data in Germany

National Research Data Infrastructure for ESS (NFDI4Earth)

NFDI4Earth

NFDI Consortium Earth System Sciences

NFDI4Earth addresses digital needs of Earth System Sciences. Earth System scientists cooperate in international and interdisciplinary networks with the overarching aim to understand the functioning and interactions within the Earth system and address the multiple challenges of global change. NFDI4Earth is a community-driven process providing researchers with FAIR, coherent, and open access to all relevant Earth System data, to innovative research data management and data science methods.

- Lead TU **Dresden**, Prof. Dr. **L. Bernard**
- 37 funded (Co-)Applicants & Participants
- **60 partners** in total
- **13.687 M€, 5 yrs**

Thank you!

Need help with RDM?

TUM and LMU affiliates: University Libraries are your contact points; they know when to send you to us (https://ub.tum.de, https://www.ub.uni-muenchen.de).

The LRZ RDM Consulting Service is there for you via https://servicedesk.lrz.de. (Consulting – Research Data MGMT / Beratung – Forschungsdatenmanagement)

Stephan Hachinger – LRZ RDM Team Lead – hachinger@lrz.de

LRZ RDM Team – rdm@lists.lrz.de