Navigating the Data Patchwork: Strategies for Integrating Metadata Catalogs, Data Publications, and Archives

NOBUGS, Grenoble, September 2024

O. Knodel, D. Pape, M. Voigt, T. Gruber, S. Müller and G. Juckeland

Outline

The HZDR

Research areas in the fields:
Energy, Health, Matter

RDM Ecosystem and Strategy

Policy and data publication systems and services at HZDR

(Meta)Data Patchwork

Approach to combine metadata sources from various different systems

Our Research Facility and our Large Scale Research Infrastructures

The Helmholtz-Zentrum Dresden - Rossendorf

— Employees approx. **1,470**. Thereof **670** scientists.

— Member of:

HELMHOLTZ

RESEARCH FOR GRAND CHALLENGES

ELBE – Center for High-Power Radiation Sources

— Electron accelerator, free-electron lasers & THz source.

— Positrons, protons, neutrons as well as X-ray and gamma radiation.

Dresden High Magnetic Field Laboratory (HLD)

— Europe's highest pulsed magnetic fields.

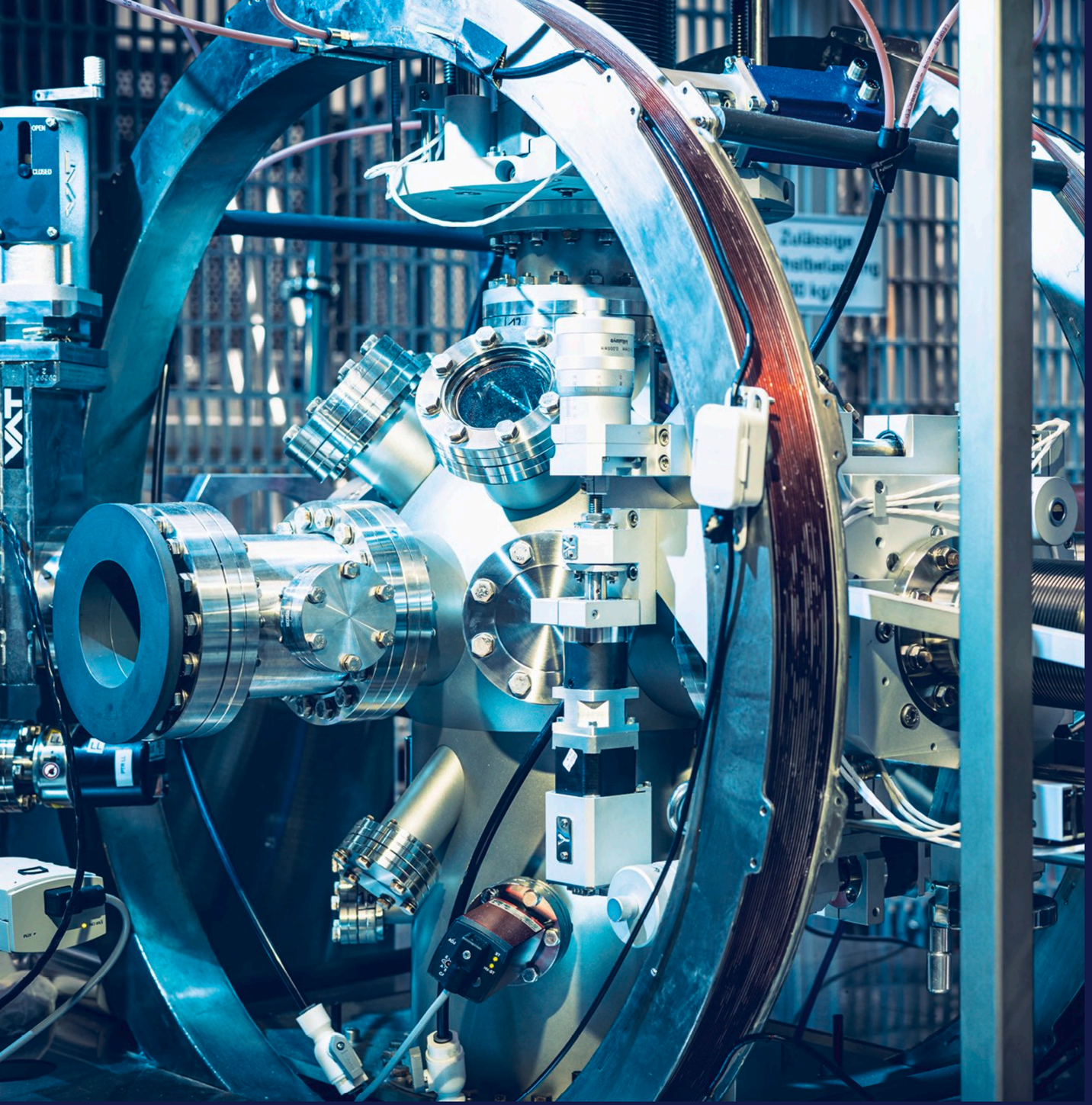
Ion Beam Center (IBC)

— Nanoscale surface analysis and modification.

DREsden Sodium facility for DYNamo (DRESDYN)

— Large scale liquid sodium experiments with geo- and astrophysical background.

The Rossendorf Beamline at ESRF (ROBL)



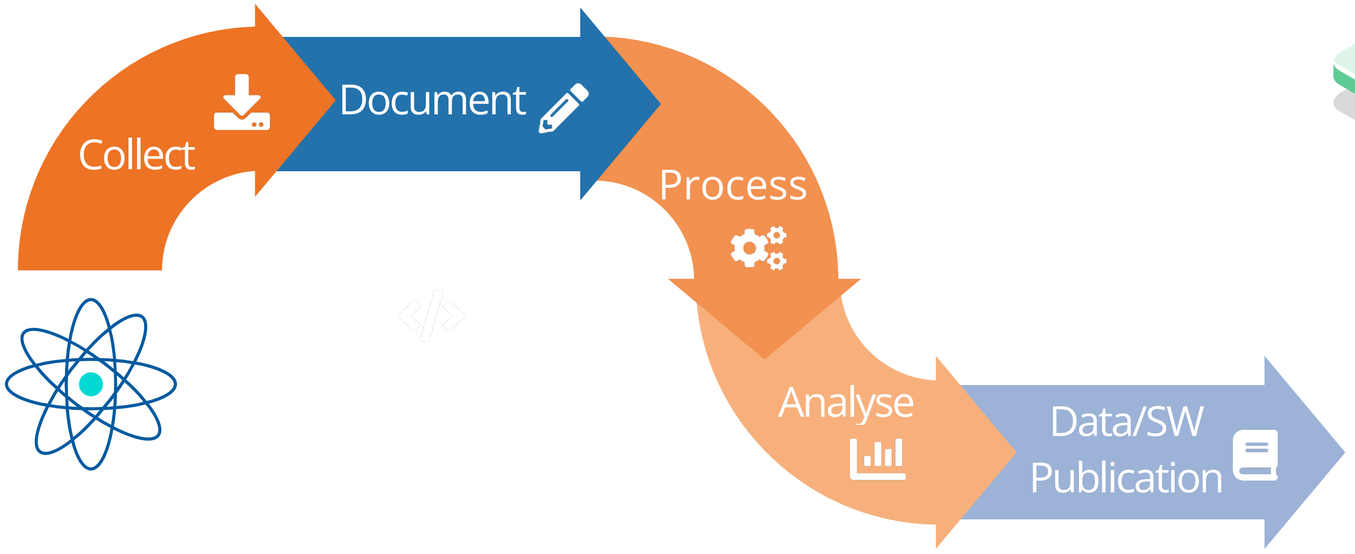
Our RDM **Ecosystem** and **Strategy**

The Mission for the Data Management and HPC Group

— Supporting our scientists and research experiment with tools:
  - electronic lab books,
  - interactive analysis,
  - publication repositories for data and software,
  - scientific workflow management,
  - PID (handle) generation and management.

— Providing support in connecting experiments, detectors and diagnostics to our infrastructure.

— Establishment of analysis workflows and HPC applications.



Development of the HZDR

Data Management Ecosystem

Scientific Computing Department at HZDR — 2017

Establishment of the Scientific Computing division within the IT as a bridge to the science

Data Policy — 2018

Legal framework for the institutional data management at HZDR.

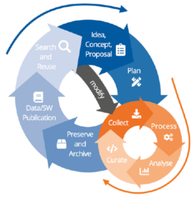
Data Repository RODARE — 2018

Provision of the Rosendorf Data Repository for data and software publications.

RODARE
ROSSENDORF DATA REPOSITORY

RDM Landscape and Data Management Group — 2019

Analysis of data/service landscape at HZDR and development of an uniform data lifecycle.



HZDR Metadata Catalogue — 2024

Metadata catalogue for additional experiment-specific metadata as extension of RODARE.

RO MECA
Rossendorf Metadata Catalogue

HZDR RDM Strategy— 2025

Identification of the need to develop a data management strategy together with our scientists.

HZDR

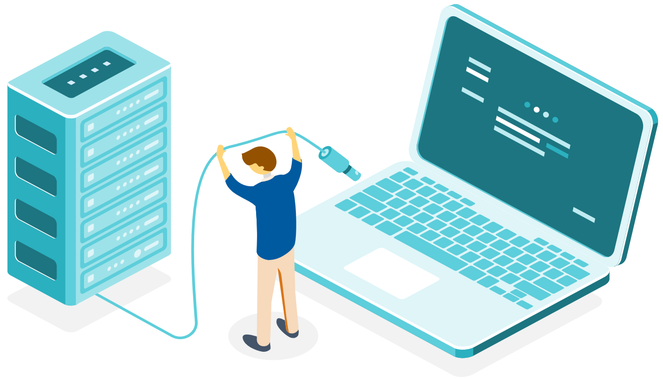
Our Motivation for Developing a Full Digital Data Lifecycle:

— The HZDR has a data policy since May 2018.

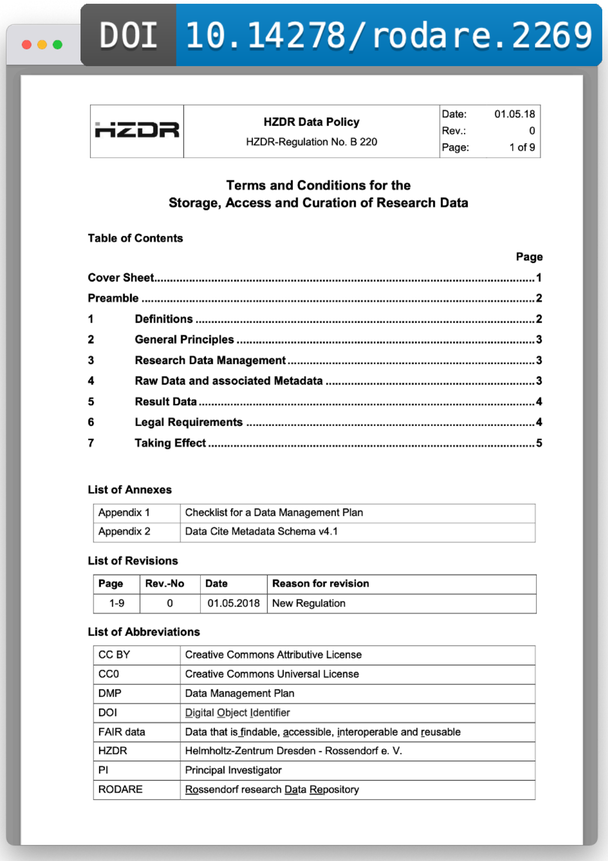
— Reasons for the development:

- Establishment and legitimisation of publication, handling and deletion of data generated or taken at the HZDR,
- Legal framework for data management and…

… to further develop data management services together with our scientists!



The HZDR Data Policy



DOI 10.14278/rodare.2269

HZDR Data Policy
HZDR-Regulation No. B 220

Date: 01.05.18
Rev.: 0
Page: 1 of 9

Terms and Conditions for the Storage, Access and Curation of Research Data

Table of Contents

List of Annexes

Appendix 1	Checklist for a Data Management Plan
Appendix 2	Data Cite Metadata Schema v4.1

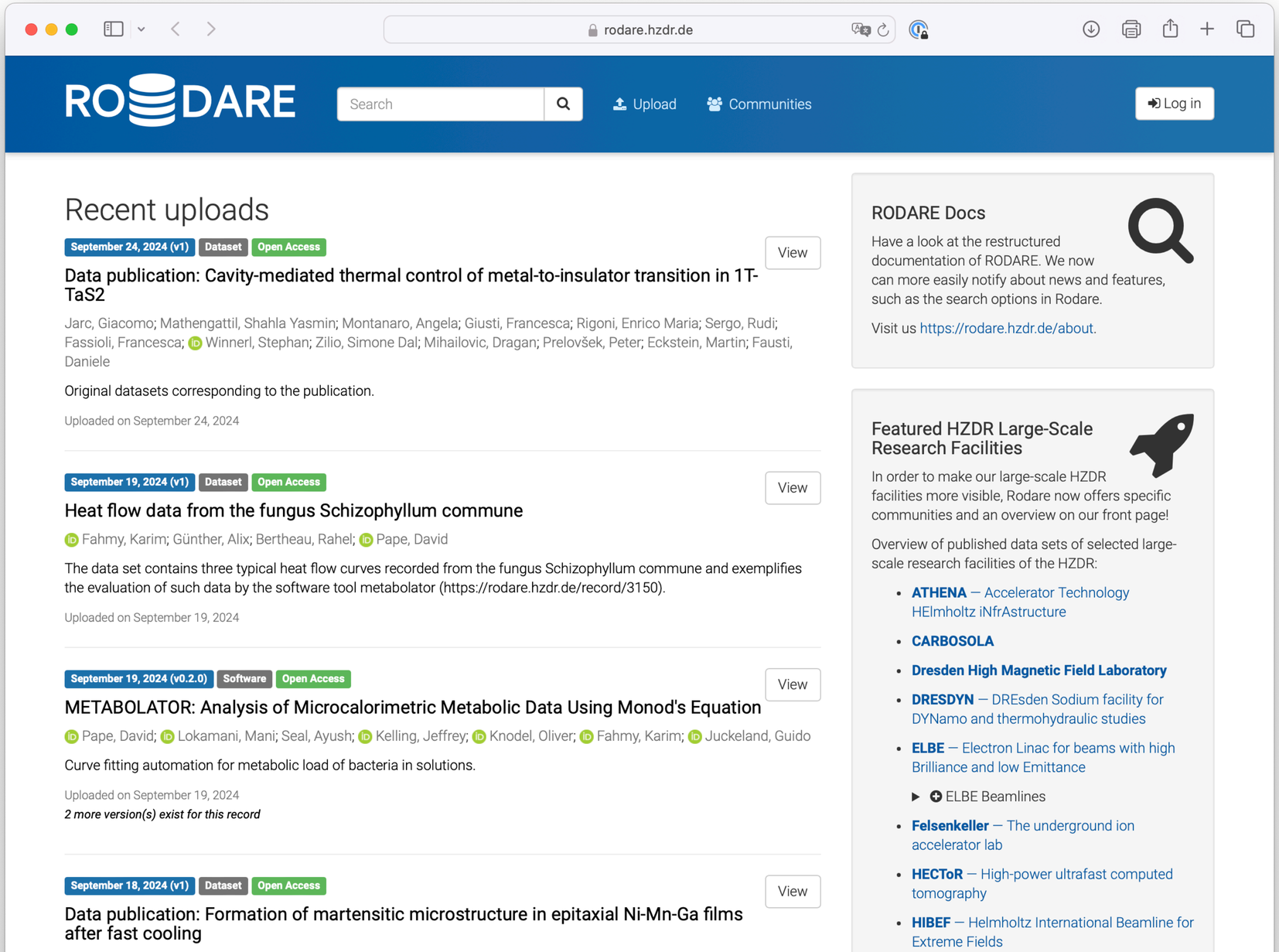
List of Revisions

Page	Rev.-No	Date	Reason for revision
1-9	0	01.05.2018	New Regulation

List of Abbreviations

CC BY	Creative Commons Attributive License
CC0	Creative Commons Universal License
DMP	Data Management Plan
DOI	Digital Object Identifier
FAIR data	Data that is findable, accessible, interoperable and reusable
HZDR	Helmholtz-Zentrum Dresden - Rossendorf e. V.
PI	Principal Investigator
RODARE	Rossendorf research Data Repository

Our Contribution to Support FAIR Research: RODARE



Powered by:

INVENIO RDM (in 2025)

Harvested via OAI-PMH by:

B2FIND EUDAT

Registered in:

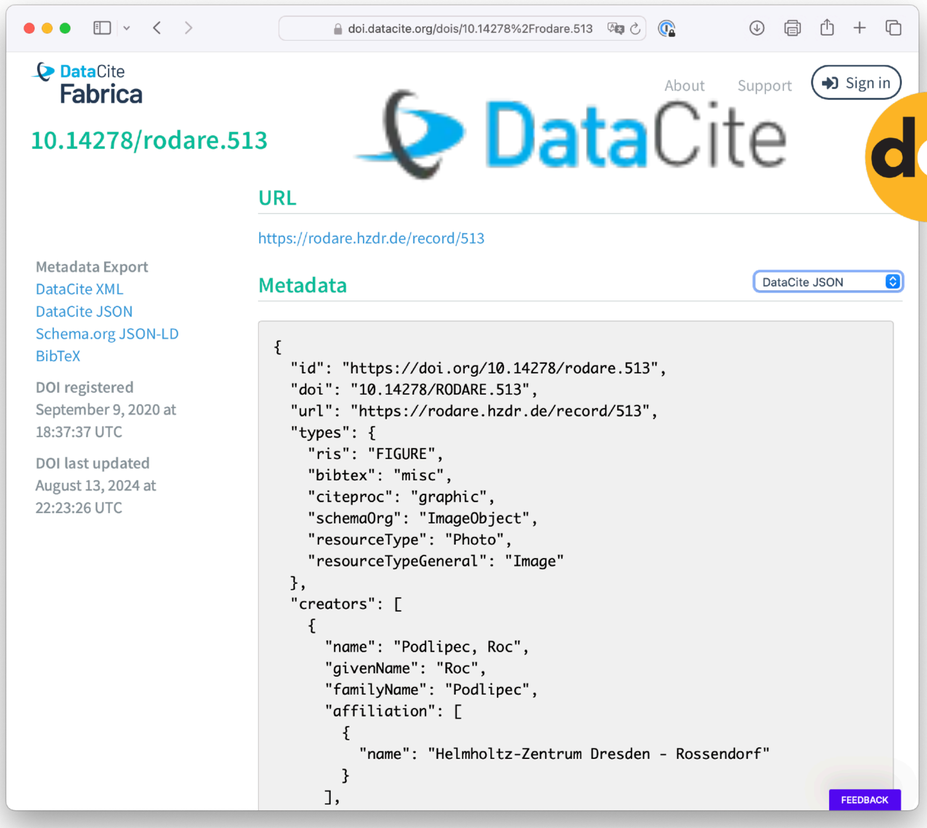
re3data.org REGISTRY OF RESEARCH DATA REPOSITORIES

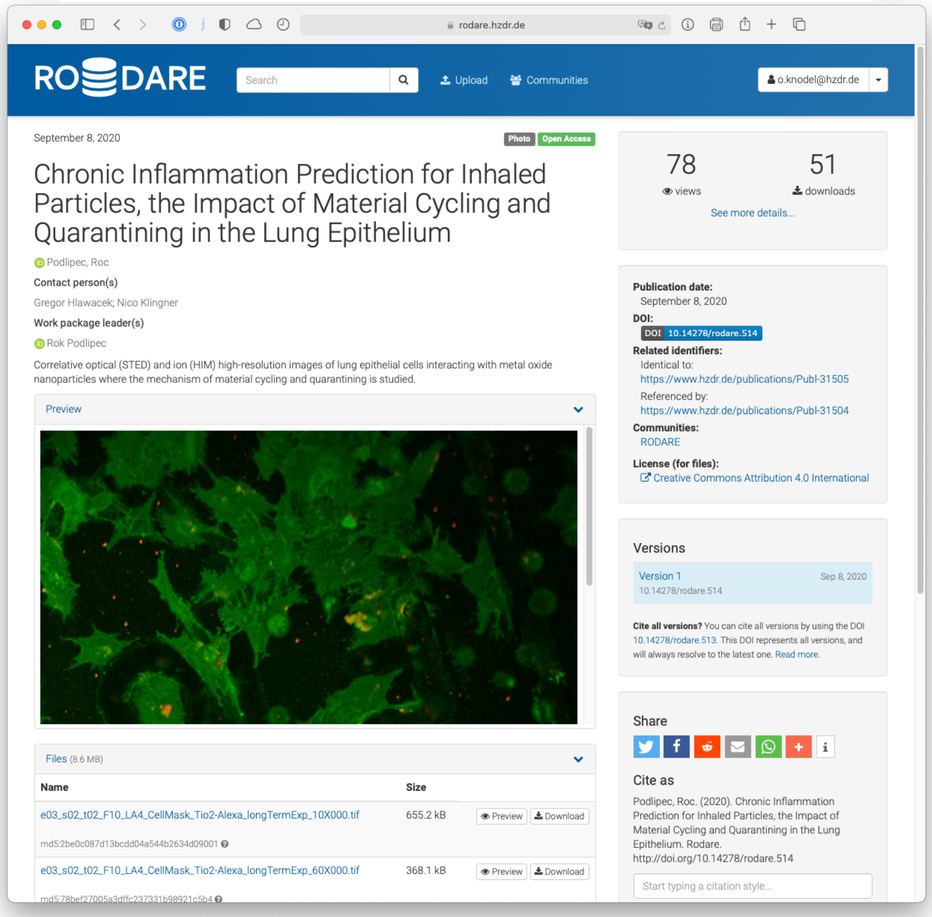
http://doi.org/10.17616/R3BR40

DRESDEN concept HZDR

Metadata and Data in RODARE

- In our data publication system the datasets are described via **DataCite** metadata to be FAIR.
- The DataCite metadata is attached to the DOI and harvested via portals, such as **B2Find**.





File Access:
Rest API

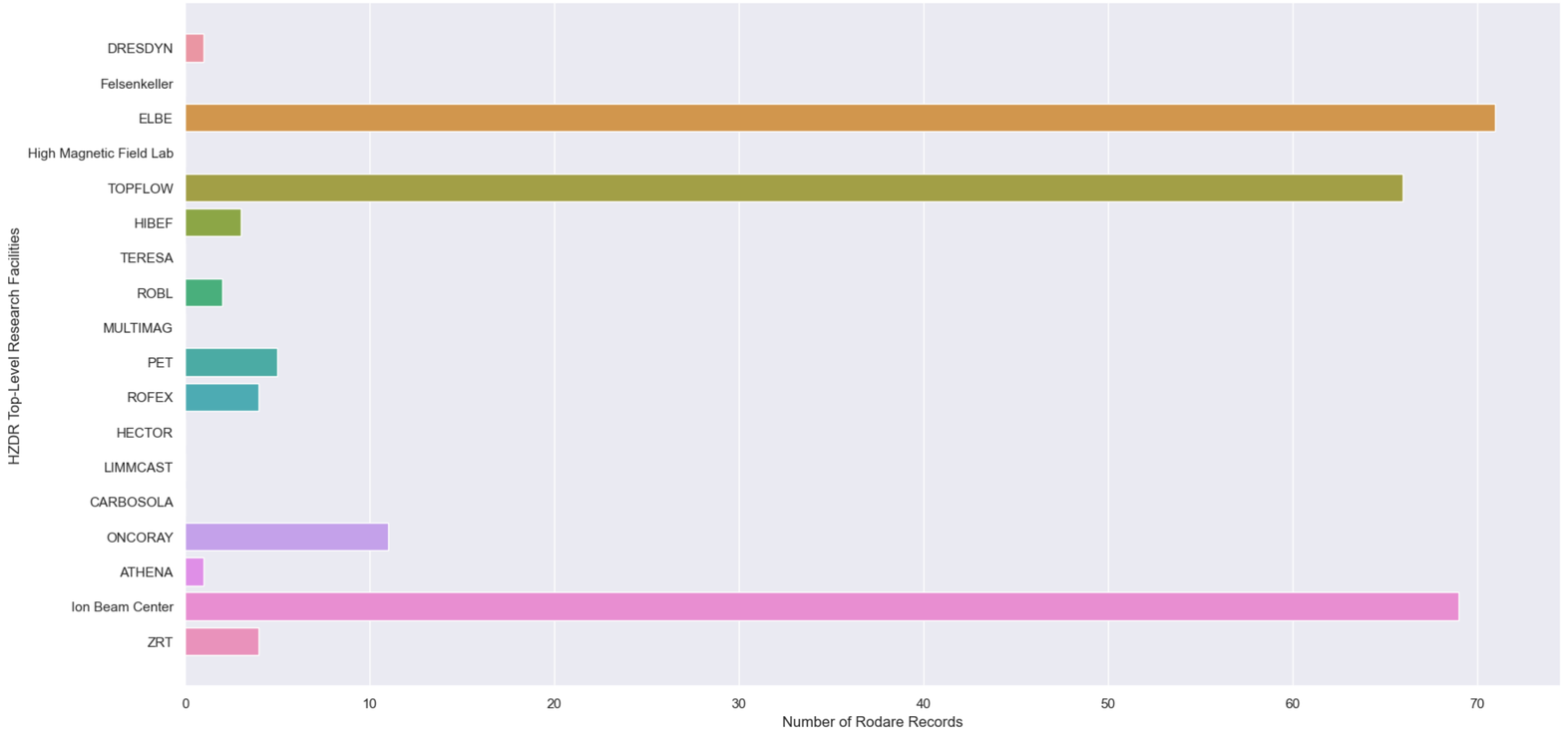
Filesystem

Metadata API:
Rest API, OAI-PMH

B2FIND
EUDAT
HZDR

Additional Metadata Through Communities

- Experiment-specific metadata is difficult because the RODARE metadata is the same for every entry.
- We integrated the facility information throughout Communities for our large-scale facilities:
  - **18** facilities from different research areas…
  - …and one facility (ELBE radiation source) with **9** beamlines.



Featured HZDR Large-Scale Research Facilities

In order to make our large-scale HZDR facilities more visible, Rodare now offers specific communities and an overview on our front page!

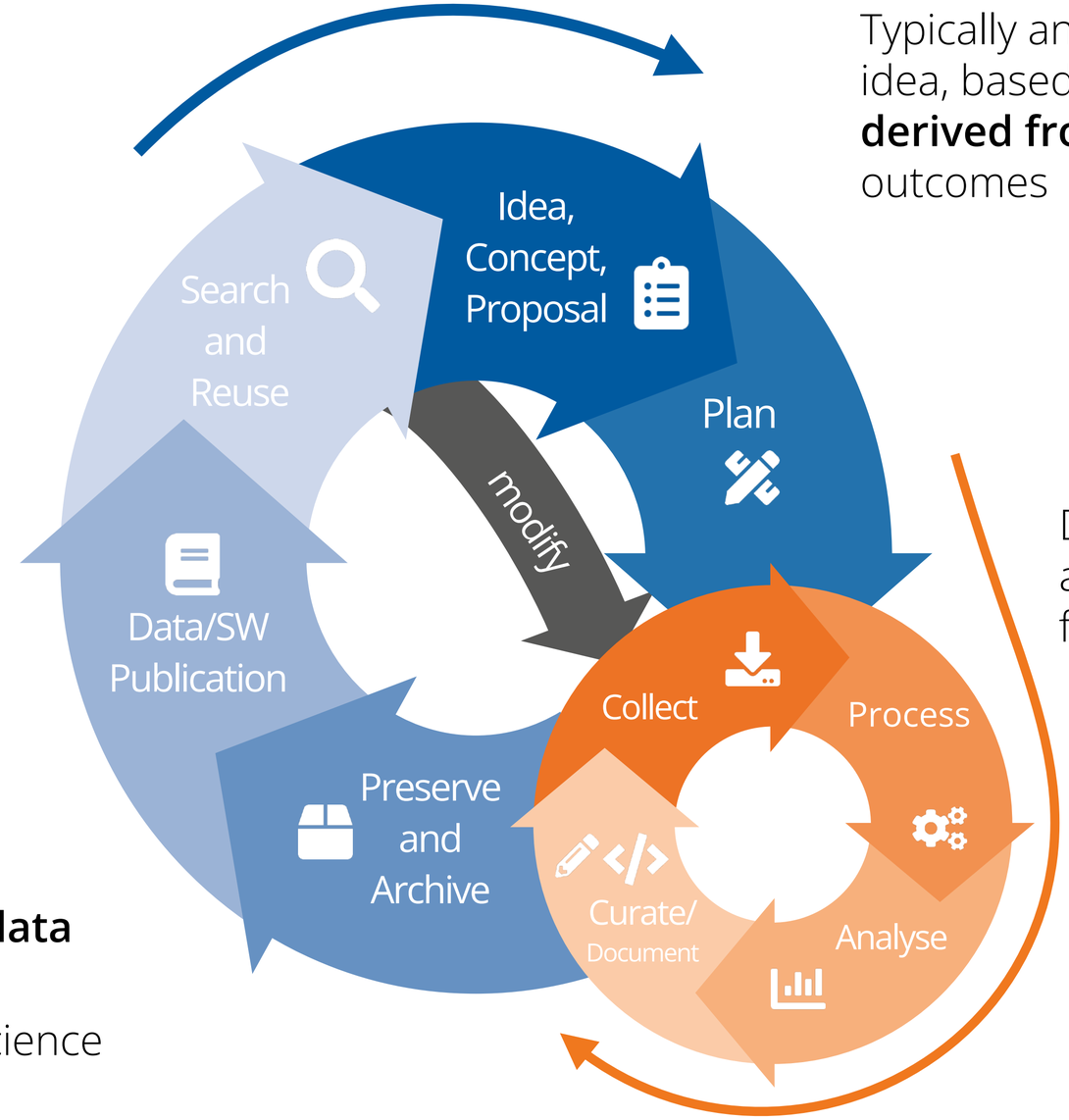
Overview of published data sets of selected large-scale research facilities of the HZDR:

- **ATHENA** — Accelerator Technology HElmholtz iNfrAstructure
- **CARBOSOLA**
- **Dresden High Magnetic Field Laboratory**
- **DRESDYN** — DREsden Sodium facility for DYNamo and thermohydraulic studies
- **ELBE** — Electron Linac for beams with high Brilliance and low Emittance
  - ELBE Beamlines
- **HECToR** — High-power ultrafast computed tomography
- **HIBEF** — Helmholtz ernational Beamline for me lds

Data Management Lifecycle

Publication of **data** and **metadata** is at the end of a typical research lifecycle and can be the starting point for new iteration cycles…

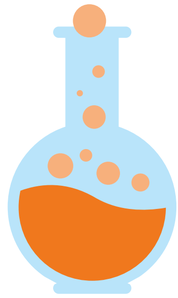
Archive or (additional) **data publication** to provide sustainable and open science



Typically an experiment starts with an idea, based on a research question, **derived from available research** outcomes

→ Valuable metadata in our proposal systems

During experiment **documentation** and **metadata** curation are essential for a **comprehensible** experiment



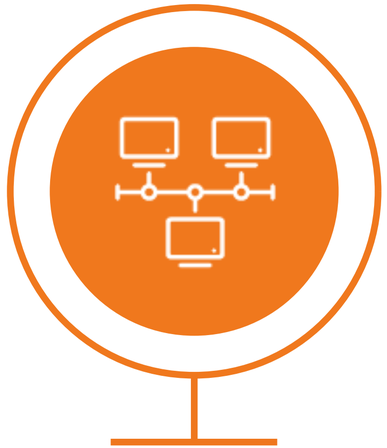
Gap Analysis of our Metadata Ecosystem

GAP

Some approach to provide additional searchable experiment- and discipline-specific metadata

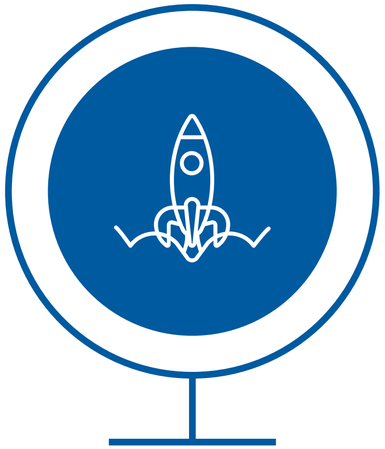
CURRENT STATE

DataCite metadata with only a few domain/experiment-specific fields (e.g. facility) in our data repository RODARE



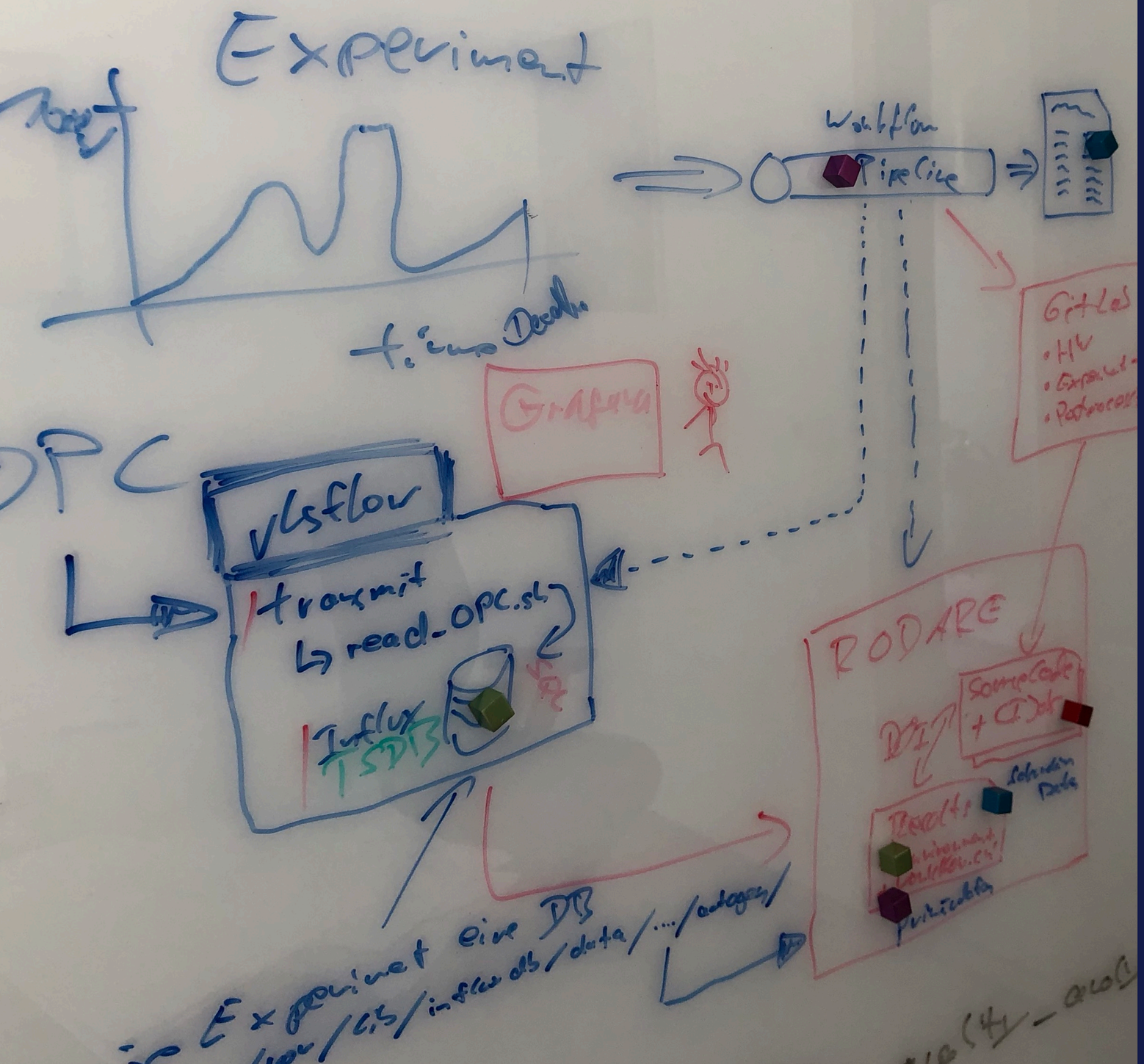
PLAN

Integrate SciCat in between RODARE and the experiment, transfer of experiment-specific metadata using automated pipelines and link the final data in RODARE, archive or filesystem



DESIRED STATE

Additional community-specific metadata to search for and reference to the datasets (in RODARE, our archive or filesystem)



Our (Meta)**Data Patchwork**

Data, Metadata and Publications at HZDR

— Data Repository RODARE for citable **data/software publications** with **DOIs**

— **SciCat** for public findable metadata describing data and it's different locations with "HZDR Handles" as **PID**:

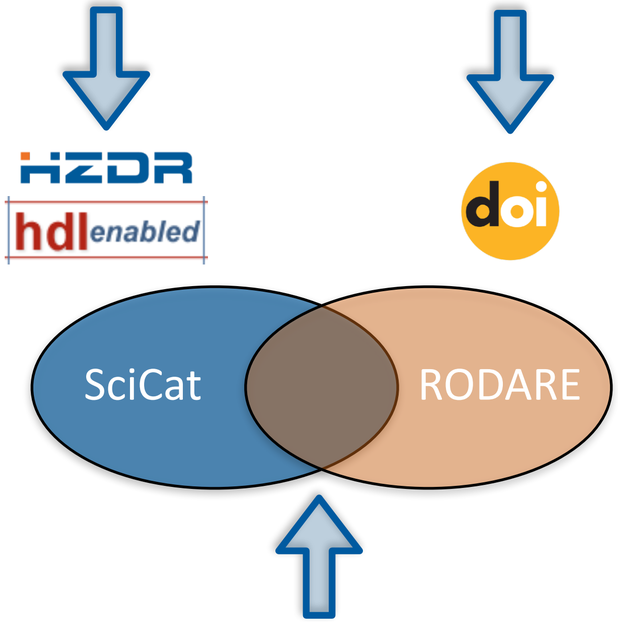
- RODARE, if the data is published and used in scientific publications
- Anywhere on our internal filesystems or cloud storage
- In our internal archive for long-term storage

SciCat metadata for describing unpublished RAW or derived data located in our filesystems or archives

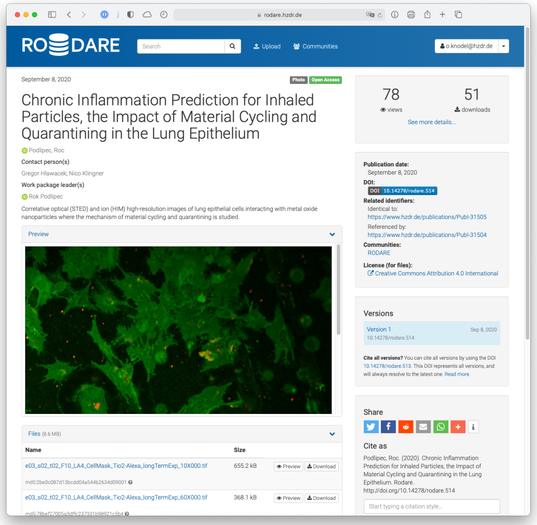
(external access can be granted)

Data and software publications in **RODARE**

(Open, closed, restricted or embargoed access)

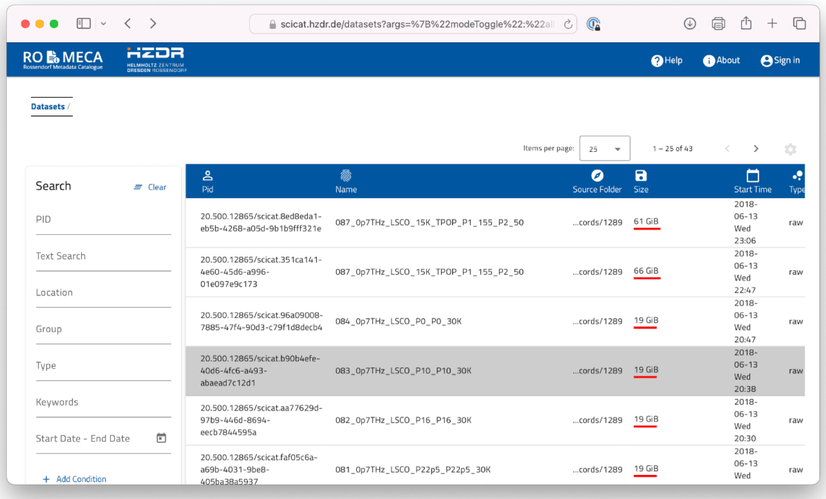


SciCat metadata describing data publication in **RODARE**



Data publication system RODARE

rodare.hzdr.de



SciCat metadata catalogue

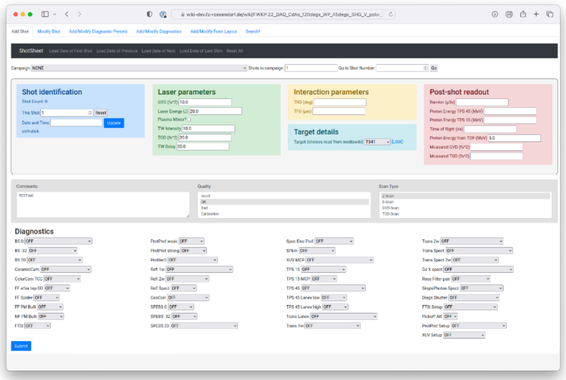
scicat.hzdr.de

DRESDEN concept HZDR

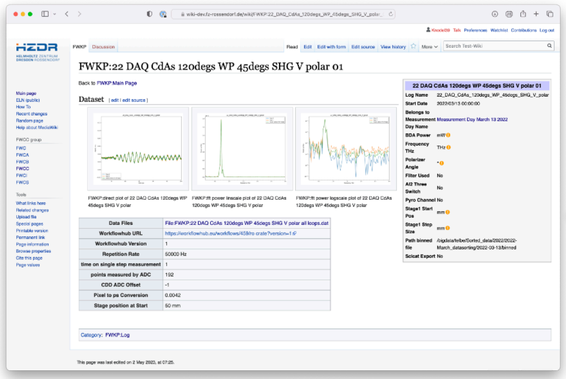
From the Experiment over the Metadata to the Publication of Data

Curated Metadata Sources

ExperimentLogging app (ExL)



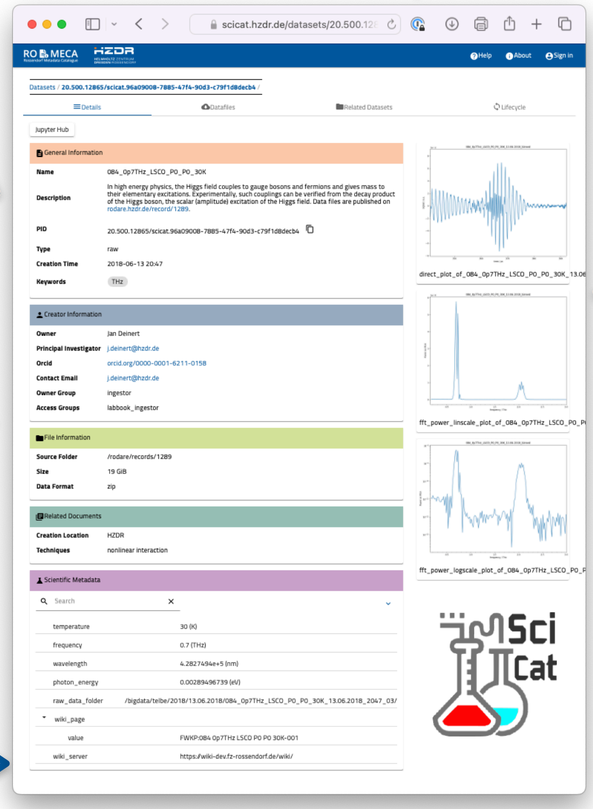
E-Logbook



Metadata from Experiment/Simulation

Public Metadata Catalogue

HZDR hdlenabled **SciCat**

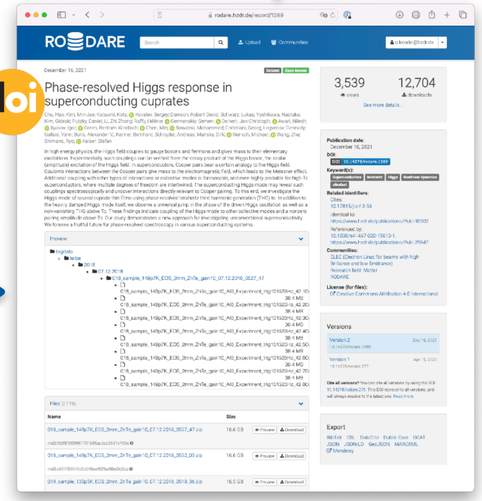


Subsequent Access to Data

panosc data portal

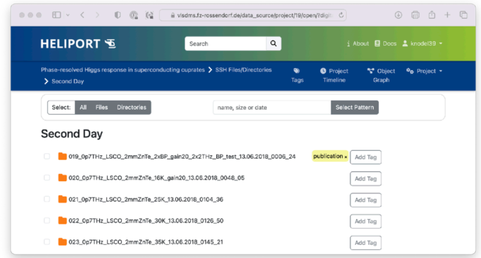
B2FIND EUDAT

RODARE (Invenio)



Dataset

HZDR hdlenabled



Filesystem

Tape Archive

Fully Automated Process for DRACO

Access: Private/Internal

Access: Public

Conclusions

- With RODARE, we have a data and software repository for all scientific research areas at the HZDR.
- RODARE only provides bibliographic metadata for description and search in the data sets.
- More detailed, experiment-specific, searchable metadata in SciCat can help scientists in their daily research.

Collect
Automated data importing

Annotate
Enrich your data

Publish
Add DOI and share

Find
Browse and search your data

SciCat

RODARE
ROSSENDORF DATA REPOSITORY

B2FIND

EUDAT

Current step: Provision of a PaNOSC search API endpoint at the HZDR!

panosc
data portal

HZDR