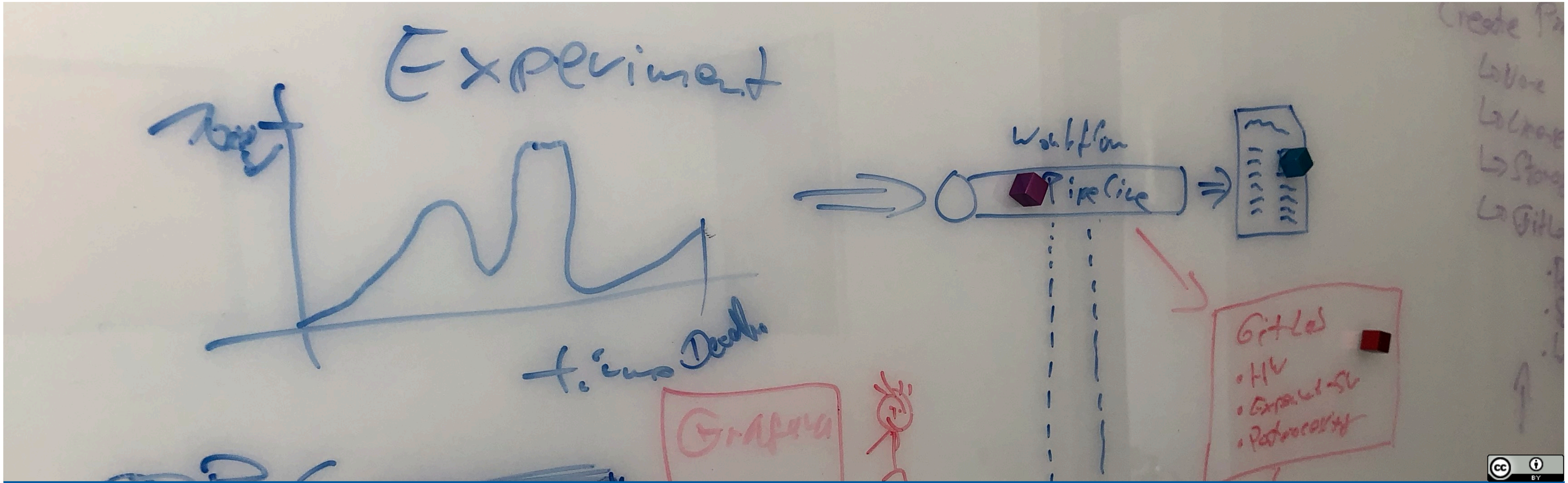


Connecting Processes to Data via Meta-Data

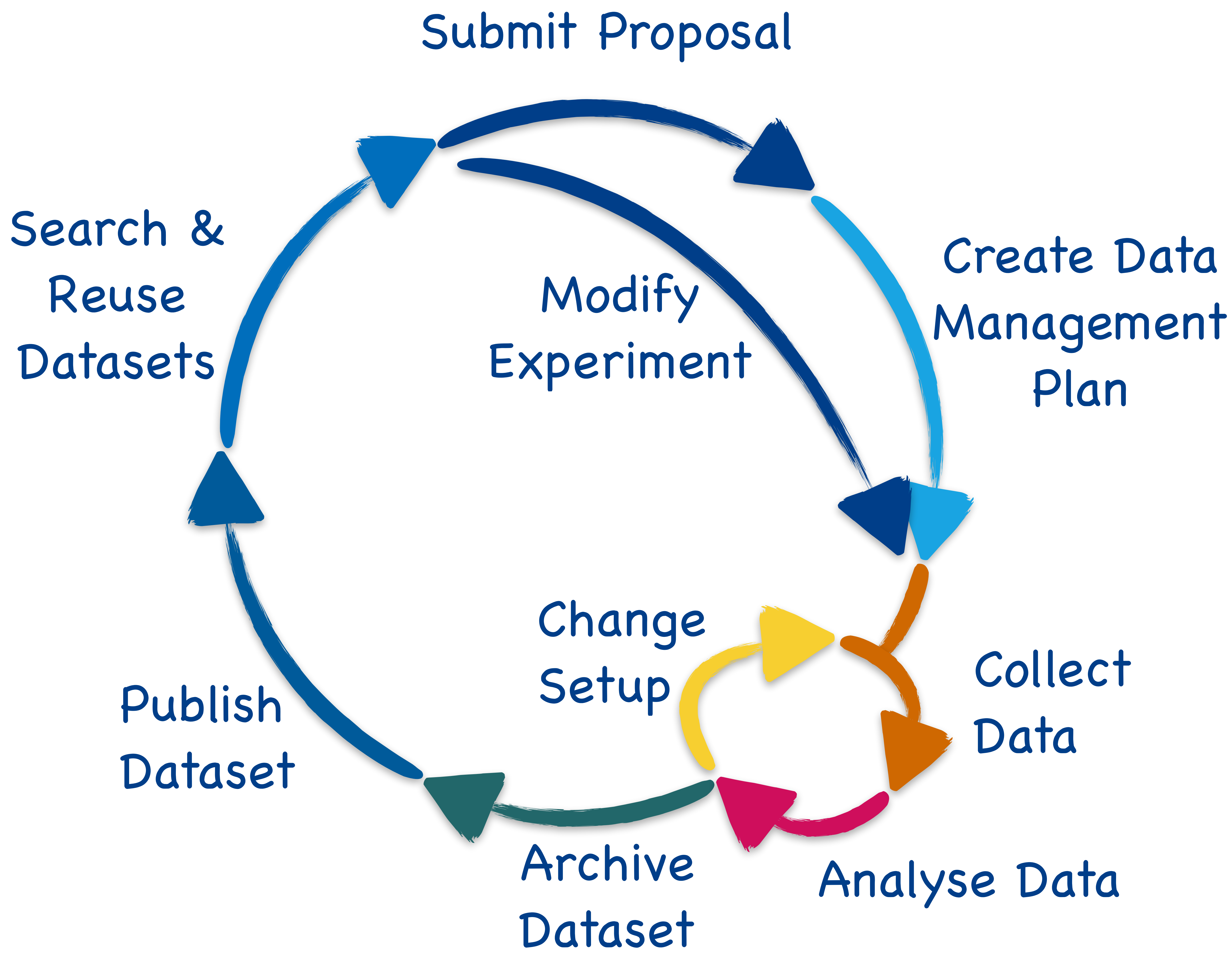
Perspectives of Data collection and Exchange

Thrill WP5 ML Workshop, HZDR, February 2024

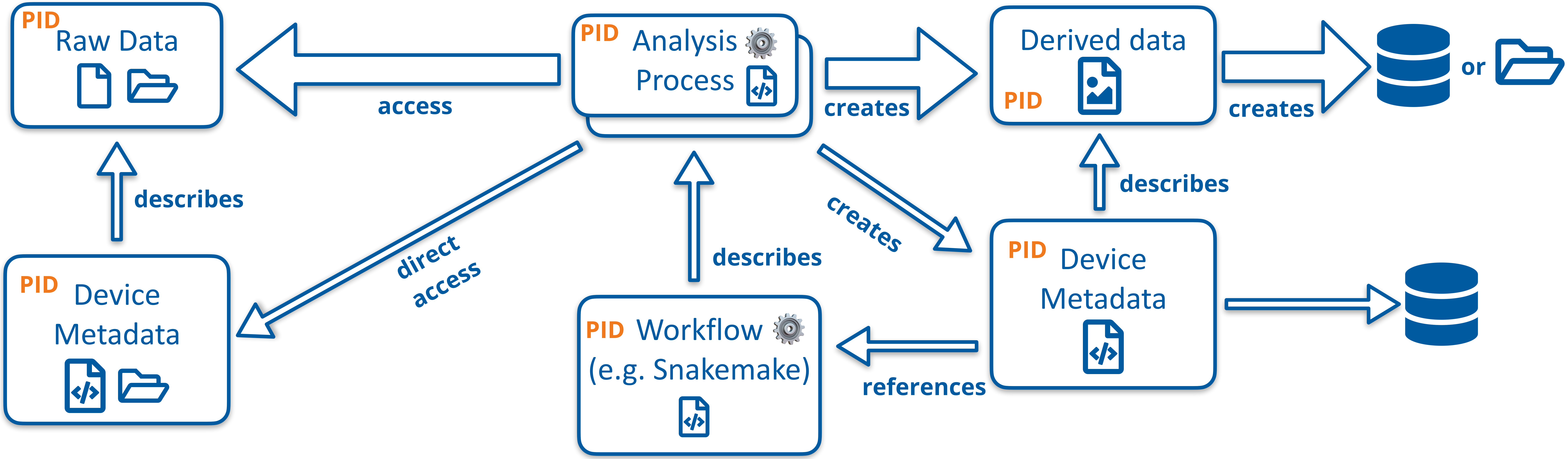
Oliver Knodel // contact: o.knodel@hzdr.de

Our Challenge: An End-to-End Digital Data Lifecycle

- We support many steps of our different research experiment (matter, energy and health) with tools:
  - Electronic lab notebook (**E-Logbook**),
  - Interactive analysis,
  - Publication of datasets,
  - Scientific workflow management,
  - Handle generation and management.
- A uniform and smooth access to and between all services and **processes** in a digital ecosystem is necessary.
- The description and interconnection between all linked resources through metadata is essential to create a **comprehensible** and **FAIR** experiment.

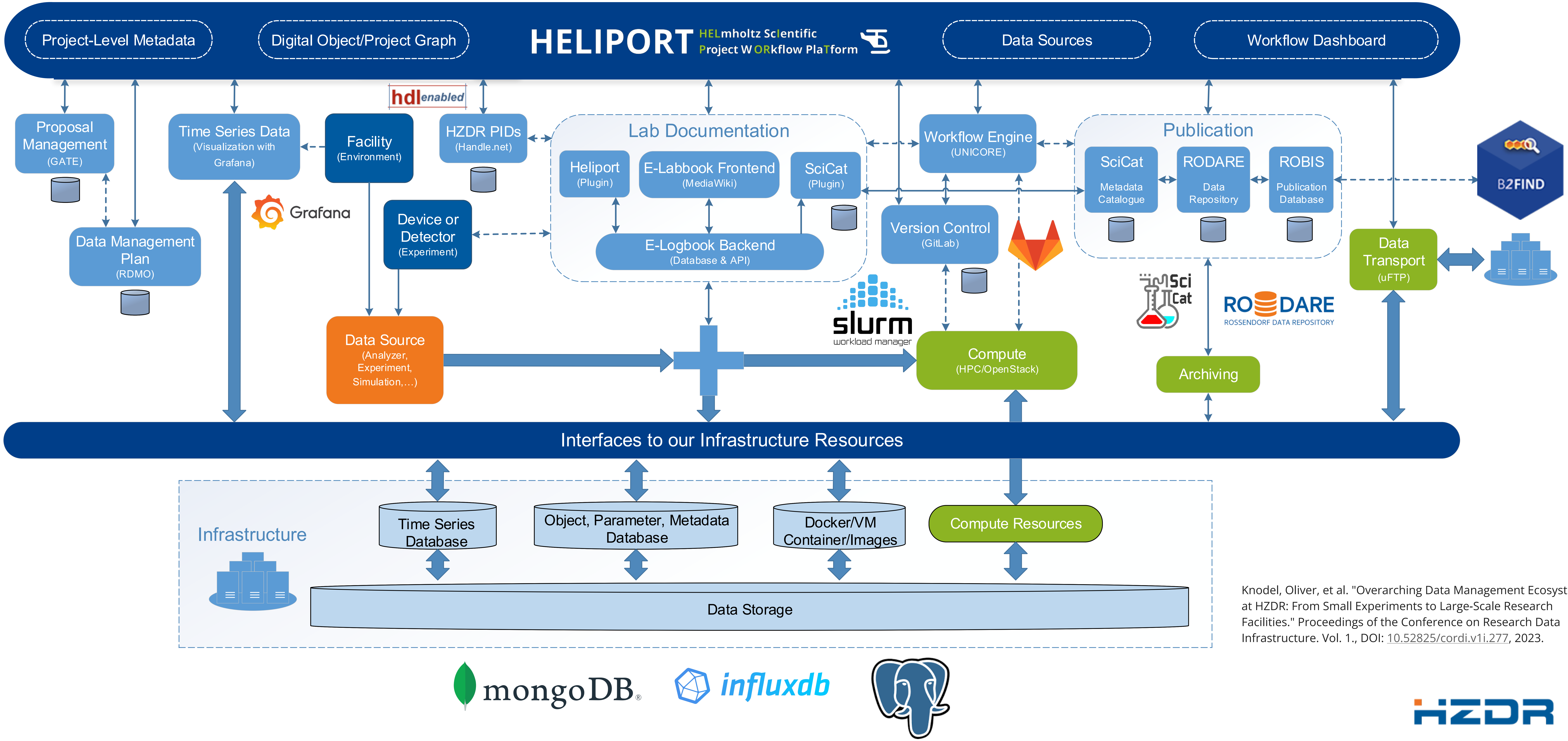


Connecting Processes to Data via Meta-Data



- The entry point is usually the raw data set that is generated from the test setup (e.g. detector, camera)
- Each data product should be described by a (standardised) metadata schema
- Process (or workflows) can access the RAW data via the corresponding metadata (with associated identifier)
- Workflows themselves should be described regarding the FAIR principles
- Derived data should contain additional descriptive metadata and an exchangeable data format (e.g. HELPMI)
- **PIDs** (e.g. Handles, DOIS) are essential to provide persistent identifier for data and metadata

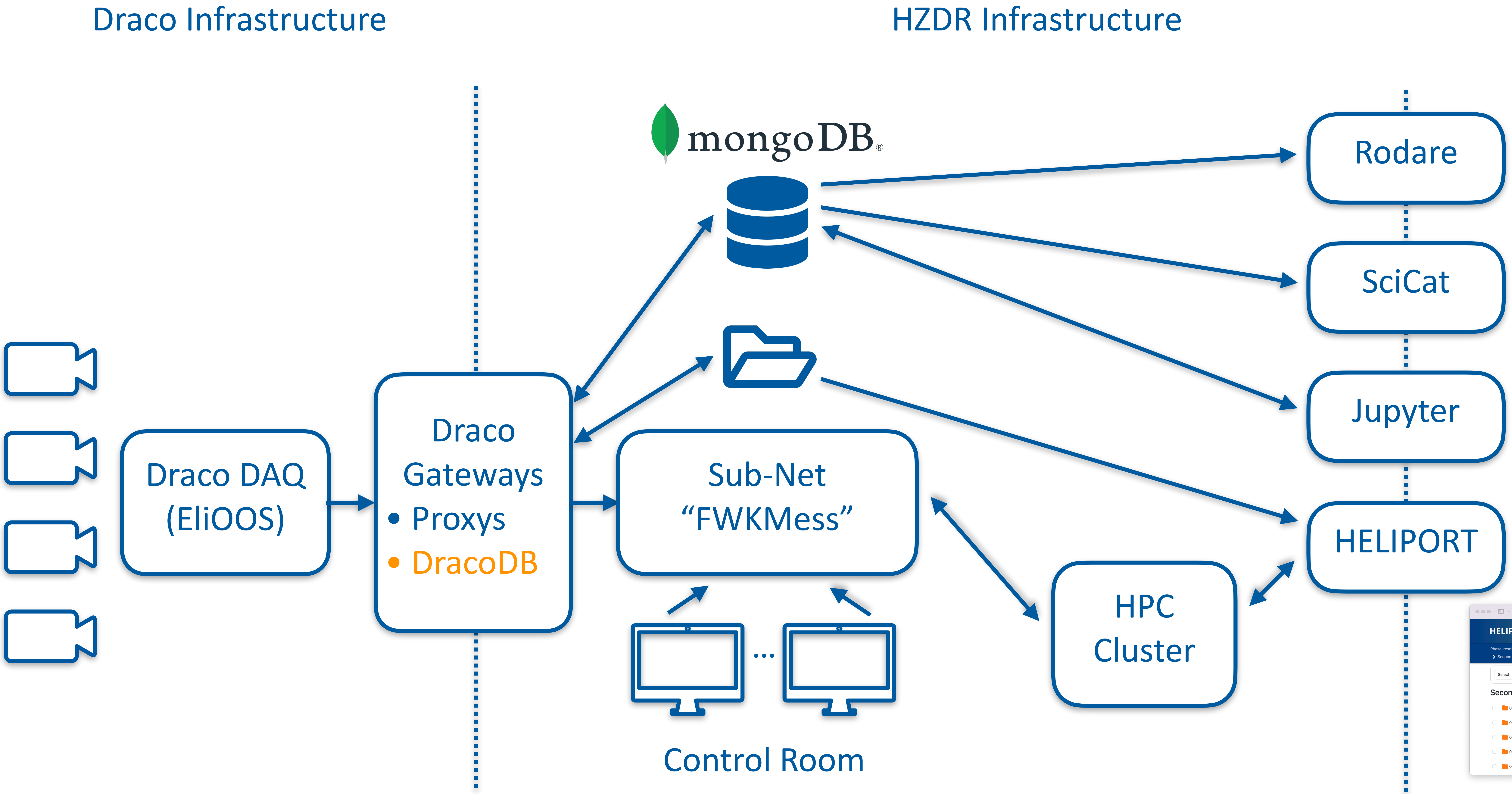
Digital Research Landscapes at HZDR

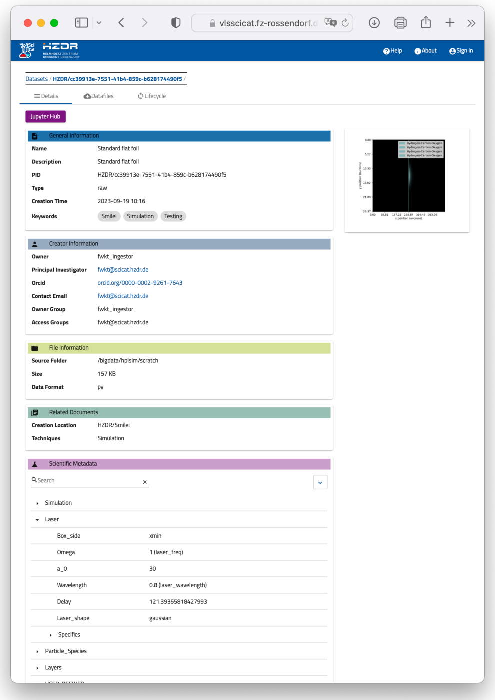


Project-Level Metadata
Digital Object/Project Graph
HELIPORT HELmholtz ScIentific Project WORkflow PlaTform
Data Sources
Workflow Dashboard
hdl enabled
Proposal Management (GATE)
Time Series Data (Visualization with Grafana)
Facility (Environment)
HZDR PIDs (Handle.net)
Lab Documentation
Heliport (Plugin)
E-Labbook Frontend (MediaWiki)
SciCat (Plugin)
E-Logbook Backend (Database & API)
Workflow Engine (UNICORE)
Version Control (GitLab)
Publication
SciCat Metadata Catalogue
RODARE Data Repository
ROBIS Publication Database
B2FIND
Grafana
Data Management Plan (RDMO)
Device or Detector (Experiment)
Data Source (Analyzer, Experiment, Simulation,…)
slurm workload manager
Compute (HPC/OpenStack)
SciCat
RODARE ROSSENDORF DATA REPOSITORY
Archiving
Data Transport (uFTP)
Interfaces to our Infrastructure Resources
Infrastructure
Time Series Database
Object, Parameter, Metadata Database
Docker/VM Container/Images
Compute Resources
Data Storage
mongoDB
influxdb

Knodel, Oliver, et al. "Overarching Data Management Ecosystem at HZDR: From Small Experiments to Large-Scale Research Facilities." Proceedings of the Conference on Research Data Infrastructure. Vol. 1., DOI: 10.52825/cordi.v1i.277, 2023.

Draco Data Flow for Advanced Data Management





Datasets / HZDR/cc39913e-7551-41b4-859c-b628174490f5 /

General Information

Name	Standard flat foil
Description	Standard flat foil
PID	HZDR/cc39913e-7551-41b4-859c-b628174490f5
Type	raw
Creation Time	2023-09-19 10:16
Keywords	Smilei Simulation Testing

Creator Information

Owner	fwkt_ingestor
Principal Investigator	fwkt@scicat.hzdr.de
Orcid	orcid.org/0000-0002-9261-7643
Contact Email	fwkt@scicat.hzdr.de
Owner Group	fwkt_ingestor
Access Groups	fwkt@scicat.hzdr.de

File Information

Source Folder	/bigdata/hplsim/scratch
Size	157 KB
Data Format	py

Related Documents

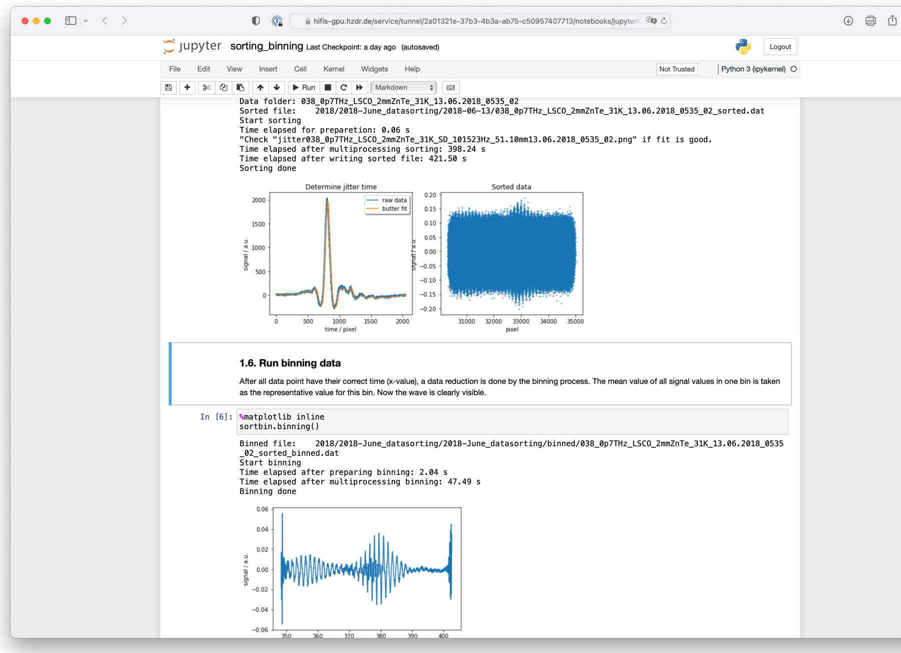
Creation Location	HZDR/Smilei
Techniques	Simulation

Scientific Metadata

- Simulation
- Laser

Box_side	xmin
Omega	1 (laser_freq)
a_0	30
Wavelength	0.8 (laser_wavelength)
Delay	121.39355818427993
Laser_shape	gaussian

- Specifics
- Particle_Species
- Layers



sorting_binning Last Checkpoint: a day ago (autosaved)

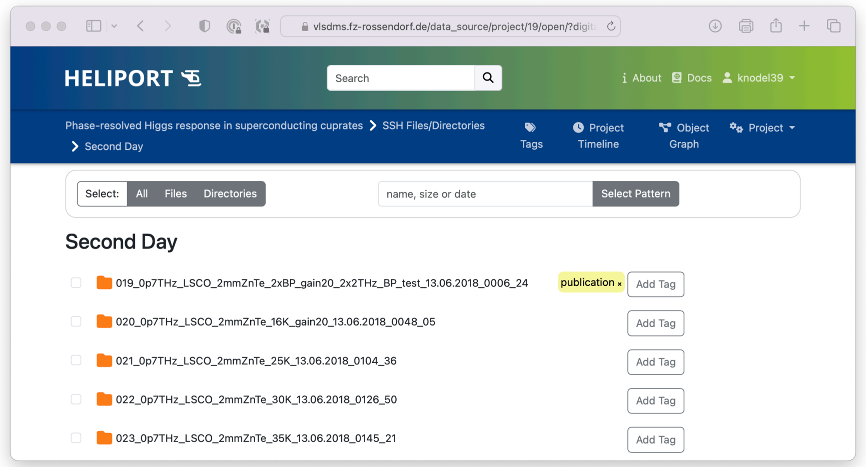
```
Data folder: 038_0p7THz_LSCO_2mmZnTe_31K_13.06.2018_0535_02
Sorted file:    2018/2018-June_datasorting/2018-06-13/038_0p7THz_LSCO_2mmZnTe_31K_13.06.2018_0535_02_sorted.dat
Start sorting
Time elapsed for preparetion: 0.06 s
"Check "jitter038_0p7THz_LSCO_2mmZnTe_31K_SD_101523Hz_51.10mm13.06.2018_0535_02.png" if fit is good.
Time elapsed after multiprocessing sorting: 398.24 s
Time elapsed after writing sorted file: 421.50 s
Sorting done
```

1.6. Run binning data

After all data point have their correct time (x-value), a data reduction is done by the binning process. The mean value of all signal values in one bin is taken as the representative value for this bin. Now the wave is clearly visible.

```
In [6]: %matplotlib inline
        sortbin.binning()
```

```
Binned file:    2018/2018-June_datasorting/2018-June_datasorting/binned/038_0p7THz_LSCO_2mmZnTe_31K_13.06.2018_0535
_02_sorted_binned.dat
Start binning
Time elapsed after preparing binning: 2.04 s
Time elapsed after multiprocessing binning: 47.49 s
Binning done
```

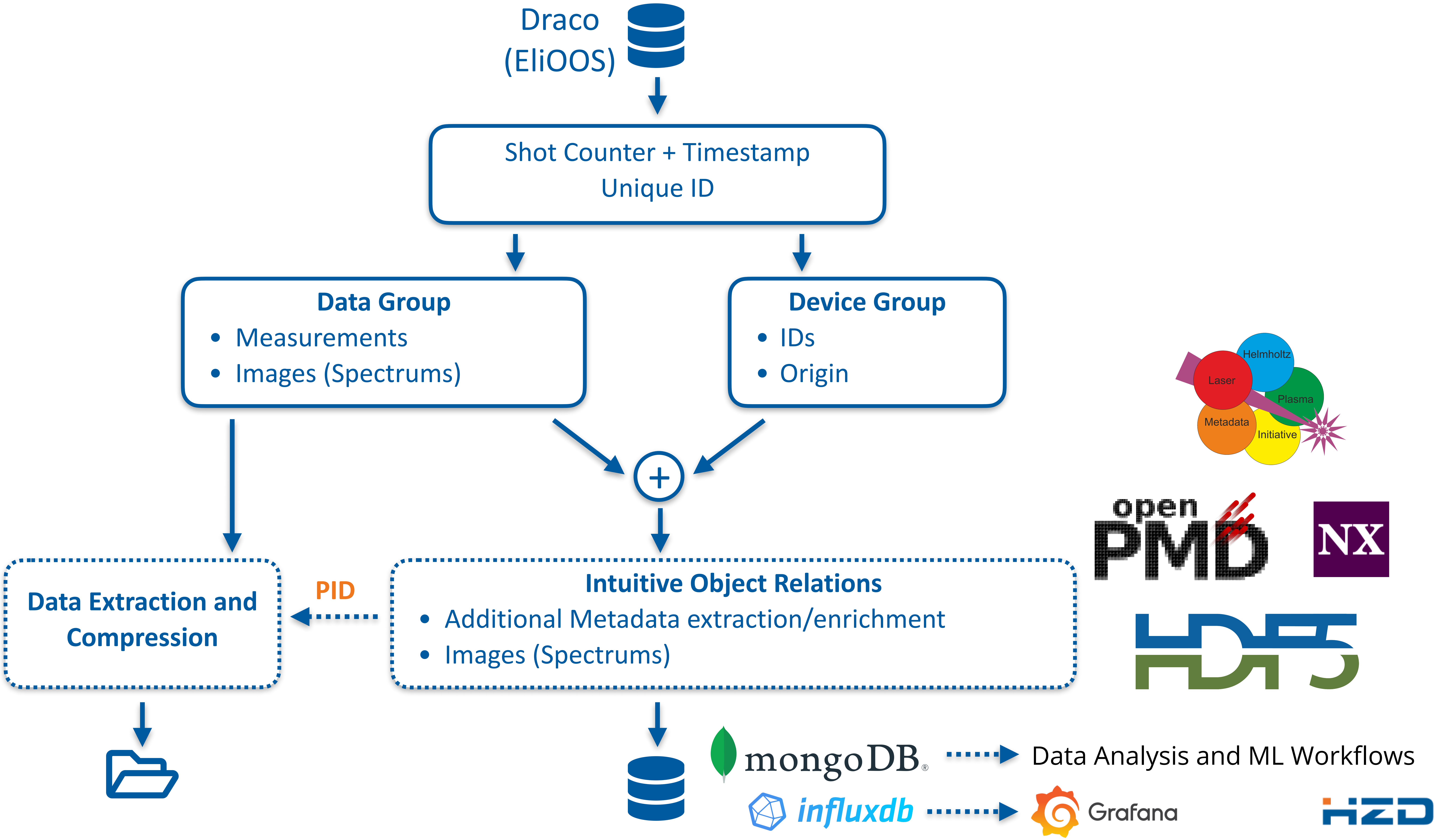


Phase-resolved Higgs response in superconducting cuprates > SSH Files/Directories > Second Day

Second Day

- 019_0p7THz_LSCO_2mmZnTe_2xBP_gain20_2x2THz_BP_test_13.06.2018_0006_24 publication ×
- 020_0p7THz_LSCO_2mmZnTe_16K_gain20_13.06.2018_0048_05
- 021_0p7THz_LSCO_2mmZnTe_25K_13.06.2018_0104_36
- 022_0p7THz_LSCO_2mmZnTe_30K_13.06.2018_0126_50
- 023_0p7THz_LSCO_2mmZnTe_35K_13.06.2018_0145_21

Extended Draco Data Pipeline (to fully take advantage of the HZDR infrastructure)



Draco (EliOOS)

Shot Counter + Timestamp
Unique ID

Data Group
- Measurements
- Images (Spectrums)

Device Group
- IDs
- Origin

Data Extraction and Compression

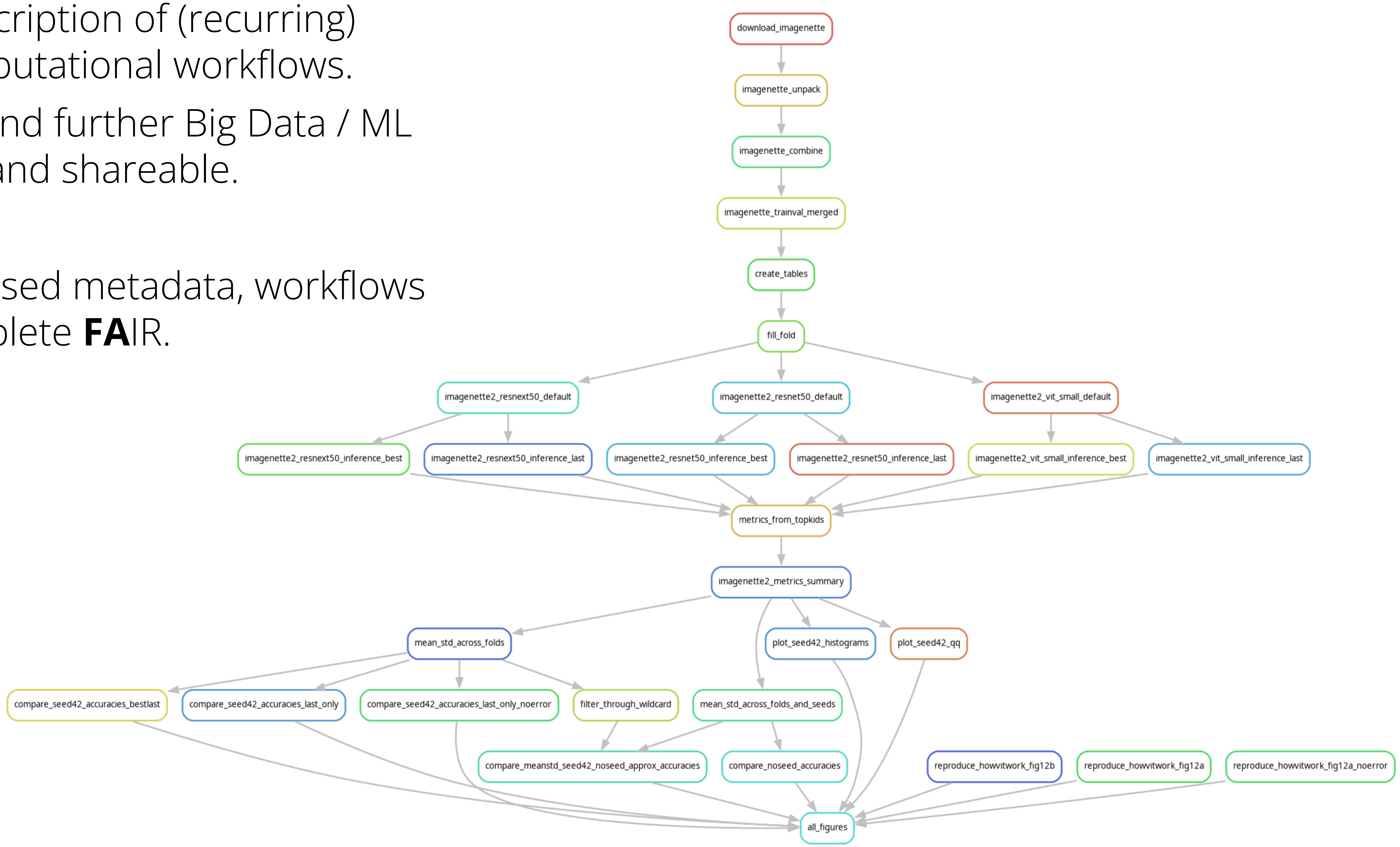
PID

Intuitive Object Relations
- Additional Metadata extraction/enrichment
- Images (Spectrums)

Data Analysis and ML Workflows

Computational Workflows (Processes) and Data Management (MetaData)

- In our HZDR infrastructure, the description of (recurring) work can be automatised with computational workflows.
- Workflows enable deeper insights and further Big Data / ML methods that are comprehensible and shareable.
- Workflows enable FA**IR**.
- With interchangeable, standards-based metadata, workflows can be used in different RIs to complete **FA**IR.

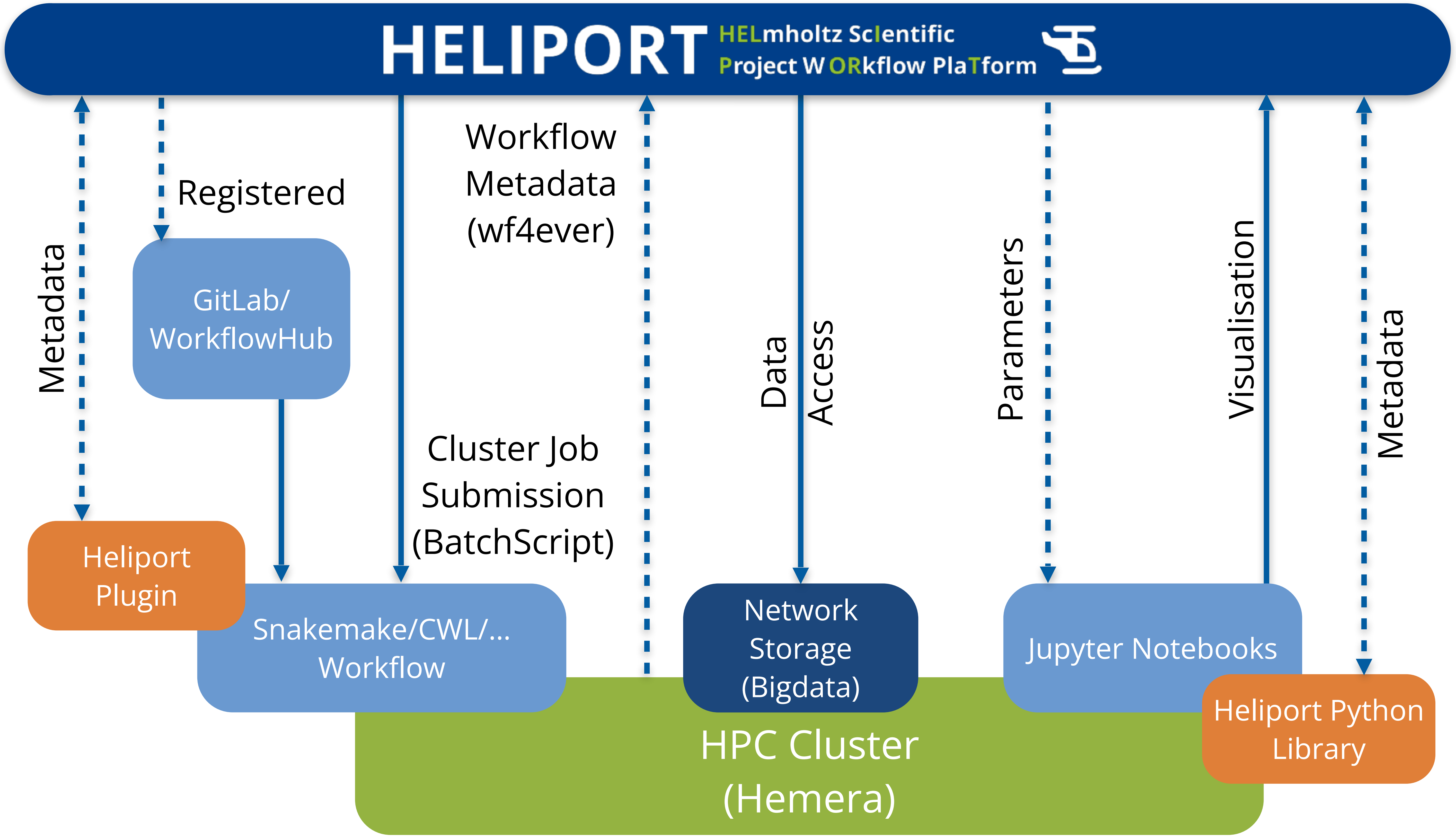


download_imagenette
imagenette_unpack
imagenette_combine
imagenette_trainval_merged
create_tables
fill_fold
imagenette2_resnext50_default
imagenette2_resnet50_default
imagenette2_vit_small_default
imagenette2_resnext50_inference_best
imagenette2_resnext50_inference_last
imagenette2_resnet50_inference_best
imagenette2_resnet50_inference_last
imagenette2_vit_small_inference_best
imagenette2_vit_small_inference_last
metrics_from_topkids
imagenette2_metrics_summary
mean_std_across_folds
plot_seed42_histograms
plot_seed42_qq
compare_seed42_accuracies_bestlast
compare_seed42_accuracies_last_only
compare_seed42_accuracies_last_only_noerror
filter_through_wildcard
mean_std_across_folds_and_seeds
compare_meanstd_seed42_noseed_approx_accuracies
compare_noseed_accuracies
reproduce_howitwork_fig12b
reproduce_howitwork_fig12a
reproduce_howitwork_fig12a_noerror
all_figures

DAG of an ML Workflow in Snakemake based on “Machine Learning State-of-the-Art with Uncertainties”; DOI: 2204.05173

Workflow Architecture at HZDR (in development)

- HELIPORT offers an infrastructure which permits the integration of various workflow languages and access modes to HPC infrastructures.
- The infrastructure keeps track of and collects the metadata and enables access to all resources involved.
- Next steps:
  - Python library sending workflow information directly to HELIPORT,
  - Provision of provenance information from Jupyter notebooks,
  - Use case: **PIConGPU**



HELIPORT HELmholtz ScIentific Project WORkflow PlaTform

Metadata

Registered

GitLab/ WorkflowHub

Workflow Metadata (wf4ever)

Cluster Job Submission (BatchScript)

Data Access

Parameters

Visualisation

Metadata

Heliport Plugin

Snakemake/CWL/… Workflow

Network Storage (Bigdata)

Jupyter Notebooks

Heliport Python Library

HPC Cluster (Hemera)

HELIPORT HELmholtz ScIentific Project WORkflow PlaTform

heliport.hzdr.de

> The HELIPORT project aims at developing a platform which accommodates the **complete life cycle** of a scientific project and links all corresponding programs, systems and workflows to create a more **FAIR** and comprehensible project description using **APIs**.

Project Members:

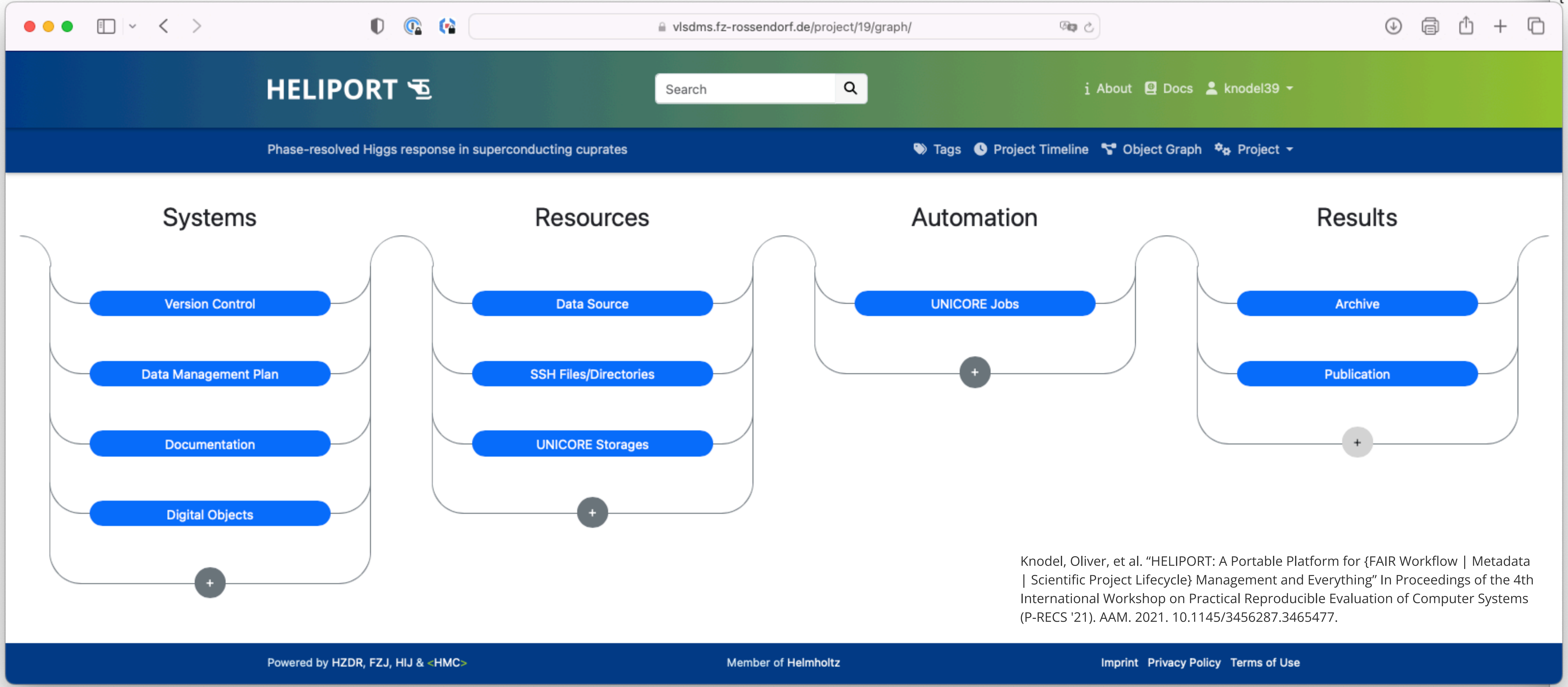
HZDR HELMHOLTZ ZENTRUM DRESDEN ROSSENDORF

HI JENA Helmholtz Institute Jena

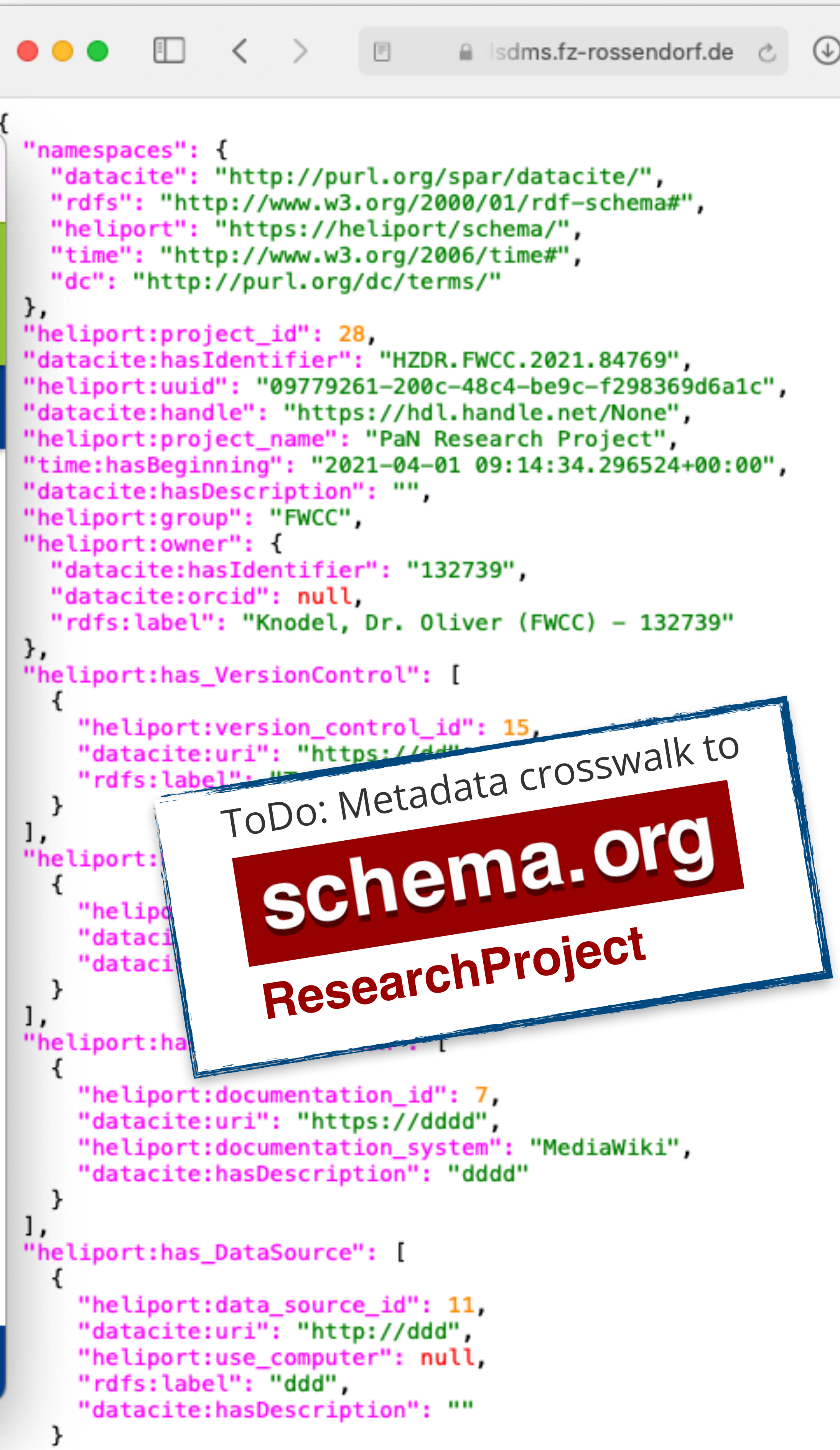
JÜLICH FORSCHUNGSZENTRUM

Funded by:

<HMC> HELMHOLTZ Metadata Collaboration



Knodel, Oliver, et al. "HELIPORT: A Portable Platform for {FAIR Workflow | Metadata | Scientific Project Lifecycle} Management and Everything" In Proceedings of the 4th International Workshop on Practical Reproducible Evaluation of Computer Systems (P-RECS '21). AAM. 2021. 10.1145/3456287.3465477.



```
{
  "namespaces": {
    "datacite": "http://purl.org/spar/datacite/",
    "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
    "heliport": "https://heliport/schema/",
    "time": "http://www.w3.org/2006/time#",
    "dc": "http://purl.org/dc/terms/"
  },
  "heliport:project_id": 28,
  "datacite:hasIdentifier": "HZDR.FWCC.2021.84769",
  "heliport:uuid": "09779261-200c-48c4-be9c-f298369d6a1c",
  "datacite:handle": "https://hdl.handle.net/None",
  "heliport:project_name": "PaN Research Project",
  "time:hasBeginning": "2021-04-01 09:14:34.296524+00:00",
  "datacite:hasDescription": "",
  "heliport:group": "FWCC",
  "heliport:owner": {
    "datacite:hasIdentifier": "132739",
    "datacite:orcid": null,
    "rdfs:label": "Knodel, Dr. Oliver (FWCC) - 132739"
  },
  "heliport:has_VersionControl": [
    {
      "heliport:version_control_id": 15,
      "datacite:uri": "https://dd
      "rdfs:label": "
    }
  ],
  "heliport:
    {
      "helipo
      "dataci
      "dataci
    }
  ],
  "heliport:ha
    {
      "heliport:documentation_id": 7,
      "datacite:uri": "https://dddd",
      "heliport:documentation_system": "MediaWiki",
      "datacite:hasDescription": "dddd"
    }
  ],
  "heliport:has_DataSource": [
    {
      "heliport:data_source_id": 11,
      "datacite:uri": "http://ddd",
      "heliport:use_computer": null,
      "rdfs:label": "ddd",
      "datacite:hasDescription": ""
    }
```

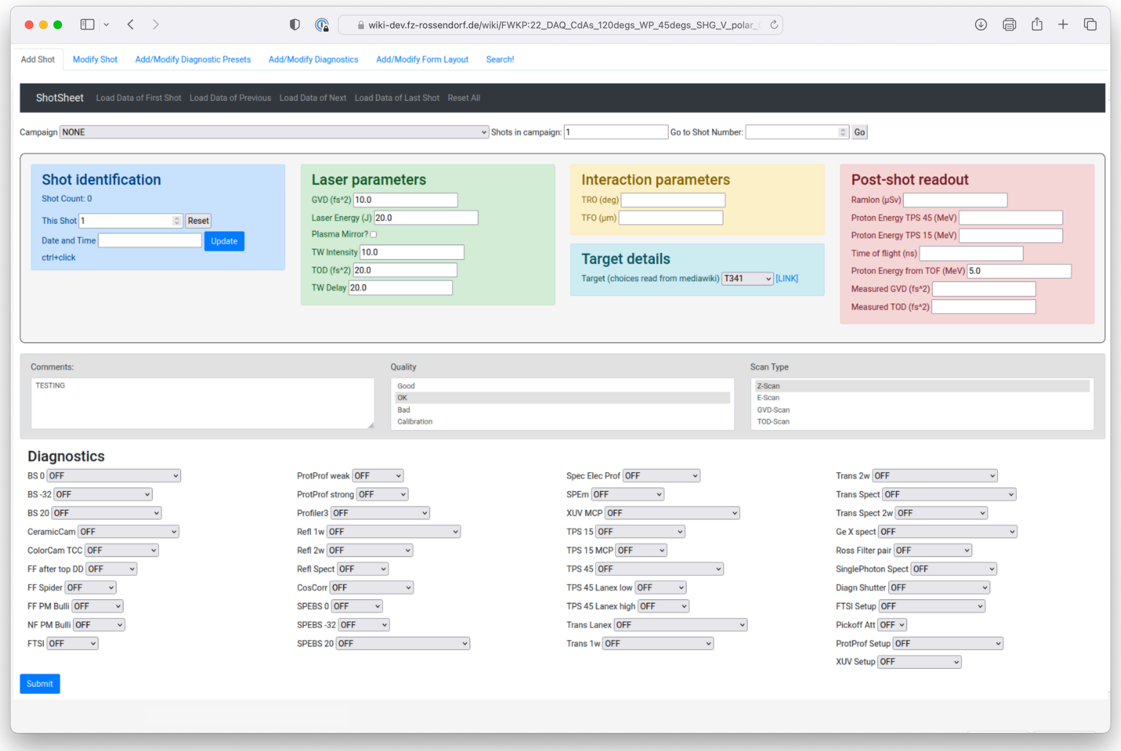
ToDo: Metadata crosswalk to **schema.org** **ResearchProject**

Metadata Catalogue SciCat and Data Repository RODARE (Draft)

Metadata from Experiment/Simulation

Curated Metadata Source

ExperimentLogging app (ExL)



wiki-dev.fz-rossendorf.de/wiki/FWKP:22_DAQ_CdAs_120degs_WP_45degs_SHG_V_polar_

Add Shot | Modify Shot | Add/Modify Diagnostic Presets | Add/Modify Diagnostics | Add/Modify Form Layout | Search!

ShotSheet | Load Data of First Shot | Load Data of Previous | Load Data of Next | Load Data of Last Shot | Reset All

Campaign: NONE | Shots in campaign: 1 | Go to Shot Number: | Go

Shot identification
Shot Count: 0
This Shot: 1 | Reset
Date and Time: | Update
ctrl+click

Laser parameters
GVD (fs^2): 10.0
Laser Energy (J): 20.0
Plasma Mirror? □
TW Intensity: 10.0
TOD (fs^2): 20.0
TW Delay: 20.0

Interaction parameters
TRO (deg):
TFO (µm):

Target details
Target (choices read from mediawiki): T341 [LINK]

Post-shot readout
Ramlon (µSv):
Proton Energy TPS 45 (MeV):
Proton Energy TPS 15 (MeV):
Time of flight (ns):
Proton Energy from TOF (MeV): 5.0
Measured GVD (fs^2):
Measured TOD (fs^2):

Comments: TESTING

Quality: Good / OK / Bad / Calibration

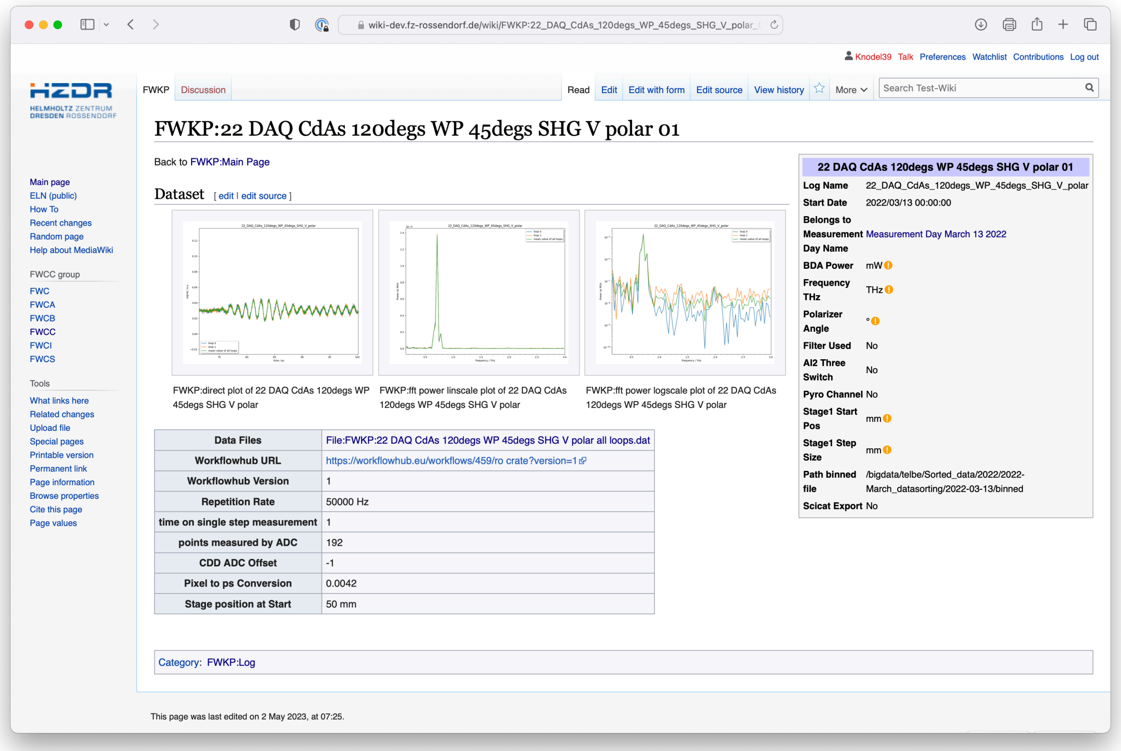
Scan Type: Z-Scan / E-Scan / GVD-Scan / TOD-Scan

Diagnostics

BS 0	OFF	ProtProf weak	OFF	Spec Elec Prof	OFF	Trans 2w	OFF
BS -32	OFF	ProtProf strong	OFF	SPEm	OFF	Trans Spect	OFF
BS 20	OFF	Profiler3	OFF	XUV MCP	OFF	Trans Spect 2w	OFF
CeramicCam	OFF	Refl 1w	OFF	TPS 15	OFF	Ge X spect	OFF
ColorCam TCC	OFF	Refl 2w	OFF	TPS 15 MCP	OFF	Ross Filter pair	OFF
FF after top DD	OFF	Refl Spect	OFF	TPS 45	OFF	SinglePhoton Spect	OFF
FF Spider	OFF	CosCorr	OFF	TPS 45 Lanex low	OFF	Diagn Shutter	OFF
FF PM Bulli	OFF	SPEBS 0	OFF	TPS 45 Lanex high	OFF	FTSI Setup	OFF
NF PM Bulli	OFF	SPEBS -32	OFF	Trans Lanex	OFF	Pickoff Att	OFF
FTSI	OFF	SPEBS 20	OFF	Trans 1w	OFF	ProtProf Setup	OFF
						XUV Setup	OFF

Submit

E-Logbook



wiki-dev.fz-rossendorf.de/wiki/FWKP:22_DAQ_CdAs_120degs_WP_45degs_SHG_V_polar_

Knodel39 Talk Preferences Watchlist Contributions Log out

FWKP | Discussion | Read | Edit | Edit with form | Edit source | View history | More | Search Test-Wiki

Main page · ELN (public) · How To · Recent changes · Random page · Help about MediaWiki
FWCC group: FWC · FWCA · FWCB · FWCC · FWCI · FWCS
Tools: What links here · Related changes · Upload file · Special pages · Printable version · Permanent link · Page information · Browse properties · Cite this page · Page values

FWKP:22 DAQ CdAs 120degs WP 45degs SHG V polar 01

Back to FWKP:Main Page

Dataset [edit | edit source]

FWKP:direct plot of 22 DAQ CdAs 120degs WP 45degs SHG V polar

FWKP:fft power linscale plot of 22 DAQ CdAs 120degs WP 45degs SHG V polar

FWKP:fft power logscale plot of 22 DAQ CdAs 120degs WP 45degs SHG V polar

Data Files	File:FWKP:22 DAQ CdAs 120degs WP 45degs SHG V polar all loops.dat
Workflowhub URL	https://workflowhub.eu/workflows/459/ro crate?version=1
Workflowhub Version	1
Repetition Rate	50000 Hz
time on single step measurement	1
points measured by ADC	192
CDD ADC Offset	-1
Pixel to ps Conversion	0.0042
Stage position at Start	50 mm

22 DAQ CdAs 120degs WP 45degs SHG V polar 01	
Log Name	22_DAQ_CdAs_120degs_WP_45degs_SHG_V_polar
Start Date	2022/03/13 00:00:00
Belongs to Measurement Day Name	Measurement Day March 13 2022
BDA Power	mW
Frequency THz	THz
Polarizer Angle	°
Filter Used	No
Al2 Three Switch	No
Pyro Channel	No
Stage1 Start Pos	mm
Stage1 Step Size	mm
Path binned file	/bigdata/telbe/Sorted_data/2022/2022-March_datasorting/2022-03-13/binned
Scicat Export	No

Category: FWKP:Log

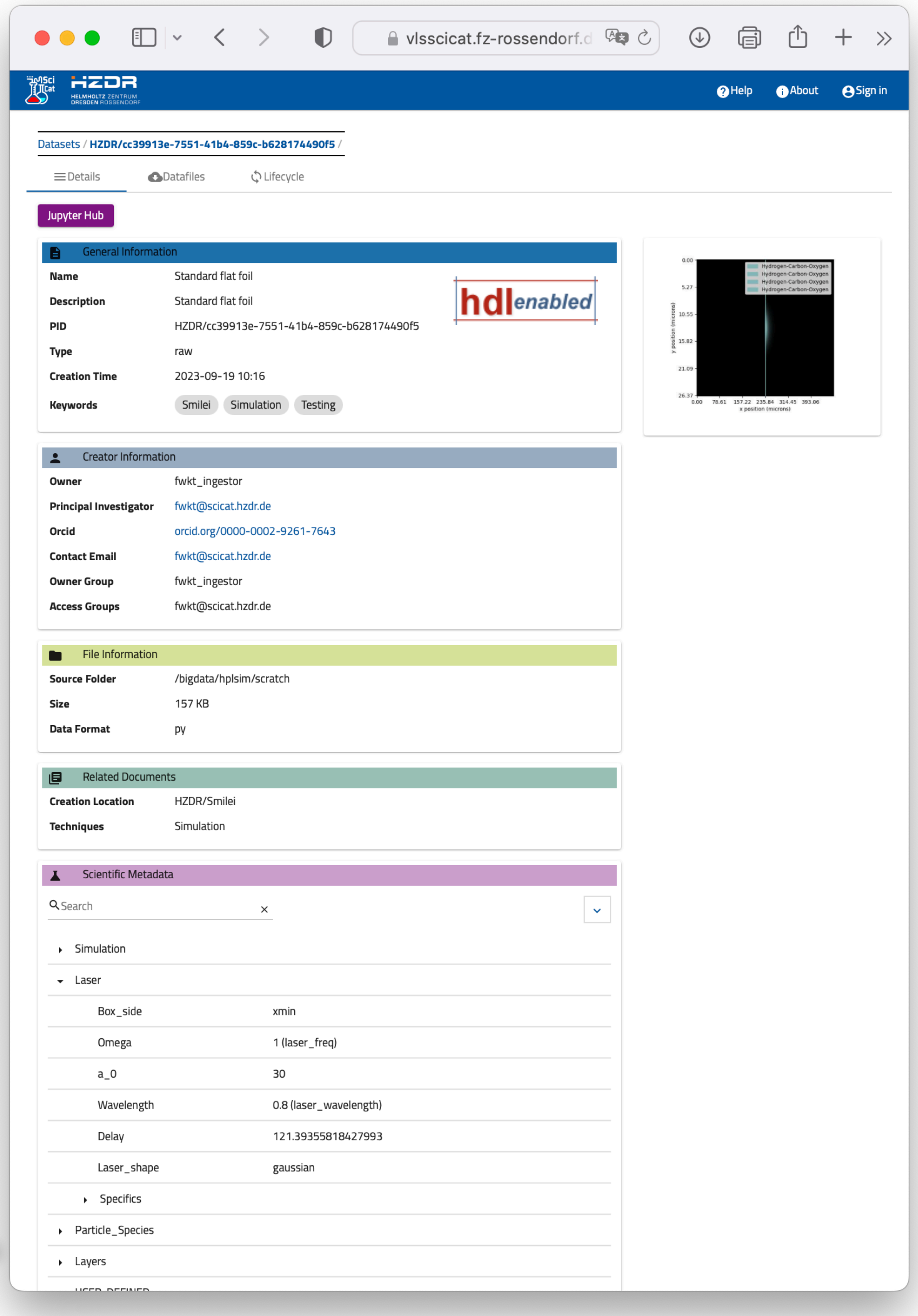
This page was last edited on 2 May 2023, at 07:25.

Fully Automated Process for DRACO

Public Metadata Catalogue

DAPHNE 4NFDI

SciCat



vlsscicat.fz-rossendorf.d

SciCat HZDR Helmholtz Zentrum Dresden Rossendorf | Help | About | Sign in

Datasets / HZDR/cc39913e-7551-41b4-859c-b628174490f5 /

Details | Datafiles | Lifecycle

Jupyter Hub

General Information

Name	Standard flat foil
Description	Standard flat foil
PID	HZDR/cc39913e-7551-41b4-859c-b628174490f5
Type	raw
Creation Time	2023-09-19 10:16
Keywords	Smilei Simulation Testing

hdl enabled

Creator Information

Owner	fwkt_ingestor
Principal Investigator	fwkt@scicat.hzdr.de
Orcid	orcid.org/0000-0002-9261-7643
Contact Email	fwkt@scicat.hzdr.de
Owner Group	fwkt_ingestor
Access Groups	fwkt@scicat.hzdr.de

File Information

Source Folder	/bigdata/hplsim/scratch
Size	157 KB
Data Format	py

Related Documents

Creation Location	HZDR/Smilei
Techniques	Simulation

Scientific Metadata

Search

- Simulation
- Laser

Box_side	xmin
Omega	1 (laser_freq)
a_0	30
Wavelength	0.8 (laser_wavelength)
Delay	121.39355818427993
Laser_shape	gaussian

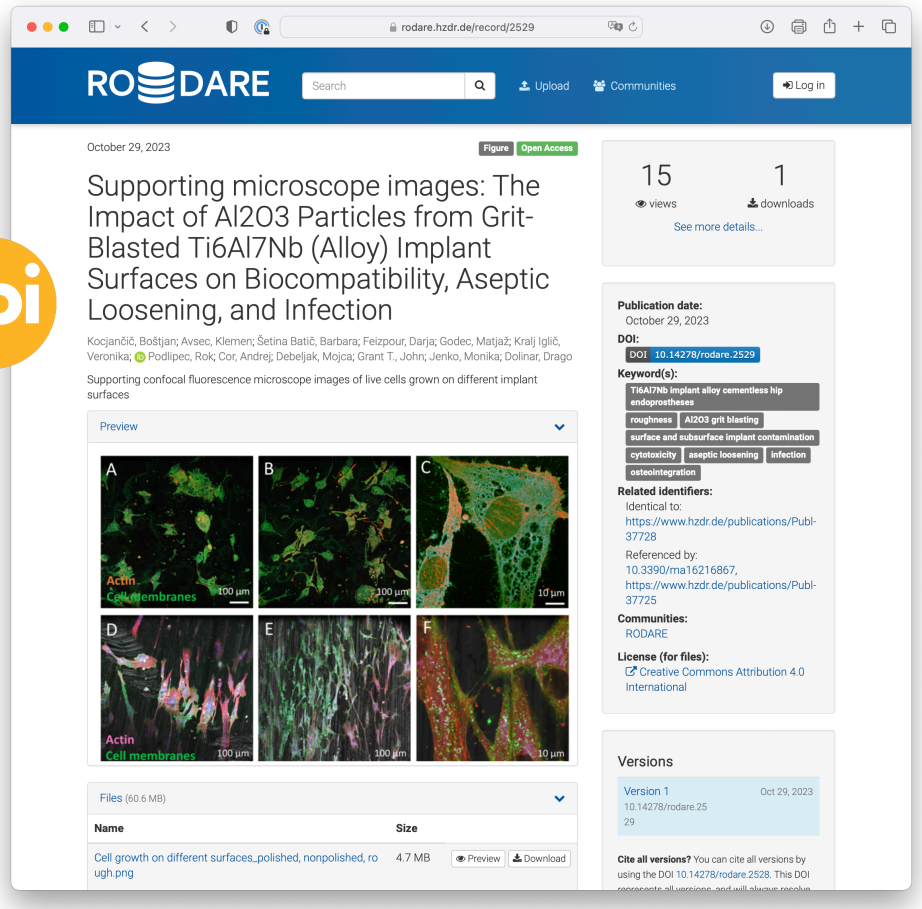
- Specifics
- Particle_Species
- Layers

Subsequent Access to Data

Findable Accessible Interoperable Reusable

RODARE

doi



rodare.hzdr.de/record/2529

RODARE | Search | Upload | Communities | Log in

October 29, 2023 | Figure | Open Access

Supporting microscope images: The Impact of Al2O3 Particles from Grit-Blasted Ti6Al7Nb (Alloy) Implant Surfaces on Biocompatibility, Aseptic Loosening, and Infection

Kocjančič, Boštjan; Avsec, Klemen; Šetina Batič, Barbara; Feizpour, Darja; Godec, Matjaž; Kralj Iglič, Veronika; Podlipec, Rok; Cor, Andrej; Debeljak, Mojca; Grant T., John; Jenko, Monika; Dolinar, Drago

Supporting confocal fluorescence microscope images of live cells grown on different implant surfaces

Preview

Files (60.6 MB)

Name	Size	
Cell growth on different surfaces_polished, nonpolished, rough.png	4.7 MB	Preview · Download

15 views · 1 downloads · See more details...

Publication date: October 29, 2023

DOI: 10.14278/rodare.2529

Keyword(s): Ti6Al7Nb implant alloy cementless hip endoprostheses · roughness · Al2O3 grit blasting · surface and subsurface implant contamination · cytotoxicity · aseptic loosening · infection · osteointegration

Related identifiers:
Identical to: https://www.hzdr.de/publications/Publ-37728
Referenced by: 10.3390/ma16216867, https://www.hzdr.de/publications/Publ-37725

Communities: RODARE

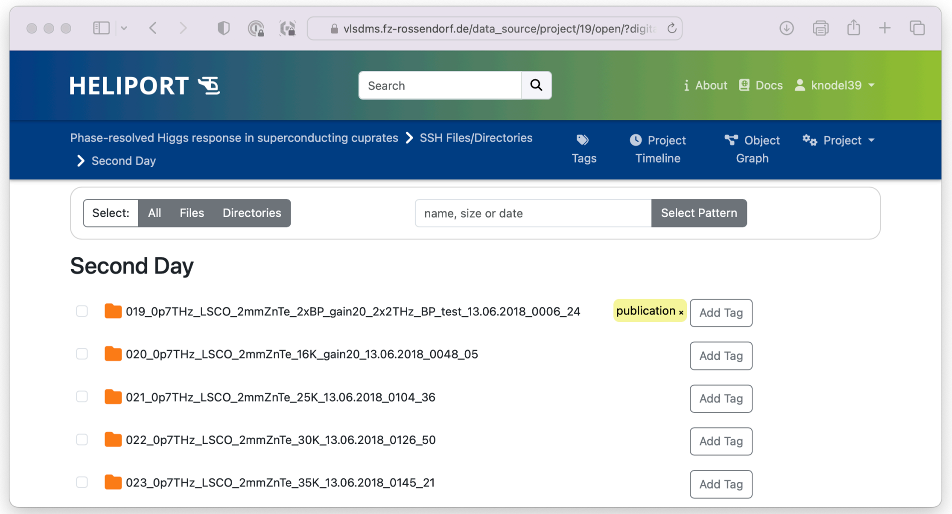
License (for files): Creative Commons Attribution 4.0 International

Versions
Version 1 — Oct 29, 2023 — 10.14278/rodare.2529

Cite all versions? You can cite all versions by using the DOI 10.14278/rodare.2528. This DOI represents all versions, and will always resolve

HELIPORT

hdl enabled



vlsdms.fz-rossendorf.de/data_source/project/19/open/?digit

HELIPORT | Search | About | Docs | knodel39

Phase-resolved Higgs response in superconducting cuprates > SSH Files/Directories > Second Day | Tags | Project Timeline | Object Graph | Project

Select: All | Files | Directories | name, size or date | Select Pattern

Second Day

- 019_0p7THz_LSCO_2mmZnTe_2xBP_gain20_2x2THz_BP_test_13.06.2018_0006_24 — publication × — Add Tag
- 020_0p7THz_LSCO_2mmZnTe_16K_gain20_13.06.2018_0048_05 — Add Tag
- 021_0p7THz_LSCO_2mmZnTe_25K_13.06.2018_0104_36 — Add Tag
- 022_0p7THz_LSCO_2mmZnTe_30K_13.06.2018_0126_50 — Add Tag
- 023_0p7THz_LSCO_2mmZnTe_35K_13.06.2018_0145_21 — Add Tag

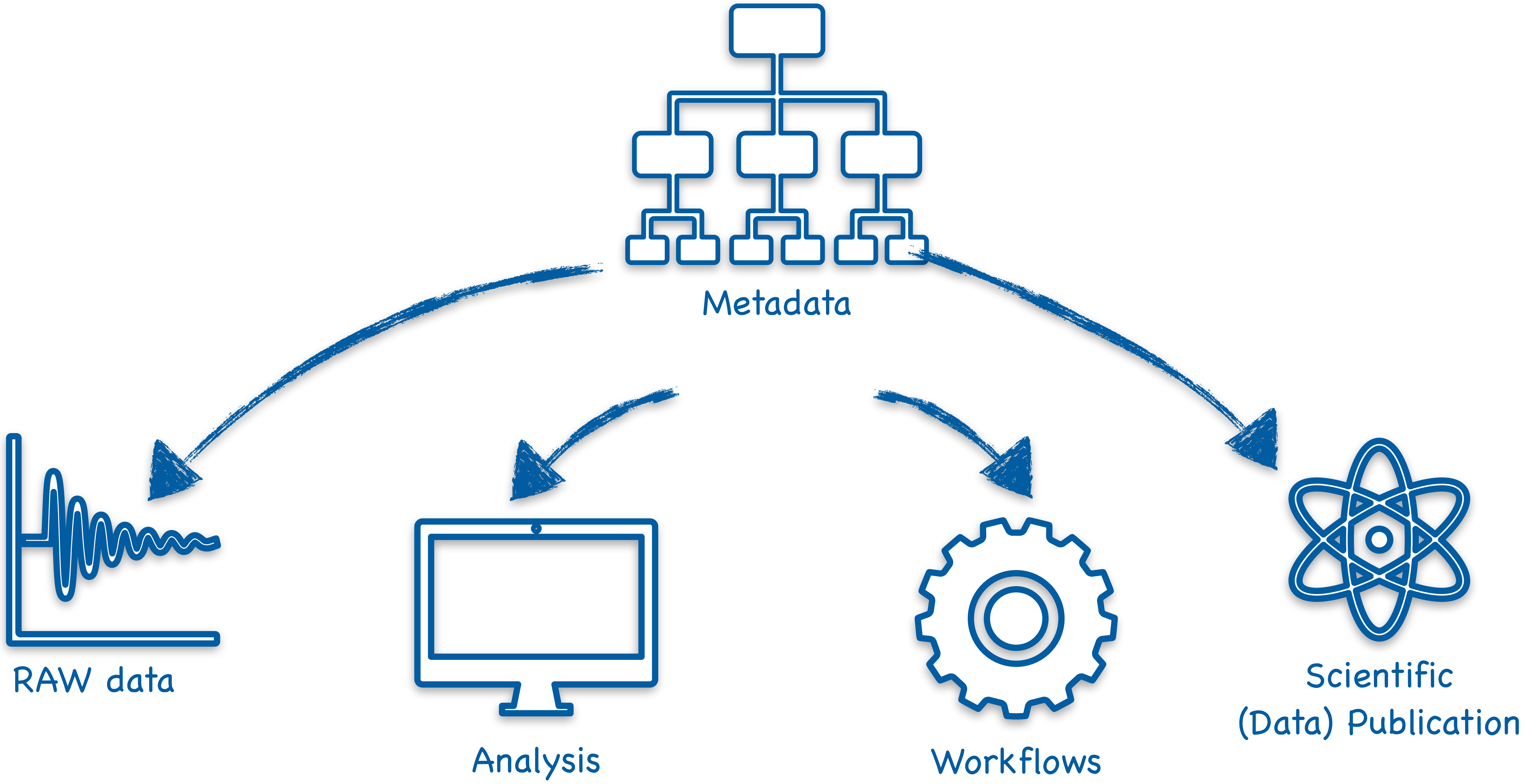
Dataset

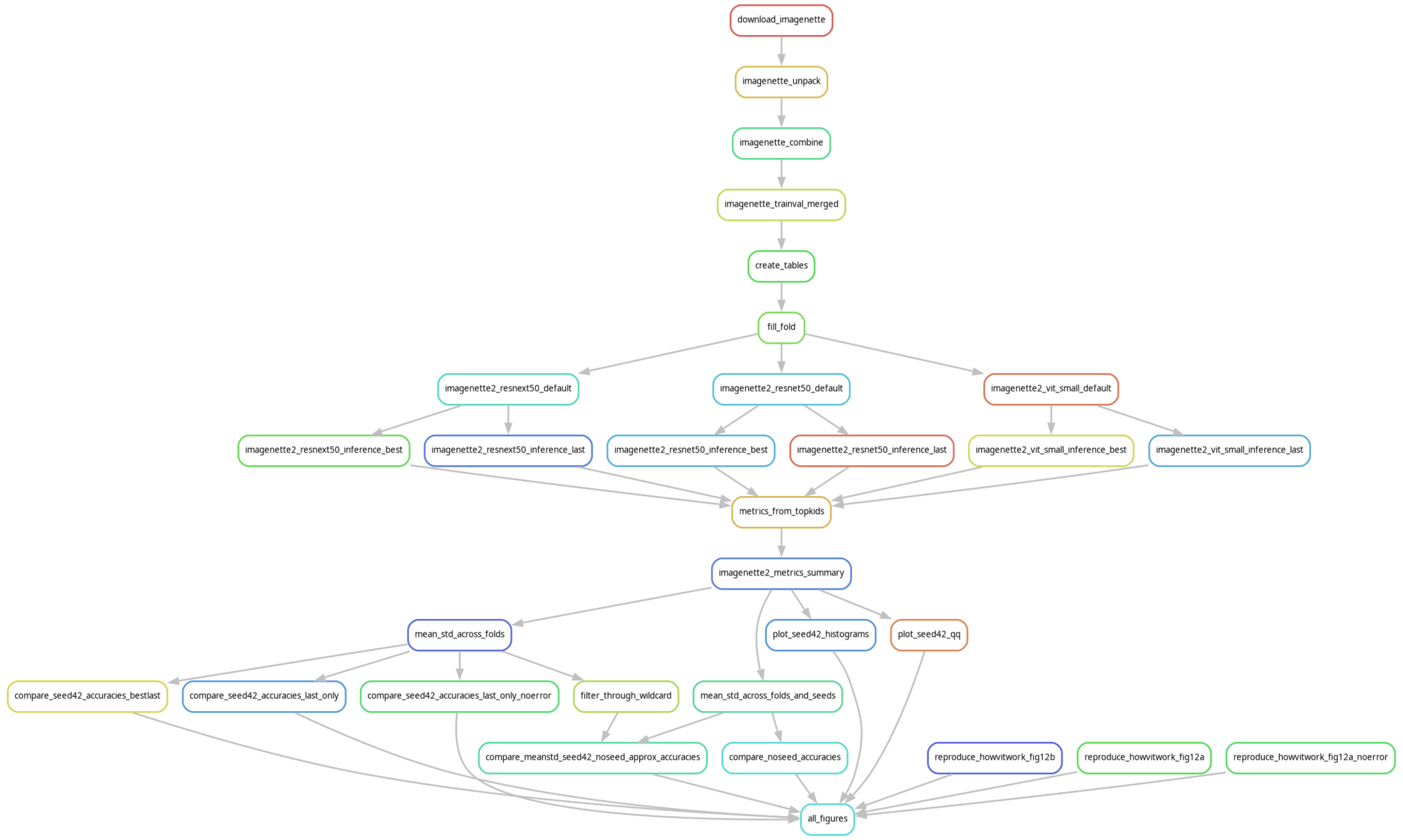
Filesystem

Tape Archive

Conclusions

- Access to data via metadata and PIDs is required to enable ML according to the FAIR criteria
- Connecting Systems, services and processes using APIs for Metadata exchange is essential.
- Metadata and data should follow data standards and schemas to allow exchange of research products and to provide **FAIR** and **comprehensible** research.
- The computational workflows are essential to automate recurring processes.




download_imagenette
imagenette_unpack
imagenette_combine
imagenette_trainval_merged
create_tables
fill_fold
imagenette2_resnext50_default
imagenette2_resnet50_default
imagenette2_vit_small_default
imagenette2_resnext50_inference_best
imagenette2_resnext50_inference_last
imagenette2_resnet50_inference_best
imagenette2_resnet50_inference_last
imagenette2_vit_small_inference_best
imagenette2_vit_small_inference_last
metrics_from_topkids
imagenette2_metrics_summary
mean_std_across_folds
plot_seed42_histograms
plot_seed42_qq
compare_seed42_accuracies_bestlast
compare_seed42_accuracies_last_only
compare_seed42_accuracies_last_only_noerror
filter_through_wildcard
mean_std_across_folds_and_seeds
compare_meanstd_seed42_noseed_approx_accuracies
compare_noseed_accuracies
reproduce_howvitwork_fig12b
reproduce_howvitwork_fig12a
reproduce_howvitwork_fig12a_noerror
all_figures