



Documenting ML Experiments in HELIPORT

David Pape, Oliver Knodel, Sebastian Starke

2024-03-07 · Helmholtz-Zentrum Dresden - Rossendorf · Information Services and Computing · David Pape · d.pape@hzdr.de · <https://heliport.hzdr.de>

Contents

And Some Disclaimers

- Data Management Guidance System HELIPORT
- Examples of „ML Experiments“
- Documentation of ML Experiments
- How HELIPORT Could Help

I am **not** a Machine Learning expert!

These are **not** turn-key solutions but initial ideas!

HELIPORT HELMholtz Scientific Project WORKflow PlatForm



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

Project Members:



Funded by:



HELMHOLTZ
Metadata
Collaboration

```
{
  "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": "Knode1, Dr. Oliver (FWCC) - 132739"
  },
  "heliport:has_VersionControl": [
    {
      "heliport:version_control_id": 15,
      "datacite:uri": "https://dd",
      "rdfs:label": "Test"
    }
  ],
  "heliport:has_DataManagementPlan": [
    {
      "heliport:data_management_plan_id": 6,
      "datacite:uri": "https://dddd",
      "datacite:hasDescription": "dddd"
    }
  ],
  "heliport:has_Documentation": [
    {
      "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": ""
    }
  ]
}
```

Basic Linking of Resources

Documentation, Publications, Software, Data Archives, ...

- Collects all resources of a project
- Helps current and future you remember
- Helps onboard new colleagues

🏠 > gELBE beamtime 21102205-ST > Publication 🏷️ Tags 🕒 Proj

Publications

ID	DOI/Handle/...	Description
10	https://doi.org/10.14278/rodare.1188	Tests of the detector system for the Stopping Target Monitor of the MU2E experiment in a high flux pulsed gamma beam

🏠 > gELBE beamtime 21102205-ST > Documentation 🏷️ Tags 🕒 Proj

Documentation

ID	Description	System
11	Experimental Setup (Room 540/109)	MediaWiki
16	Cloud storage containing Pictures, Software, Presentations related to the beamtime	Lims
12	HedgeDoc - Mu2e @ELBE Labbook	HedgeDoc

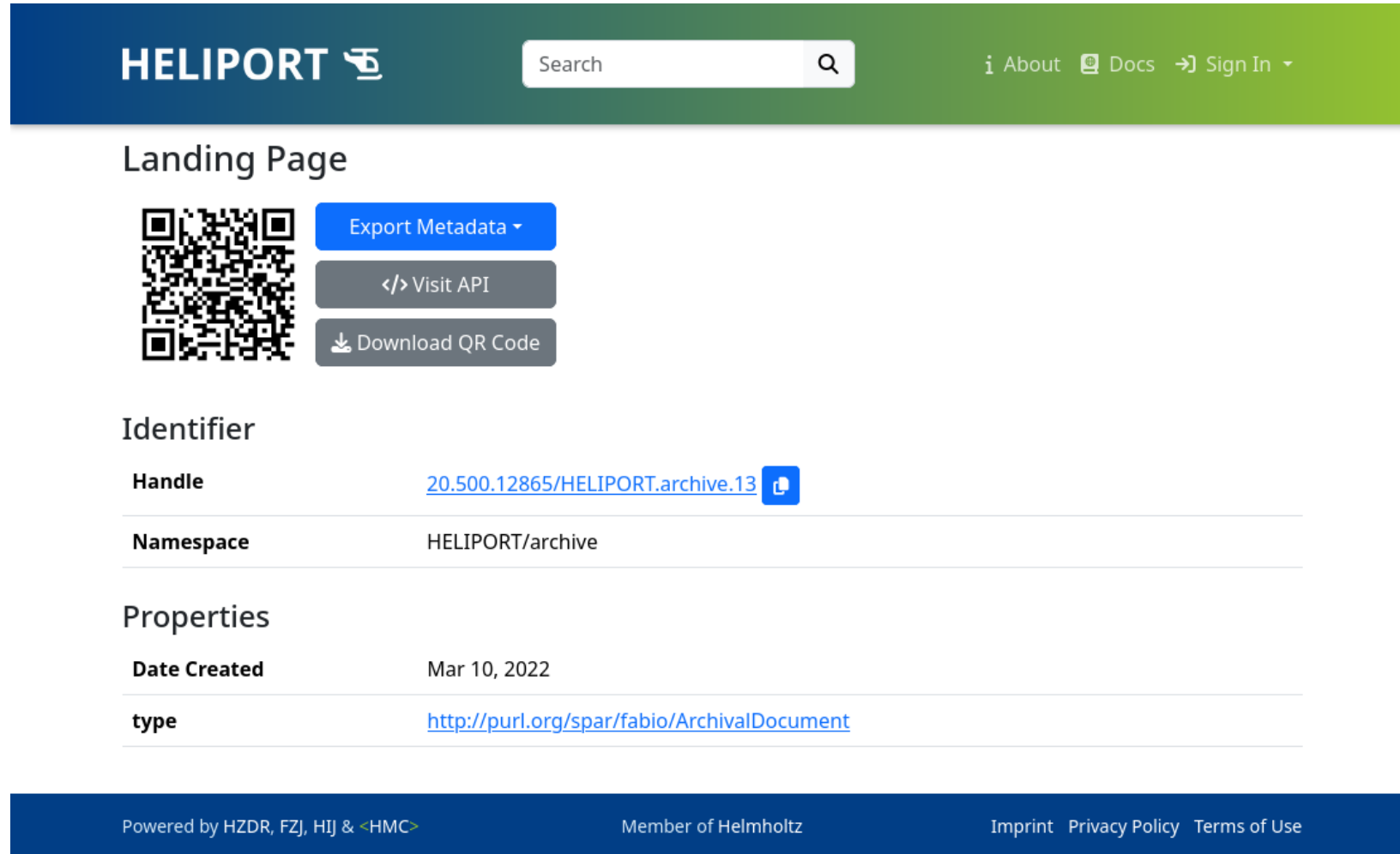
Automatic data publication to RODARE (01 Jun 2022)
Mu2e Run I Sensitivity Projections for the Neutrinoless $\mu^- \rightarrow e^- \gamma$ Conversion Search in Aluminum



Landing Pages

For All Digital Objects

- Persistent Identifiers (handle.net)
- Semantic properties
- Visibility levels for metadata
- Metadata export
 - DataCite
 - RDF (various serializations)



The screenshot shows the HELIPORT interface. At the top, there is a dark blue header with the HELIPORT logo and a search bar. Below the header, the page title "Landing Page" is displayed. A QR code is shown on the left, with three buttons to its right: "Export Metadata", "Visit API", and "Download QR Code". Below the QR code, the "Identifier" section shows the Handle "20.500.12865/HELIPORT.archive.13" with a copy icon. The "Namespace" is "HELIPORT/archive". The "Properties" section shows the "Date Created" as "Mar 10, 2022" and the "type" as "http://purl.org/spar/fabio/ArchivalDocument". At the bottom, there is a dark blue footer with the text "Powered by HZDR, FZJ, HIJ & <HMC>", "Member of Helmholtz", and "Imprint Privacy Policy Terms of Use".

HELIPORT

Search

About Docs Sign In

Landing Page

Export Metadata

Visit API

Download QR Code

Identifier

Handle [20.500.12865/HELIPORT.archive.13](https://hdl.handle.net/20.500.12865/HELIPORT.archive.13)

Namespace HELIPORT/archive

Properties

Date Created Mar 10, 2022

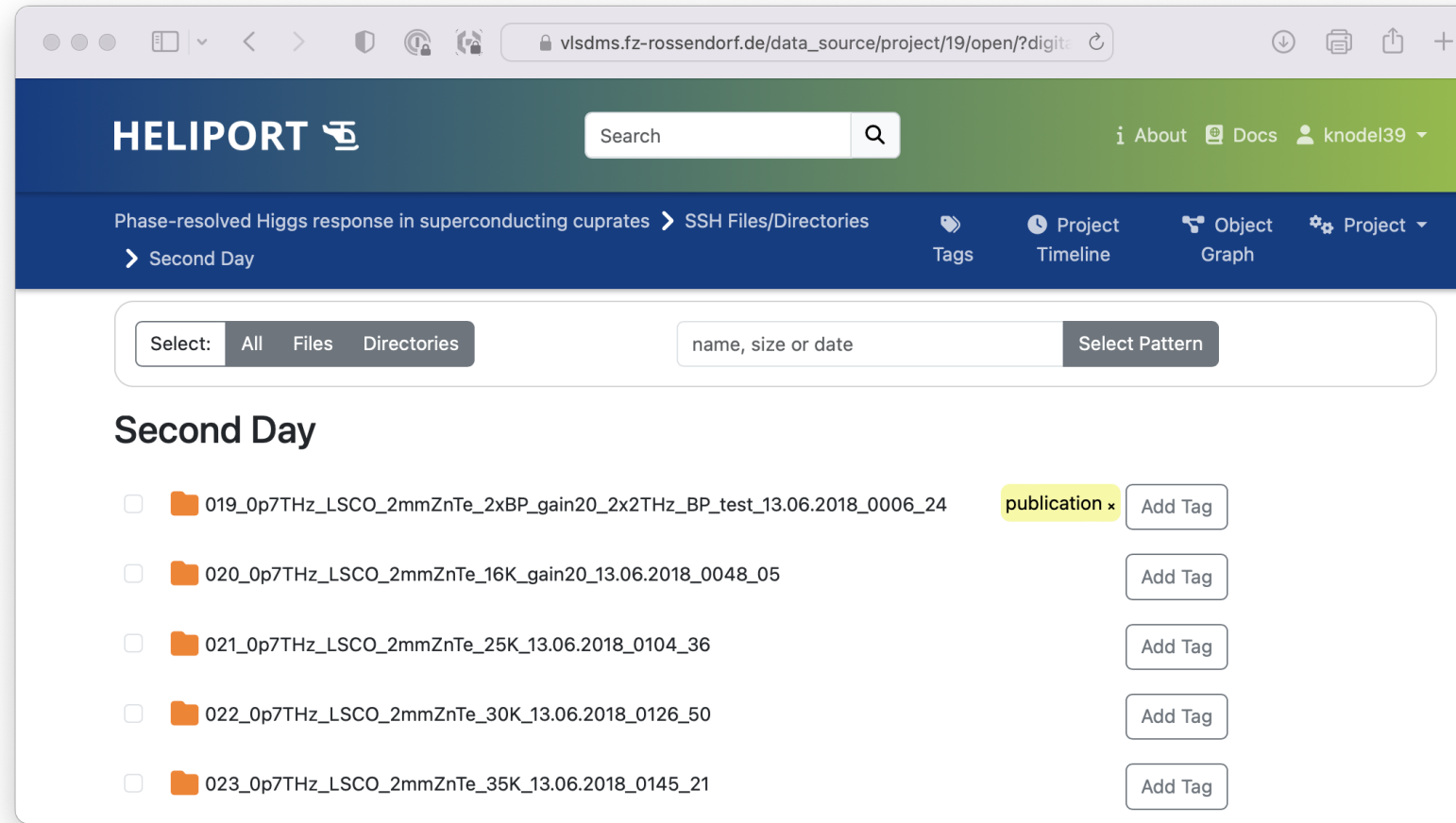
type <http://purl.org/spar/fabio/ArchivalDocument>

Powered by HZDR, FZJ, HIJ & <HMC> Member of Helmholtz Imprint Privacy Policy Terms of Use

Data Sources

File Browser and Tags

- SFTP, SMB/CIFS supported
- Can be browsed
- Files for download
- „publication“ tag allows for automated publication on Rodare



The screenshot displays the HELIPORT web interface. The browser address bar shows the URL: `vlsdms.fz-rossendorf.de/data_source/project/19/open/?digit`. The page header includes the HELIPORT logo, a search bar, and user information for 'knodel39'. The breadcrumb trail indicates the current location: 'Phase-resolved Higgs response in superconducting cuprates > SSH Files/Directories > Second Day'. The main content area shows a file browser for the 'Second Day' directory. A search filter is set to 'name, size or date'. A list of five files is displayed, each with a checkbox, a folder icon, a filename, and an 'Add Tag' button. The first file, '019_0p7THz_LSCO_2mmZnTe_2xBP_gain20_2x2THz_BP_test_13.06.2018_0006_24', has a yellow 'publication' tag applied to it.

File Name	Tag
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	

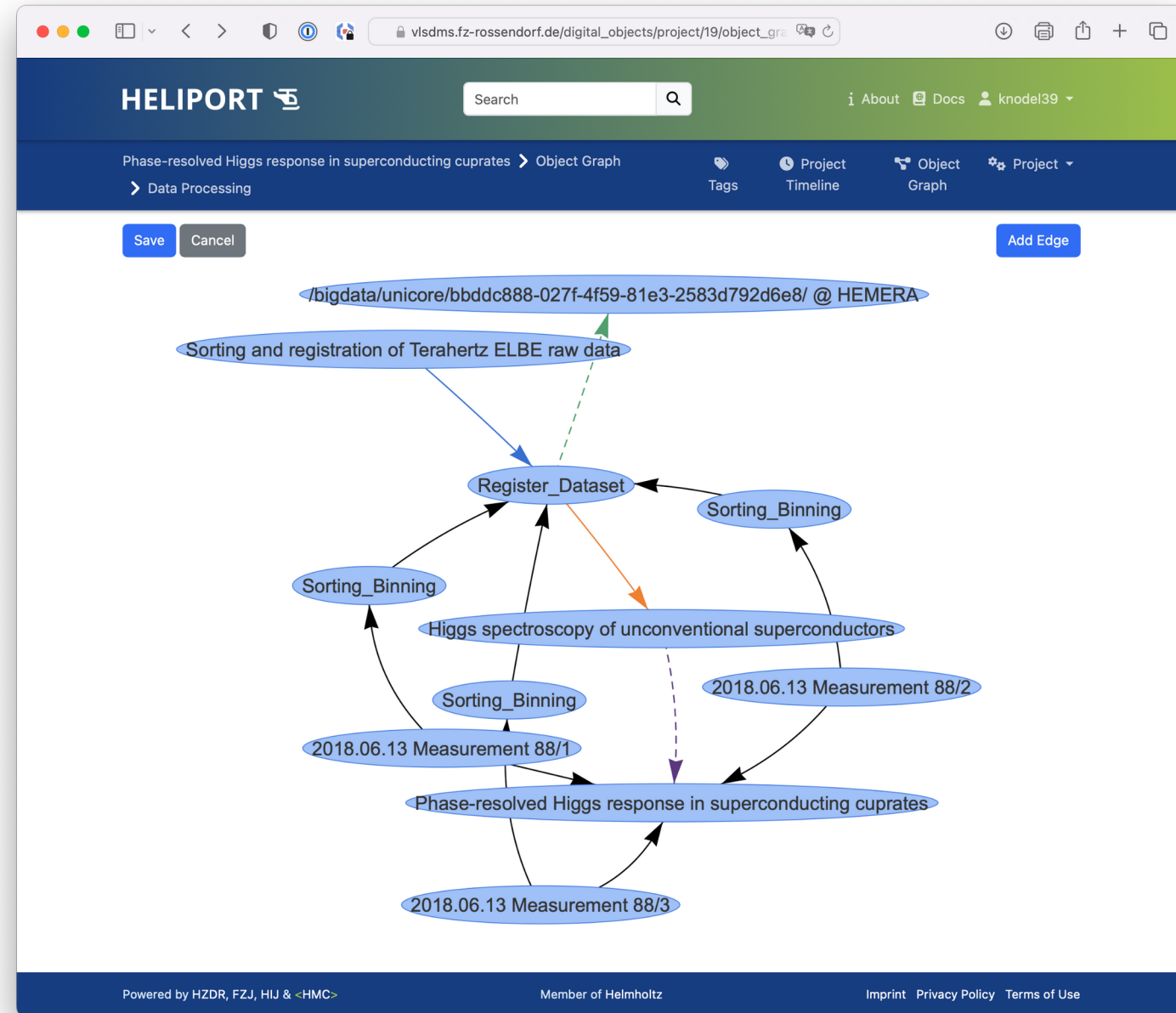
Digital Object Graphs

Relations Between Resources

- Visualization of facets of an experiment
- Currently created manually
 - Error-prone
 - Use and possibilities not clear to the user
- In the future: Automatic creation from ontology-based metadata

First use case:

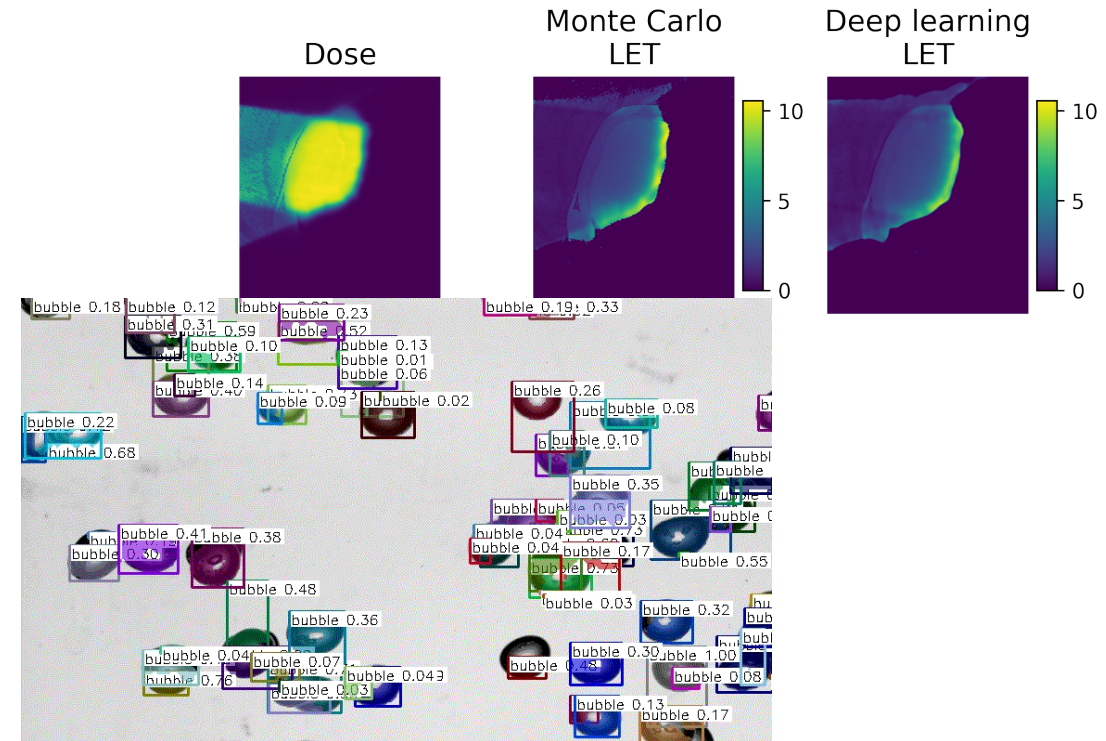
Computational Workflows in Snakemake



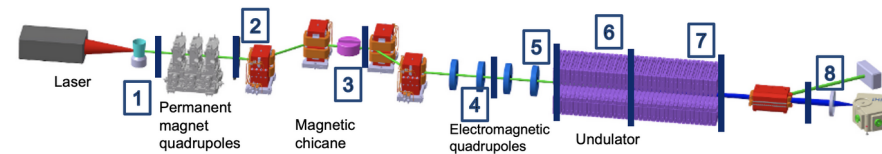
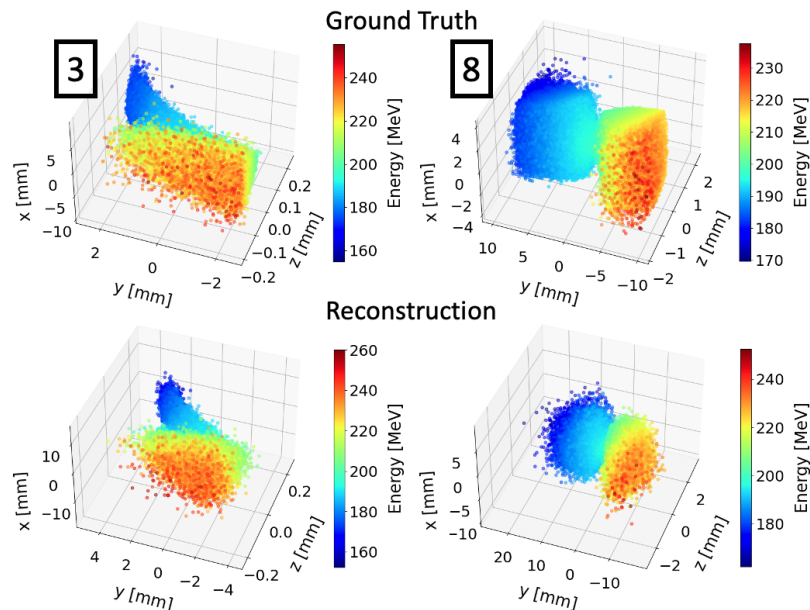
ML Experiments at HZDR

From the Local Helmholtz AI Unit

- Approximation of linear energy transfer (LET) in radiotherapy using deep learning
- Instance segmentation of bubbles (in videos)
- Surrogate models and virtual diagnostics for stages of laser-driven particle accelerators (→ digital twins)



Figures: Sebastian Starke



Figures:
Marie-Emmanuelle Couprie et al., 10.1088/1742-6596/1596/1/012040;
Anna Willmann, Helmholtz AI Conference 2023

ML Experiments Embedded in Larger Contexts

Considerations for Documentation and Relevant Metadata in HELIPORT

- Datasets are from real-world experiments (not MNIST/ImageNet) and might change!
 - Labels may be added
 - Correction of erroneous data
- (User) code often not well versioned and packaged
 - Passed around as Jupyter notebooks
 - Custom tweaks for each scientists use case
- Model outputs may influence experimental setups (e.g. laser parameter tweaks)
- Make connections between ML model and large-scale facility (digital twins)

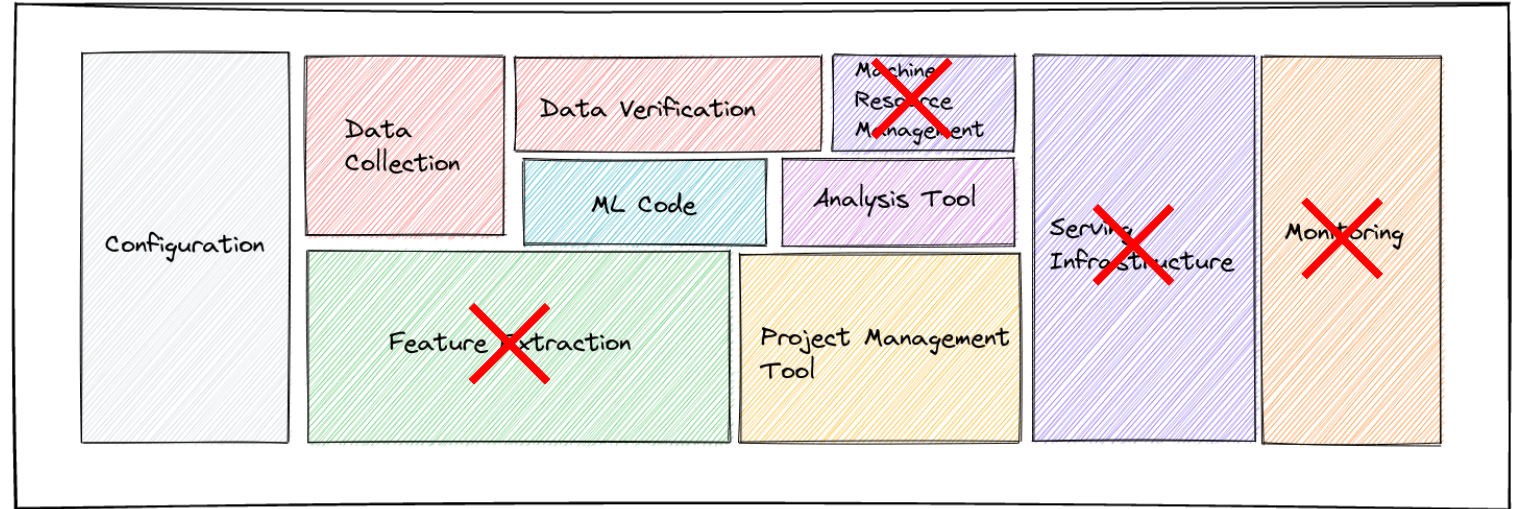
Interesting ML metadata: Model configuration and performance (e.g. accuracy, precision); compute resources used (training time, energy metrics)

Relevant digital objects: Training datasets, trained models, software used

MLOps for ML Experiments

„Can We Support Scientists By Integrating MLOps in HELIPORT?“

- Many aspects of MLOps not applicable to these ML experiments
 - Often only prototypical
 - Iterative development
 - Not service providers
 - No infrastructure to administer
- Goals rooted in open science:
 - Reproducibility
 - Transparency



Yashaswi Nayak: A Gentle Introduction to MLOps (Towards Data Science)

MLOps might come into play at a later stage when:

- Domain scientists apply model „in production“
- A common interface for model interaction is found

A Look at ML (Metadata) Visualization and Analysis Tools

From a Data Management Point of View

Tensorboard

- Visualization tool for model operation graph and metrics, and data
- [TFX metadata library](#) has artifact types
 - Arbitrary, user-defined metadata properties (key-value)

Weights & Biases

- [Artifacts](#) are file-/directory-like objects for dataset and model versioning
- Fixed set of metadata defined by the artifact model

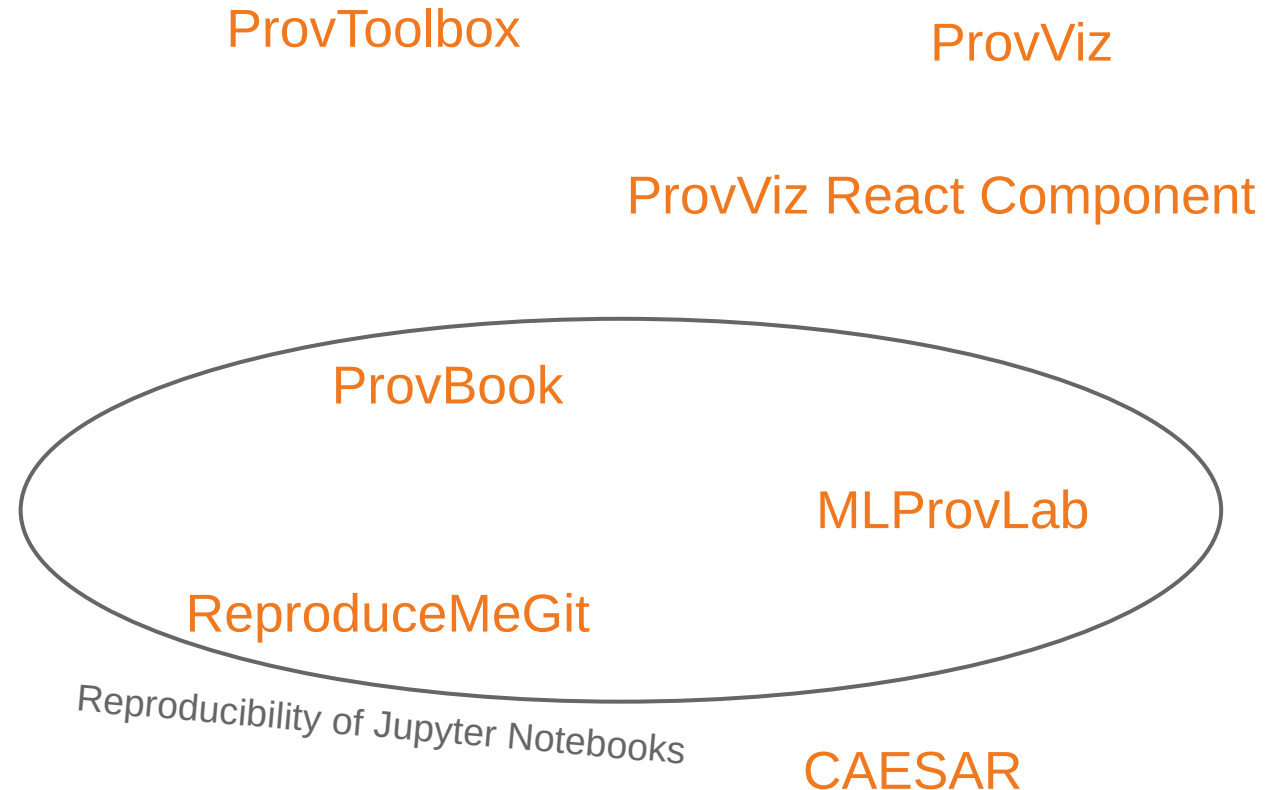
Only the machine learning domain is covered,
inputs and outputs exist without context!

A Look at (some) ML Vocabularies and Ontologies

And Associated Tools

- General description, usage, reasoning:
 - Exposé
 - SML
 - OntoDM-core
 - DMOP
 - ML-Schema
- Data provenance and reproducibility:
 - MEX Vocabulary*
 - REPRODUCE-ME*
 - ReproduceMe-ML*

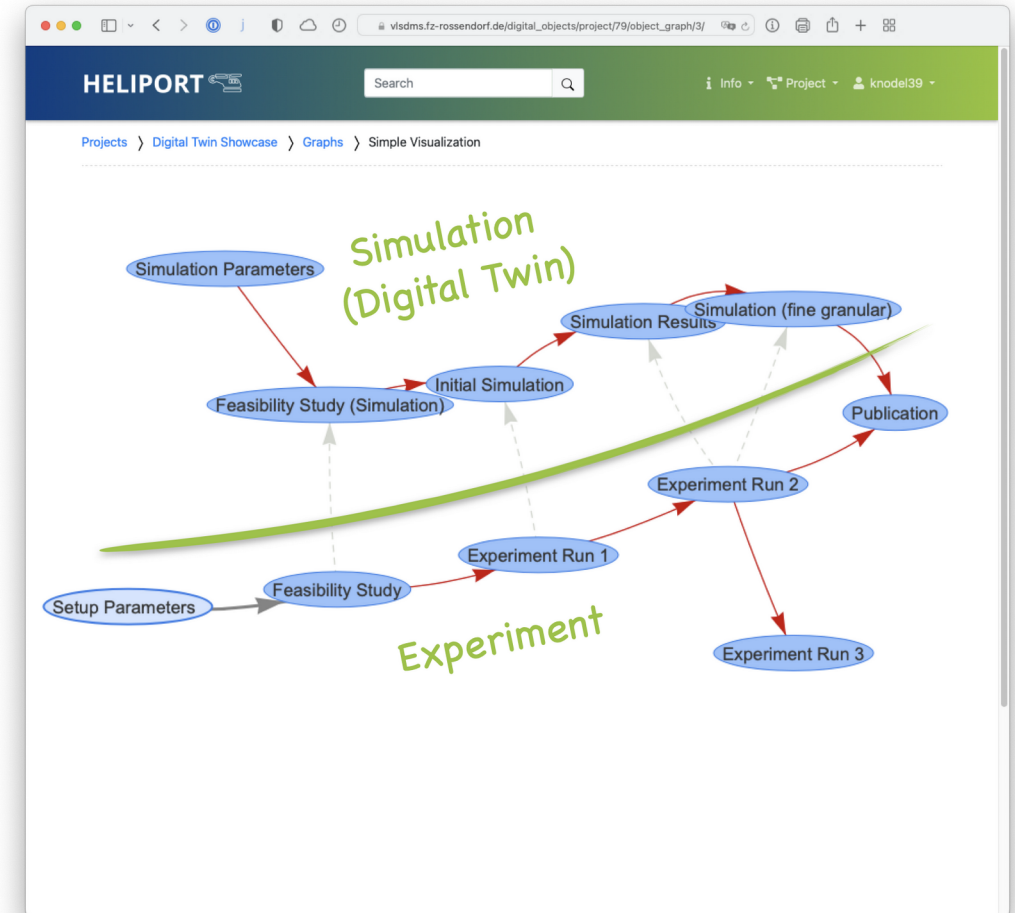
*) based on PROV-O



How HELIPORT Could Help ML Experiments

Description of Experiment within different domains

- Landing pages and globally unique persistent identifiers for all digital objects independent of domain
 - ML, data provenance, computational workflows, bibliographic metadata, ...
- Data and model provenance throughout entire project
- Identify upstream changes that affect the ML process
- Identify downstream benefits
- Automatic visualization of relationships (→ similarly to Snakemake workflows)
- Automatic creation of dataset and model cards, and visualizations for publication



Thank you for your attention!



<https://heliport.hzdr.de>