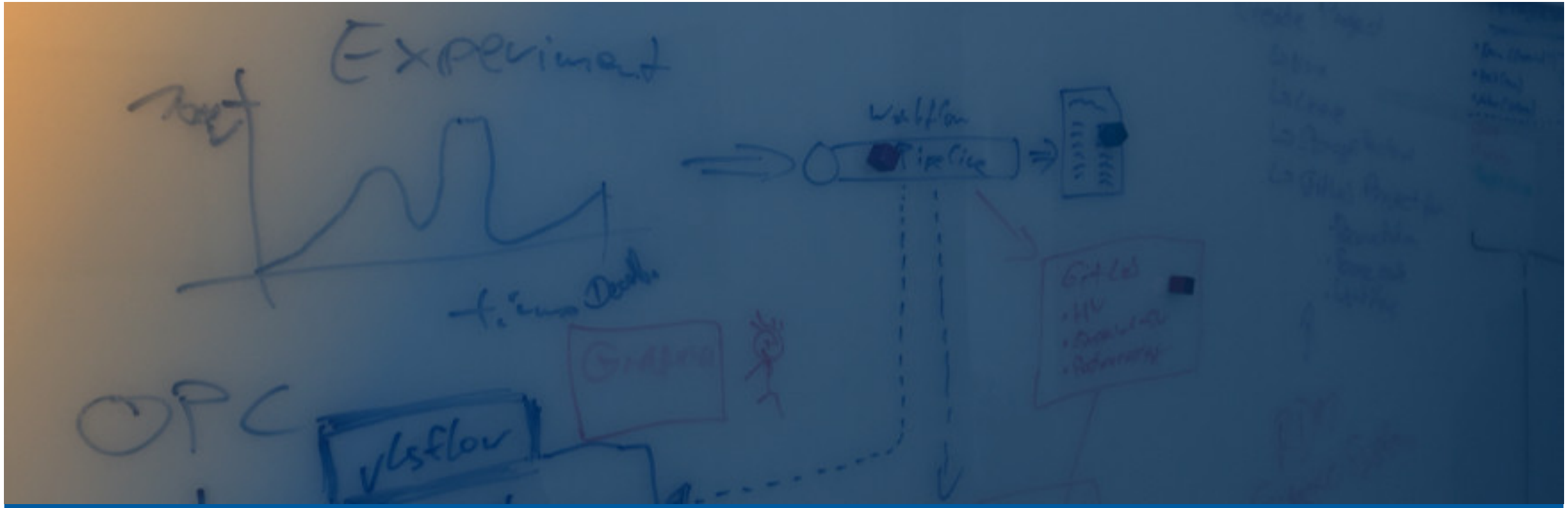
HELIPORT HELmholtz ScIentific Project W ORkflow PlaTform

HZDR HELMHOLTZ ZENTRUM DRESDEN ROSSENDORF



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

DRESDEN concept

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:

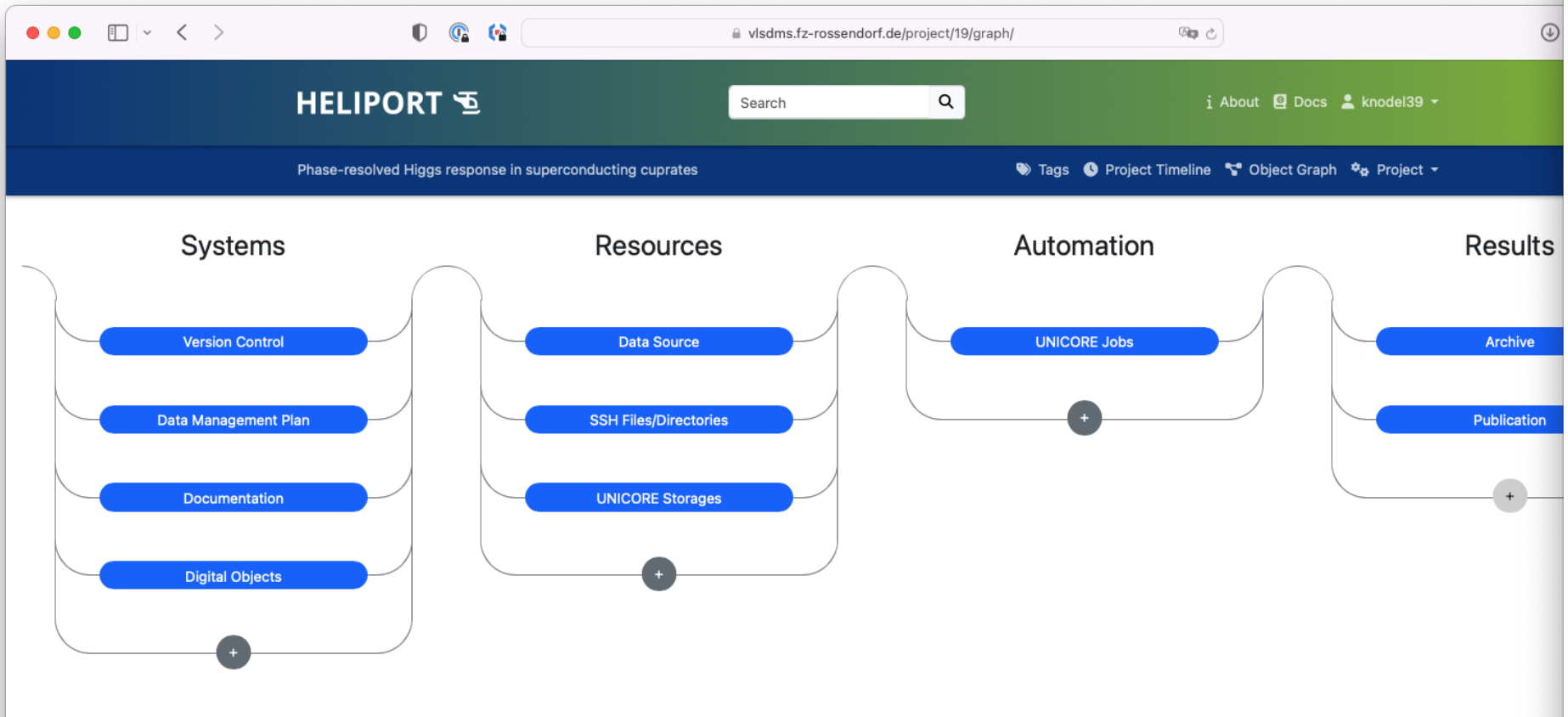
HZDR HELMHOLTZ ZENTRUM DRESDEN ROSSENDORF

HI JENA Helmholtz Institute Jena

JÜLICH FORSCHUNGSZENTRUM

Funded by:

HMC HELMHOLTZ Metadata Collaboration



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 Proje

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

gELBE beamtime 21102205-ST > Documentation Tags Pr

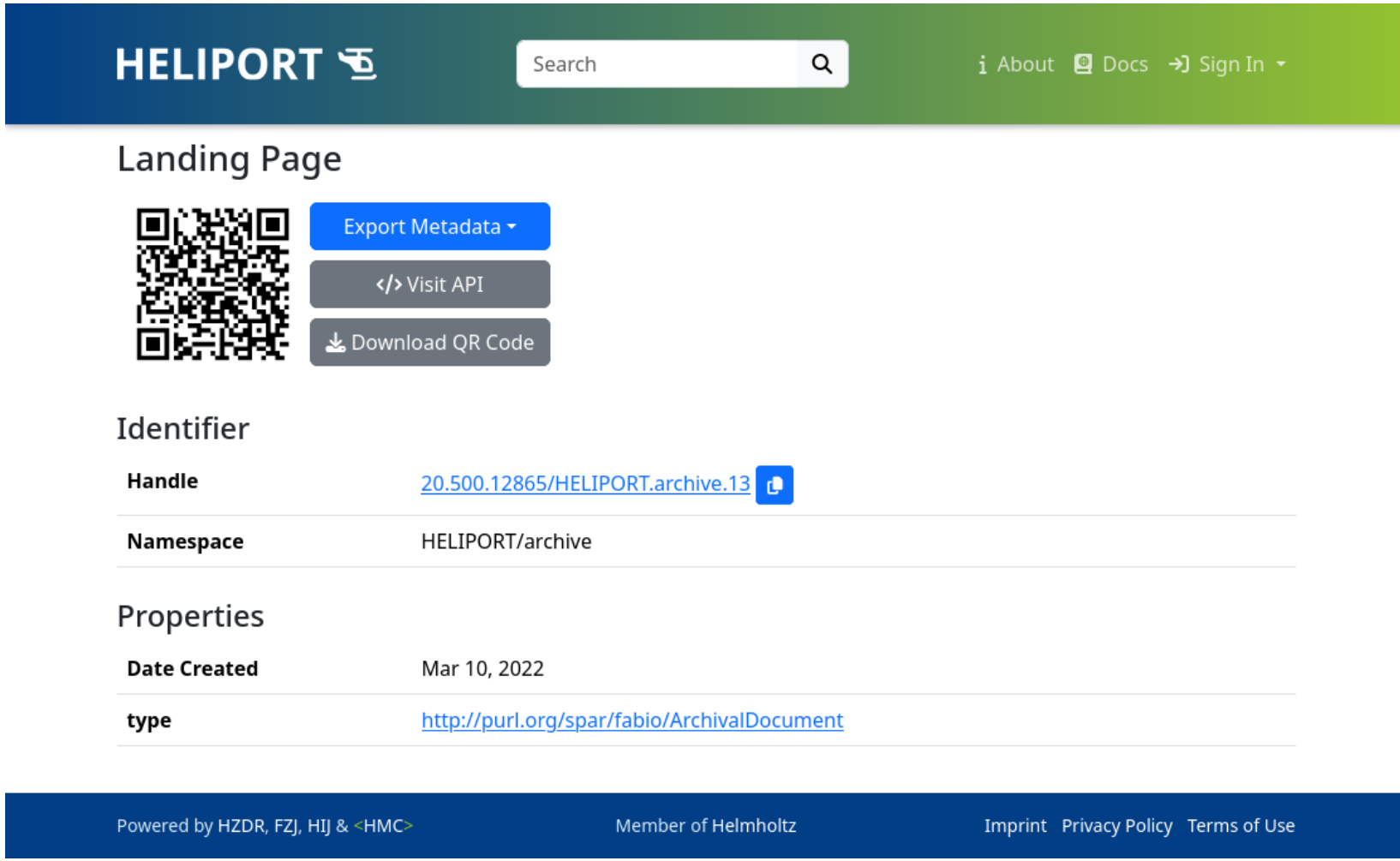
Documentation

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

Landing Pages

For All Digital Objects

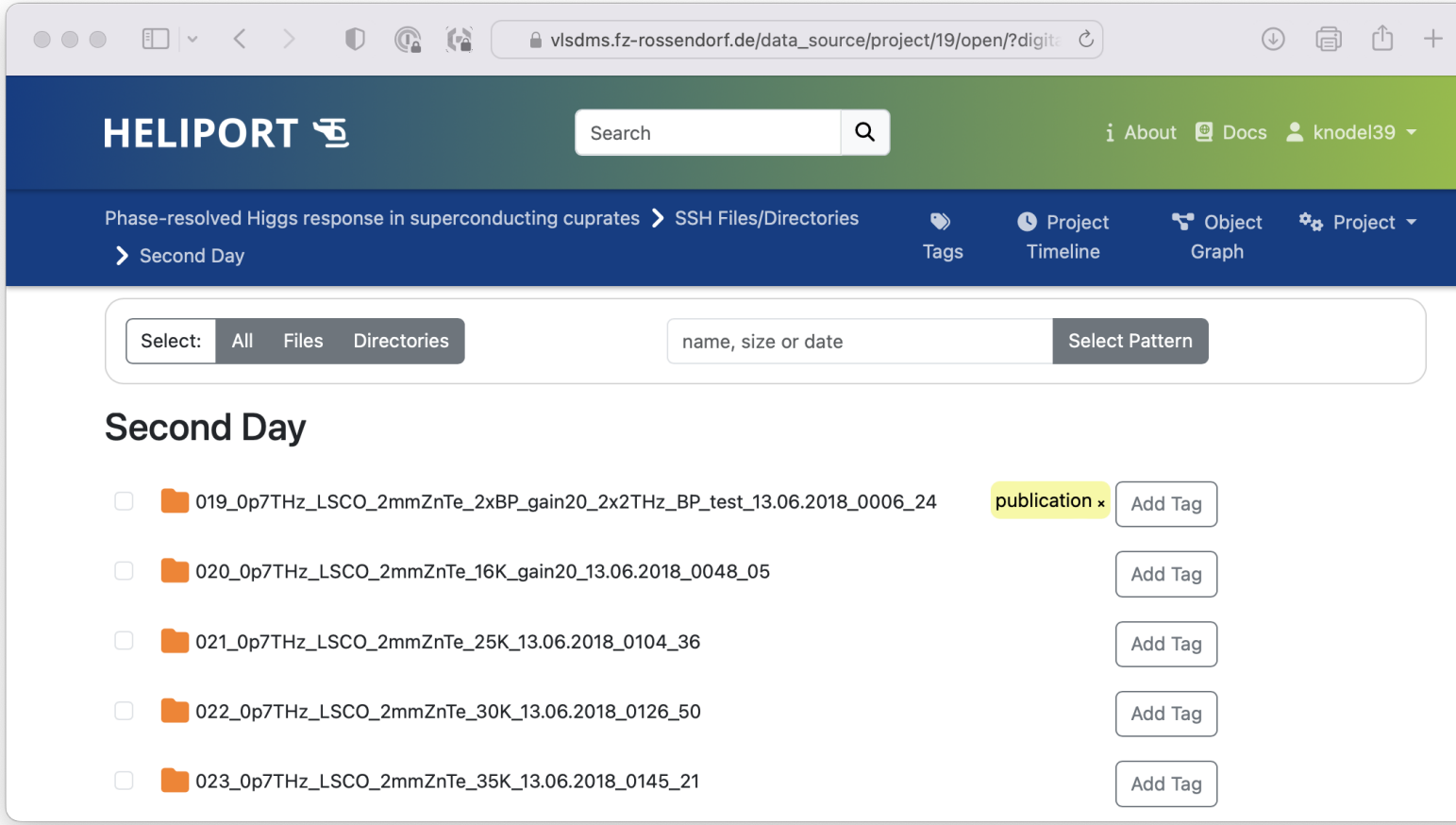
- Persistent Identifiers (handle.net)
- Semantic properties
- Visibility levels for metadata
- Metadata export
  - DataCite
  - RDF (various serializations)



Data Sources

File Browser and Tags

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



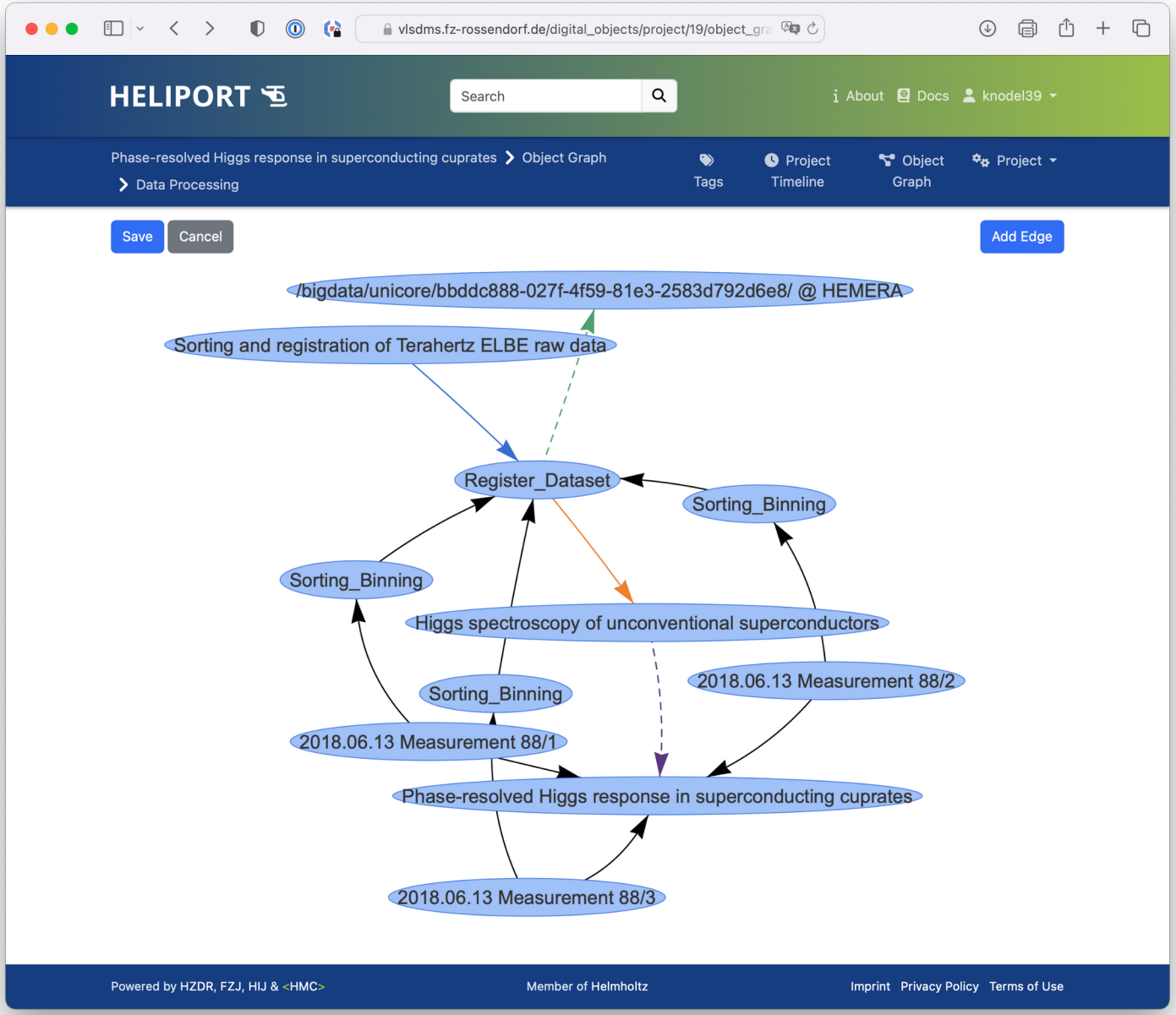
Digital Object Graphs

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:

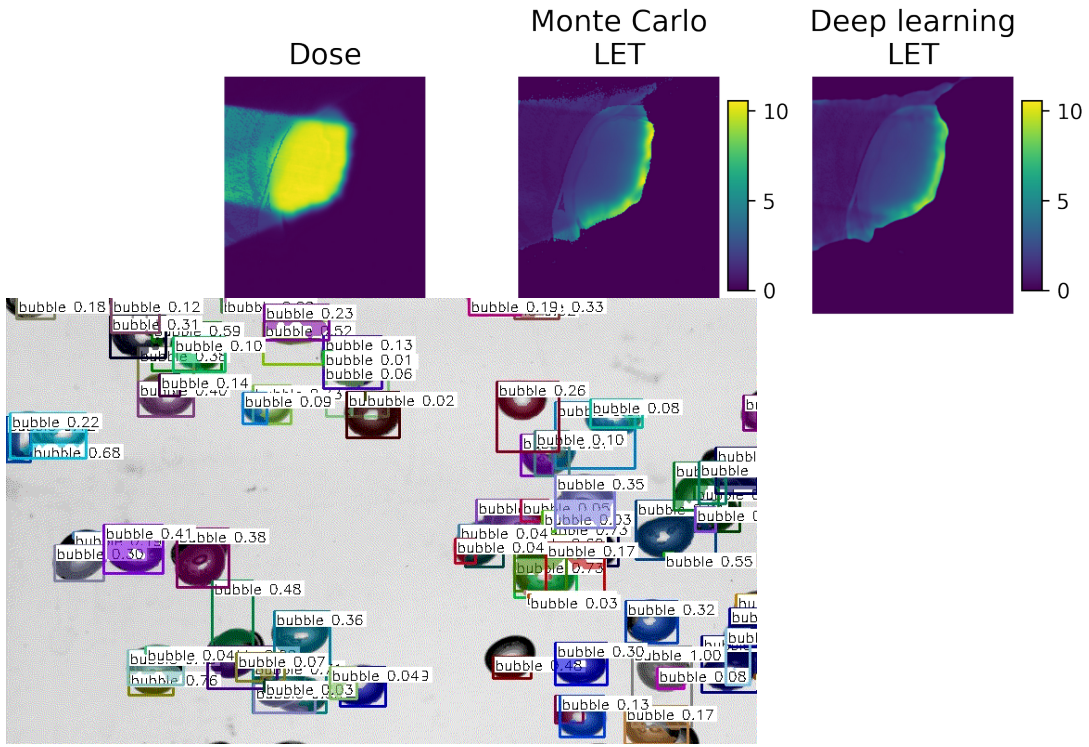
Computational Workflows in Snakemake



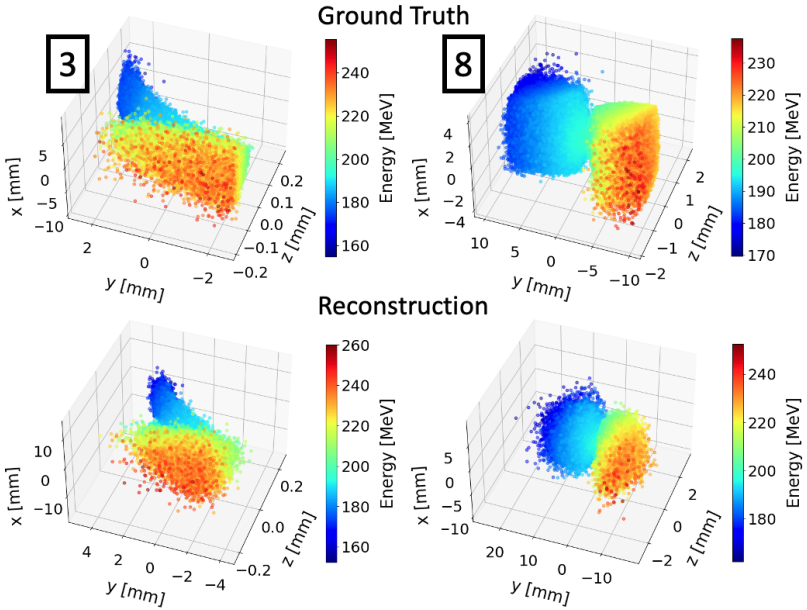
ML Experiments at HZDR

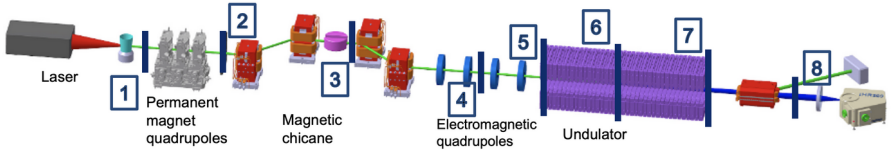
From the Local Helmholtz AI Unit

- 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)

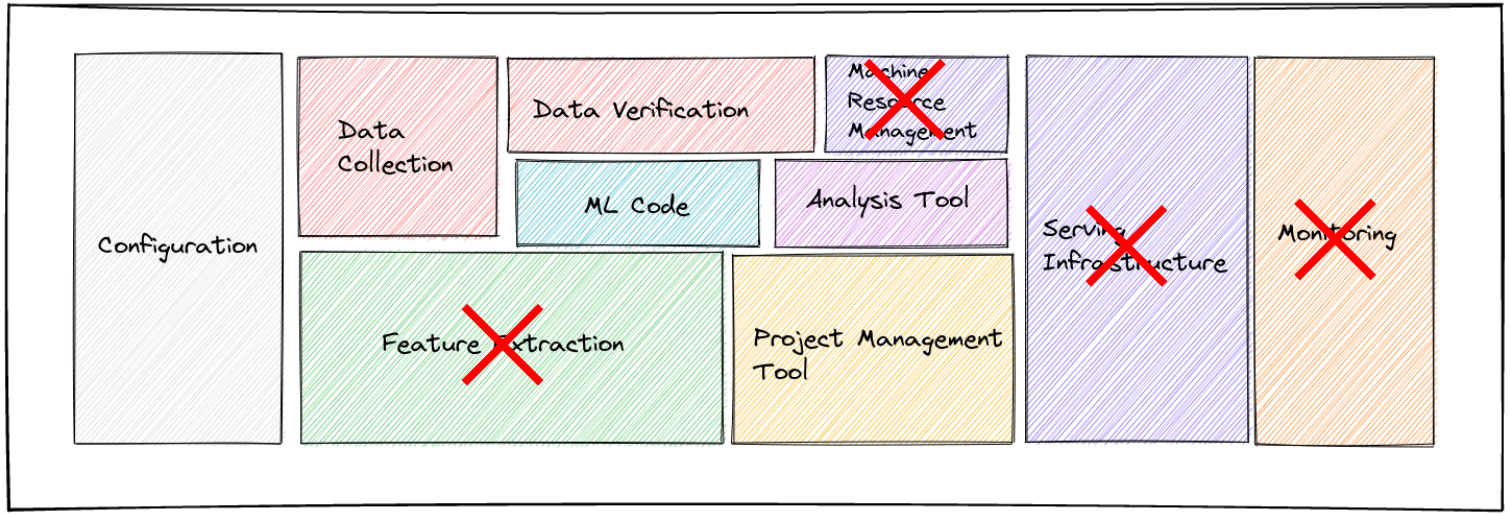
Interesting ML metadata: Model configuration and performance (e.g. accuracy, precision); compute resources used (training time, energy metrics)

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

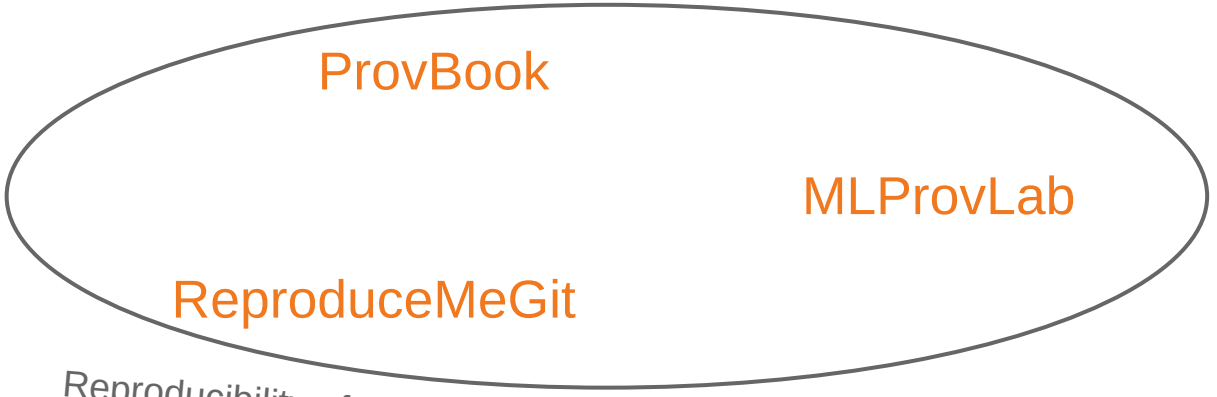
- General description, usage, reasoning:
  - Exposé
  - SML
  - OntoDM-core
  - DMOP
  - ML-Schema
- Data provenance and reproducibility:
  - MEX Vocabulary*
  - REPRODUCE-ME*
  - ReproduceMe-ML*

*) based on PROV-O

ProvToolbox

ProvViz

ProvViz React Component

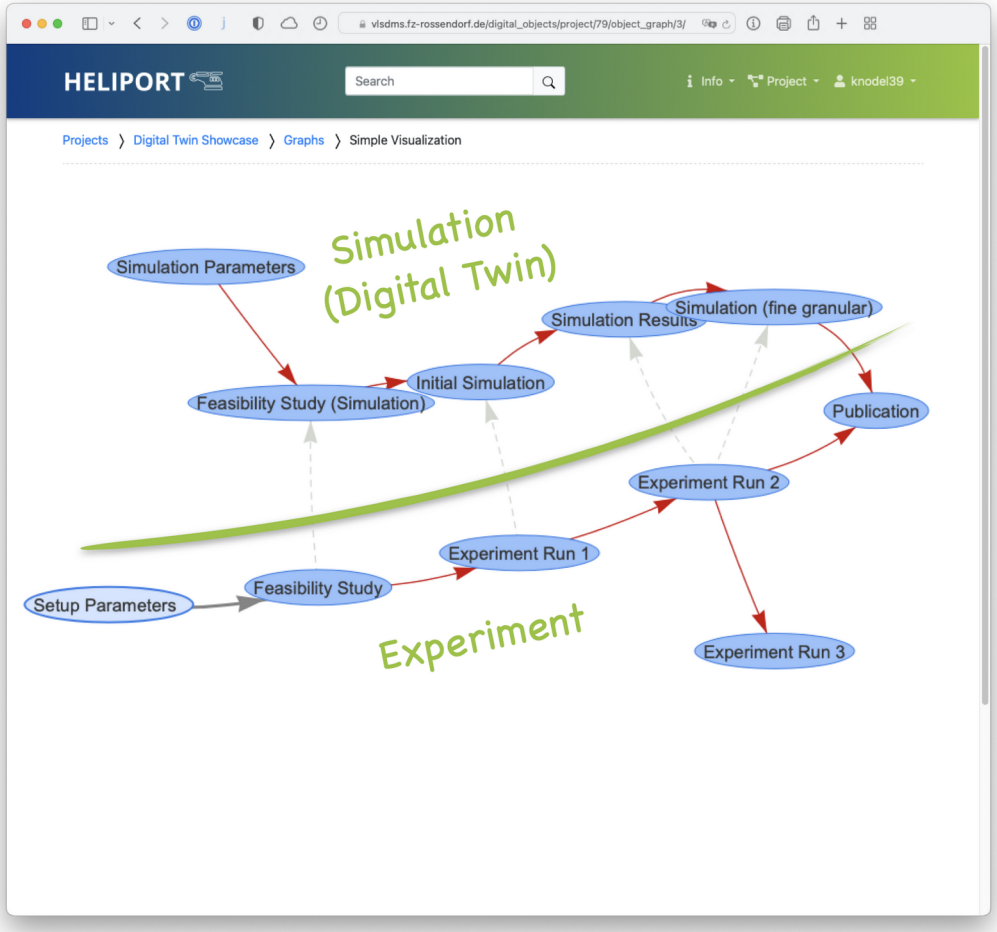


CAESAR

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



HELIPORT HELmholtz Scientific Project WORkflow PlaTform

Thank you for your attention!

https://heliport.hzdr.de