deRSE CONFERENCE WÜRZBURG 2024

Contribution ID: **90** Type: **Talk (15min + 5min)**

# Documenting ML Experiments in HELIPORT

*Thursday, March 7, 2024 9:40 AM (20 minutes)*

HELIPORT is a data management guidance system that aims at making the components and steps of the entire research experiment's life cycle findable, accessible, interoperable and reusable according to the FAIR principles. It integrates documentation, computational workflows, data sets, the final publication of the research results, and many more resources. This is achieved by gathering metadata from established tools and platforms and passing along relevant information to the next step in the experiment's life cycle. HELIPORT's high-level overview of the project allows researchers to keep all aspects of their experiment in mind.

A particularly interesting use case are machine learning projects. They are often prototypical in nature and driven by iterative development, so reproducibility and tranparency are a great concern. It is essential to keep track of the relationship between input data, choices in model parameters, the code version in use, and performance measures and generated outputs at all times. This requires a data management platform that automatically records the changes made and their effects. Existing MLOps tools (such as Weights and Biases, MLFlow) live entirely in the ML domain and start their workflow with the assumption that data is available. HELIPORT, on the other hand, takes care of the data lifecycle as well. Our envisioned platform interoperates with the domain specific tools already used by the scientists, and is able to extract relevant metadata (e.g. provenance). It can also make persistent any additional information such as papers the work was based on, documentation of software components, workflows, or failure cases. Moreover, it should be possible to publish these metadata in machine-readable formats.

The challenge arising from these aspects consists in integrating ML workflows into HELIPORT in such a way that they work on the provided data and metadata. The goal is also to enable the comprehensible development of ML models alongside the experiment documented in HELIPORT. This allows different teams (e.g. experimentalists and AI specialists) to work together on the same project in a seamless manner, and help generate FAIRer outcomes. In the long term we hope to aide in establishing digital twins of facilities, and making their maintenance a part of the data management proces.

## Slot length

**Primary authors:** PAPE, David (Helmholtz-Zentrum Dresden-Rossendorf (HZDR)); KNODEL, Oliver (Helmholtz-Zentrum Dresden-Rossendorf); STARKE, Sebastian (HZDR)

**Presenter:** PAPE, David (Helmholtz-Zentrum Dresden-Rossendorf (HZDR))

**Session Classification:** AI/ML Research Software

**Track Classification:** HPC and Massive Data: AI/ML Research Software