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

Draco Data Flow for Advanced Data Management

Draco Infrastructure

- Draco DAQ (EliOOS)
  - Draco Gateways
    - Proxys
    - DracoDB
- Sub-Net “FWKMess”
- Control Room

HZDR Infrastructure

- SciCat
- Jupyter
- HELIPORT
- Rodare
- HPC Cluster
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

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

PID

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 FAIR.
— With interchangeable, standards-based metadata, workflows can be used in different RIs to complete FAIR.

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
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.
Metadata Catalogue SciCat and Data Repository RODARE (Draft)

Curated Metadata Source
- ExperimentLogging app (ExL)
- E-Logbook

Public Metadata Catalogue
- SciCat

Subsequent Access to Data
- RODARE
- HELIPORT

Fully Automated Process for DRACO
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.