Research Data Management at HZDR
Publication Components

Scientific Results

Public. Policy

Data Policy

Horizon 2020 FAIR Data

Data

GitHub

GitLab

Software

Authors

DOI, Metadata

DataCite

RDA

Publication-Management

ROBIS

Publisher

Scientific Results

Publication

Authors

DOI, Metadata

DataCite

RDA

Publication-Management

ROBIS

Publisher
The Concept
Basic DM Concept

Data sources

- 8 kHz CT
- 1.5 MHz Cameras
- HPC Simulations
- ...

Up to 50 GB/s

- Storage (10+ PB)
- (Semi) automatic Workflows using HPC

Web frontend for data upload and workflows

Data publication system RODARE
Data Ingest

- Goal: publication-ready records already on ingest by collecting necessary metadata from available sources (e.g. GATE system, eLab-Books)
- Stepwise inclusion of facilities (T-ELBE, ROFEX)
How to deal with heterogeneous metadata schemes?

• Invenio uses jsonschemas to describe what each record has to look like
• Allows instant validation
• Plan:
  • Extract metadata schema for each resource type together with scientists
  • Characterize facilities in data-base
  • Write tools to extract metadata from existing gadgets if possible
  • convert data into a self explanatory format together with its metadata (e.g. HDF5, ADIOS or Nexus)
HZDR Data Management

- Built using Invenio (like Zenodo) → Elastic Search for large data repositories
- Start predefined compute workflows base on selected input data
- Ingest workflow output again into repo and link to input data set
Tightly coupled, even isolated

- Typical cloud (micro-)service
  - fixed workflows / machine
  - service is feature complete

- “Web”: HTTP & WebSocket
  - persistent, sharable sessions
  - browser, app, CLI

- Existing infrastructure
  - Got intranet? CMS Spawner
  - Qt expert? Try Wt
Loosely coupled, fully enabled

- Prepared Jupyter notebook
  - same user, same rights
  - transparently load: modules/containers/environments

- Lend out (G)UI design principles
  - data ↔ representation
  - client request ↔ HPC scheduling

- Transparent ("Hackable") by design
  - share full workflow as a file: migrate, re-connect
  - add cells, code, extensions, ...
  - !spack install … (pip, docker, ...)
HZDR data publication system: RODARE

- Decision for Invenio web framework

- Advantages:
  - Actively developed at CERN, also used at e.g. DESY
  - Provides the codebase for Zenodo as well
  - supports deployment via Docker → scalability
Looking at an Example: ROFEX CT
Data flow for ROFEX: Now and future

- **User**
  - Proposal and scheduling
  - Rofex CT
  - Image Reconstruction workflow
  - Data center storage

- **Data Management System**
  - Web/REST upload of raw or image data
  - Metadata
  - Offers workflow stores generated data
  - Convert to HDF5/Nexus

- **Data Flow**
  - Triggers via web form
  - Store raw data
  - Read raw data
  - Store reconstructed images
  - Store raw data via web form
Data management and analytics platform: vision

- Under current development
- Also based on Invenio modules

- Long term goals:
  - Authentication via Umbrella ID for broader audience
  - Tightly linked to web workflows using HPC resources
  - Allow tracking of processing steps for each record along its lifecycle
  - “No metadata has to be typed more than once!”
  - Extract as much metadata as possible from other available sources (e.g. GATE system, eLab-Books, …)
  - Connect this tool with all relevant HZDR facilities
  - Try to interfere with scientists workflow as little as possible
Summary

DM-platform

Experiment 1

Experiment 2

E- lab notes
Control systems

Detector

Web frontend for data upload and workflows

Data publication system RODARE

Long term archive

Data sources

Up to 50 GB/s

Storage (10+ PB)

(Semi) automatic workflows using HPC

Data sources

Data sources

Data sources

Data sources

Data sources
Backup
HZDR Data Policy: Management of Research Data

• The policy is based on the data guidelines of other data initiatives (e.g. PaN-Data, ESRF, BESSY) and the Horizon 2020 “FAIR Data” principles.

• The draft is currently being discussed in a HGF committee as a template for Helmholtz.

• General Principles:
  
  — A **Data Management Plan** (DMP) documents the responsibilities and processes
  
  — Raw data and results are stored in a trusted research data infrastructure (**RODARE**) and kept for at least 10 years. Metadata is defined according to the Data Cite Scheme.
  
  — Access to the data is initially limited to the respective user group for 5 years after completion of the experiment (**embargo period**). The HZDR acts as curator of the data.
  
  — Thereafter, the data of publicly funded research are freely published (**Open Access**). The access and licensing is defined in the DMP. Data publications as well as software publications are to be registered in the publication database (**ROBIS**).
  
  — The policy has been discussed with **user groups** of Elbe and HLD and within the **HGF and EU** projects (CALIPSO+, PanData).