HZDR Data Management Strategy

Meeting at Leibniz Institute of Polymer Research Dresden (IPF)

November 2019

Oliver Knodel, Thomas Gruber and Stefan Müller // contact: o.knodel@hzdr.de

HZDR – Facts and Figures

- **Member of the HELMHOLTZ Association**
  RESEARCH FOR GRAND CHALLENGES

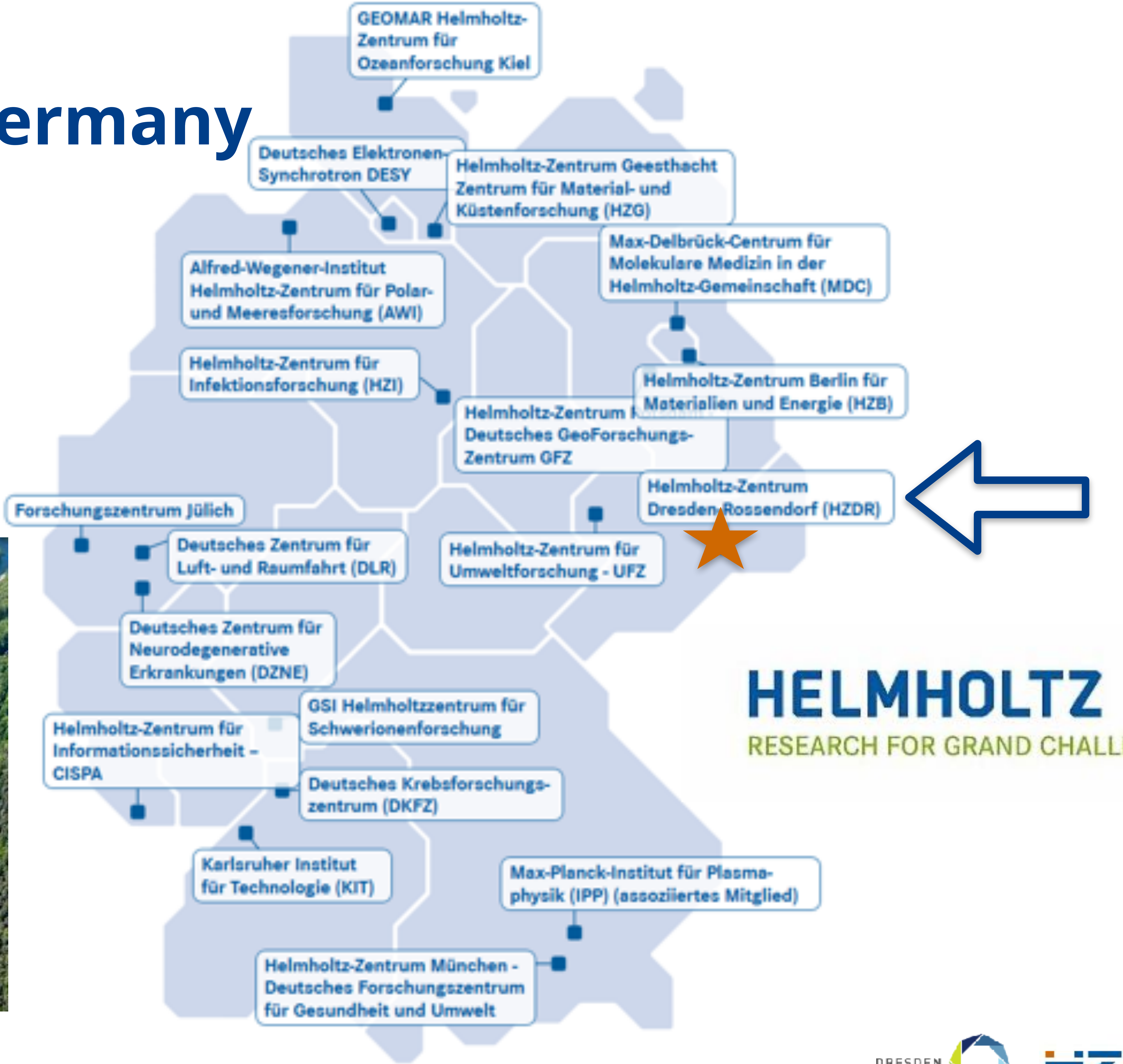
- **Employees approx. 1,200**
  including about 350 scientists
  + 150 doctoral students
  as well as employees and guest
  scientists from more than **50** countries

- **Research Sites**
  DRESDEN
  Leipzig, Freiberg, Schenefeld near Hamburg (XFEL), Grenoble (FR)

Credits: Killig, DESY, ESRF/Ginter

HZDR – Location in Germany

HZDR HELMHOLTZ ZENTRUM DRESDEN ROSSENDORF

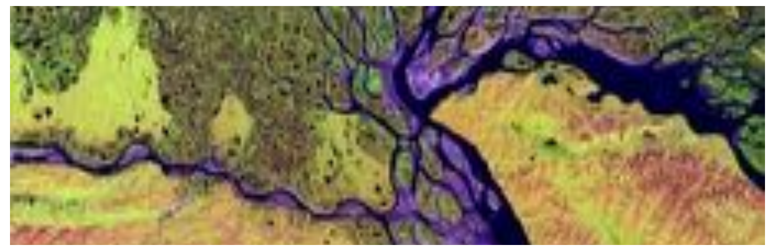

GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel
Deutsches Elektronen-Synchrotron DESY
Helmholtz-Zentrum Geesthacht Zentrum für Material- und Küstenforschung (HZG)
Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung (AWI)
Max-Delbrück-Centrum für Molekulare Medizin in der Helmholtz-Gemeinschaft (MDC)
Helmholtz-Zentrum für Infektionsforschung (HZI)
Helmholtz-Zentrum Berlin für Materialien und Energie (HZB)
Helmholtz-Zentrum Potsdam Deutsches GeoForschungs-Zentrum GFZ
Helmholtz-Zentrum Dresden-Rossendorf (HZDR)
Forschungszentrum Jülich
Deutsches Zentrum für Luft- und Raumfahrt (DLR)
Helmholtz-Zentrum für Umweltforschung - UFZ
Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE)
GSI Helmholtzzentrum für Schwerionenforschung
Helmholtz-Zentrum für Informationssicherheit - CISPA
Deutsches Krebsforschungs-zentrum (DKFZ)
Karlsruher Institut für Technologie (KIT)
Max-Planck-Institut für Plasmaphysik (IPP) (assoziiertes Mitglied)
Helmholtz-Zentrum München - Deutsches Forschungszentrum für Gesundheit und Umwelt

HELMHOLTZ
RESEARCH FOR GRAND CHALLENGES

DRESDEN concept HZDR

The six research fields of the Helmholtz Association

ENERGY ← HZDR



EARTH AND ENVIRONMENT

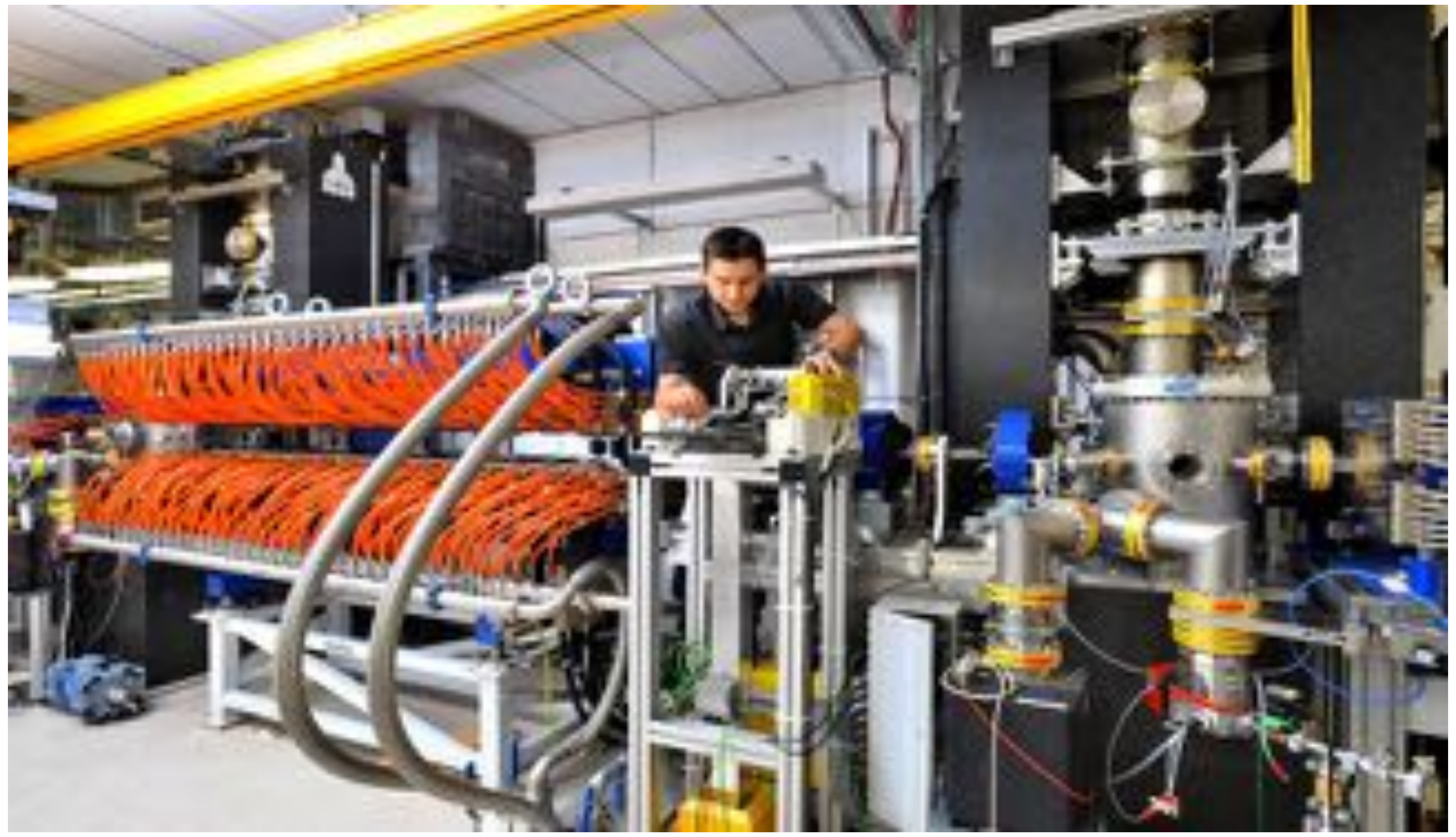
HEALTH ← HZDR

AERONAUTICS, SPACE AND TRANSPORT

KEY TECHNOLOGIES

MATTER ← HZDR

Large Research Infrastructures



Credits: Bierstedt, Killig (2 x)

ELBE – Center for High-Power Radiation Sources

- Electron accelerator ELBE feeds free-electron lasers FELBE & THz source TELBE.
- Generates positrons, protons and neutrons as well as X-ray and gamma radiation.
- High-intensity lasers (1 Petawatt) **DRACO** and **PENELOPE** (under construction).

Dresden High Magnetic Field Laboratory (HLD)

- Nanoscale Producing Europe's highest pulsed magnetic fields.

Ion Beam Center (IBC)

- Nanoscale surface analysis and modification.

One of Our Scientific Computing Challenges

Main Challenge: Pre-/post-processing and archiving of research data

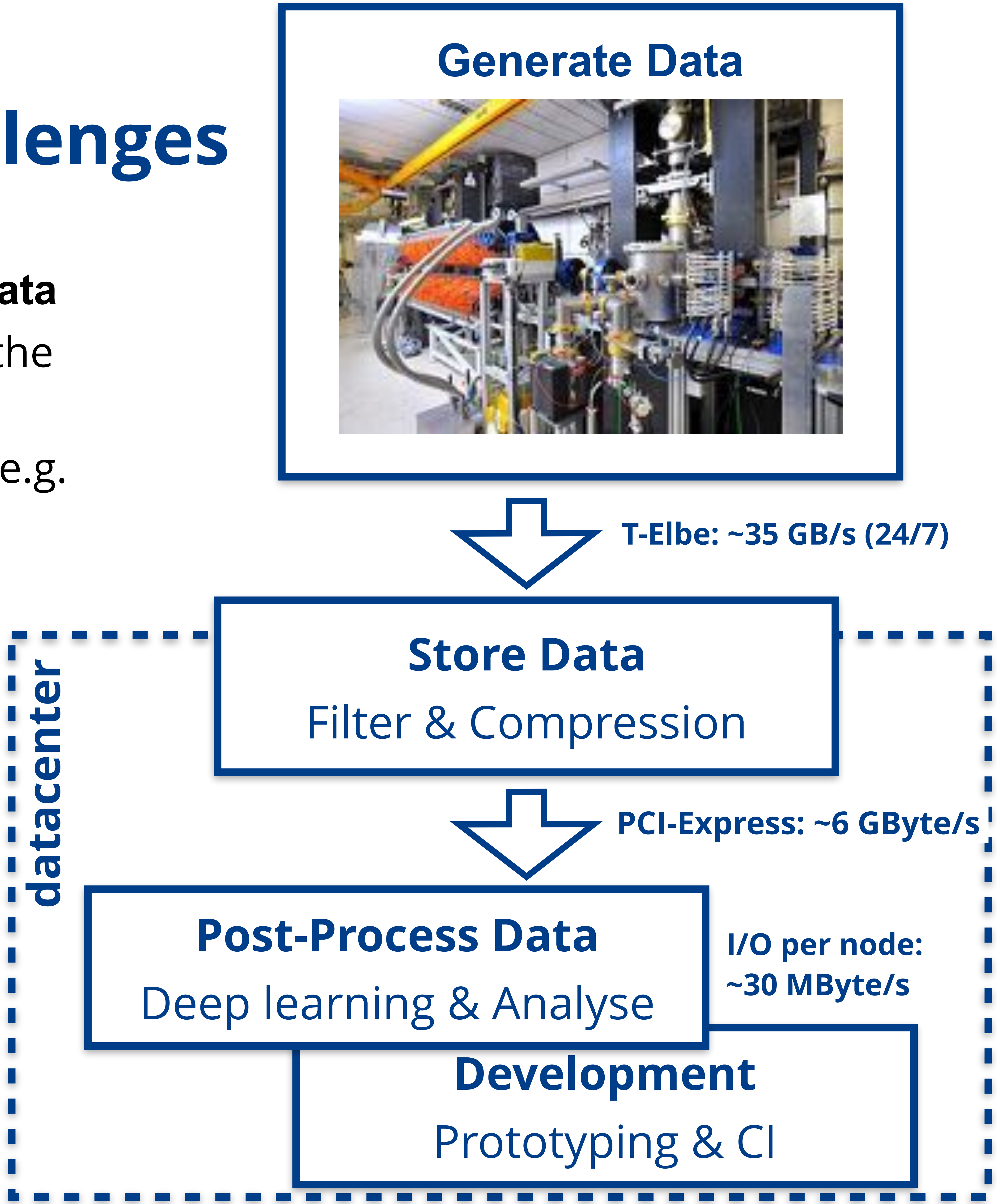
- Filter and compress measured or simulated data at the *edge* of the datacenter.
- Accelerate compute-intensive tasks with dedicated low-latency (e.g. FPGAs), high-bandwidth (e.g. GPUs) hardware.

Heterogenous Data Center

- Many research questions require compute intensive deep learning approaches suitable for our HPC Cluster with GPUs (and FPGAs).
- In the End the research data is located in the data centre anyway.

Prototyping and Continuous Integration (CI)

- We support scientific applications and workflows to improve data processing and even the full software lifecycle.
- Custom CPU, GPU and FPGA applications have to be tested and verified with every development cycle using CI.



Generate Data

T-Elbe: ~35 GB/s (24/7)

datacenter

Store Data
Filter & Compression

PCI-Express: ~6 GByte/s

Post-Process Data
Deep learning & Analyse

I/O per node:
~30 MByte/s

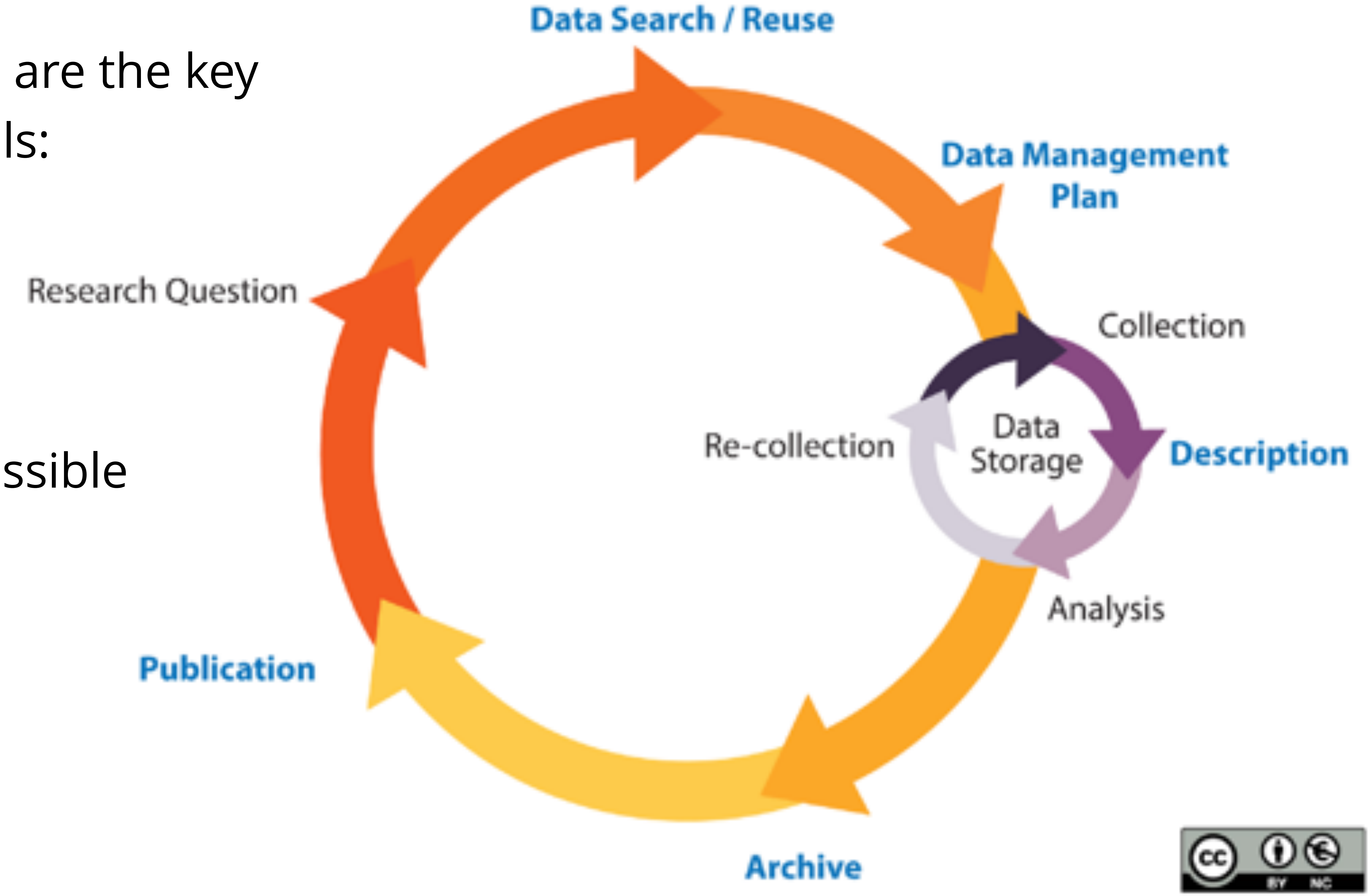
Development
Prototyping & CI

The other Challenge: An End-to-End Digital Data Lifecycle

— Data/Meta-Data standards are the key

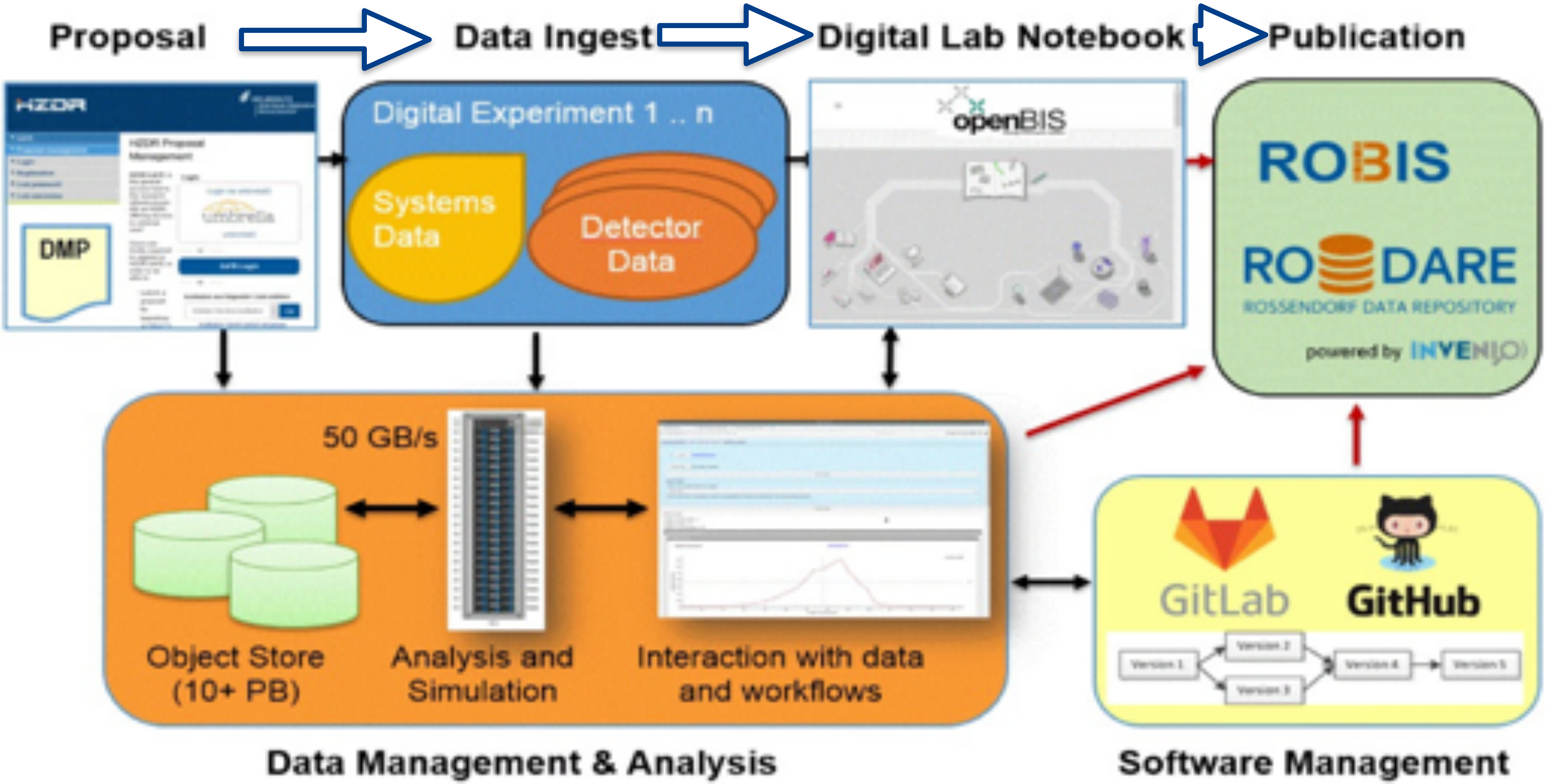
— Support all stages with tools:
  - electronic lab books,
  - interactive analysis,
  - automated publication,
  - workflow management.

— Get the data as early as possible into the data center.



Research Data Management Lifecycle taken from: https://guides.library.ucsc.edu/datamanagement

HZDR Data Management Strategy



Proposal

Data Ingest

Digital Lab Notebook

Publication

DMP

Digital Experiment 1 .. n

Systems Data

Detector Data

openBIS

ROBIS

RODARE

ROSSENDORF DATA REPOSITORY

powered by INVENIO

50 GB/s

Object Store (10+ PB)

Analysis and Simulation

Interaction with data and workflows

GitLab

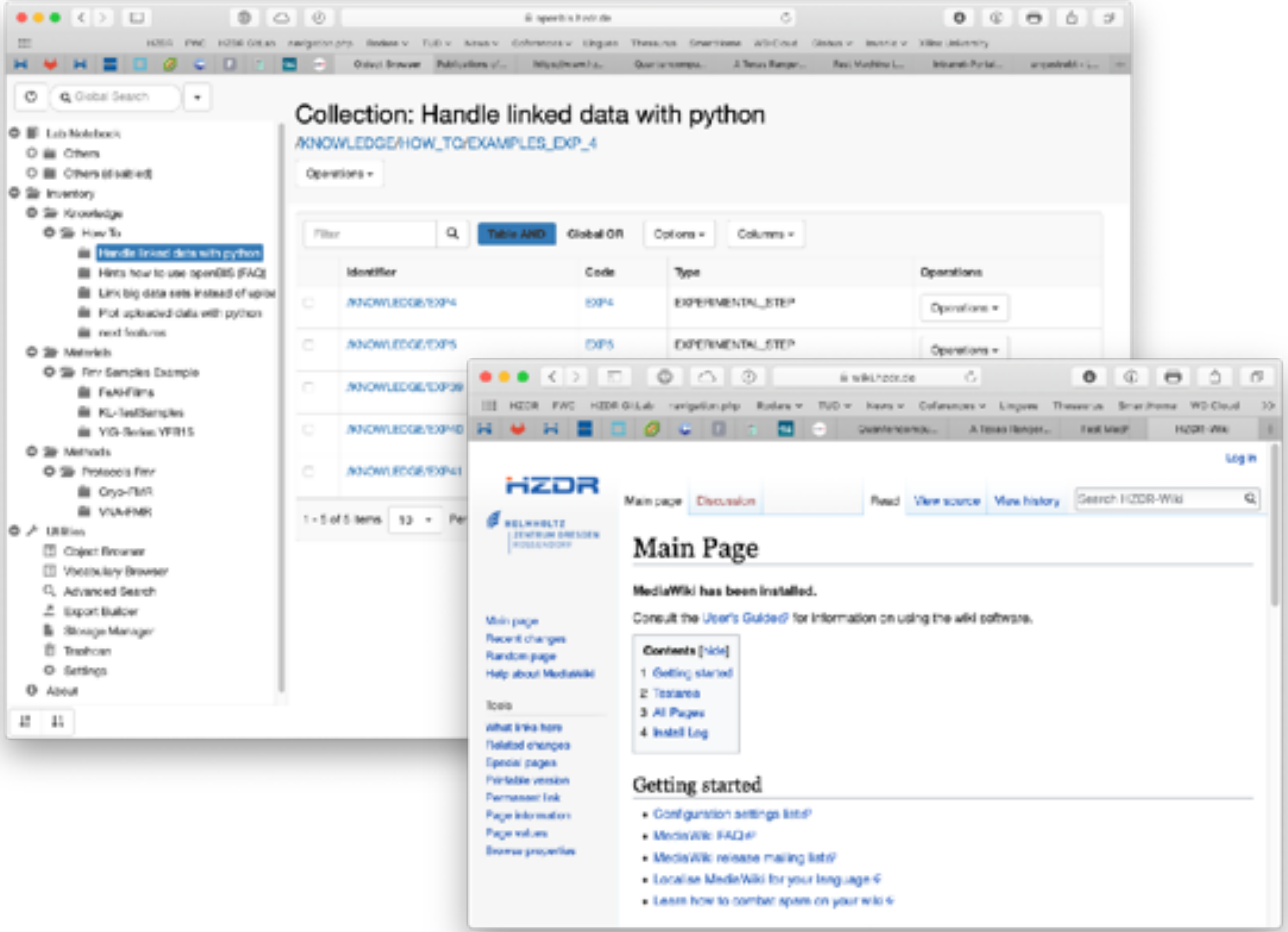
GitHub

Data Management & Analysis

Software Management

Electronic Lab Books for Better Meta-Data Management

- Long Evaluation Phase:
- Result:
  - **OpenBis** for structured Lab-Data.
  - **MediaWiki** for more free-form Documentation.
- Both are necessary to meet the requirements of the experiments at HZDR.



Lab Notebook
Others
Others (disabled)
Inventory
Knowledge
How To
Handle linked data with python
Hints how to use openBIS (FAQ)
Link big data sets instead of upload
Plot uploaded data with python
next features
Materials
Fmr Samples Example
FeNiFilms
KL-TestSamples
YIG-Series VFR15
Methods
Protocols Fmr
Cryo-FMR
VNA-FMR
Utilities
Object Browser
Vocabulary Browser
Advanced Search
Export Builder
Storage Manager
Trashcan
Settings
About

Collection: Handle linked data with python

/KNOWLEDGE/HOW_TO/EXAMPLES_EXP_4

	Identifier	Code	Type	Operations
□	/KNOWLEDGE/EXP4	EXP4	EXPERIMENTAL_STEP	Operations ▾
□	/KNOWLEDGE/EXP5	EXP5	EXPERIMENTAL_STEP	Operations ▾
□	/KNOWLEDGE/EXP39			
□	/KNOWLEDGE/EXP40			
□	/KNOWLEDGE/EXP41			

1 - 5 of 5 items 10

Main Page

MediaWiki has been installed.

Consult the User's Guide for information on using the wiki software.

Contents [hide]
1 Getting started
2 Testarea
3 All Pages
4 Install Log

Getting started

- Configuration settings list
- MediaWiki FAQ
- MediaWiki release mailing list
- Localise MediaWiki for your language
- Learn how to combat spam on your wiki

Workflow Engine for Scientific Workflows

COMMON WORKFLOW LANGUAGE

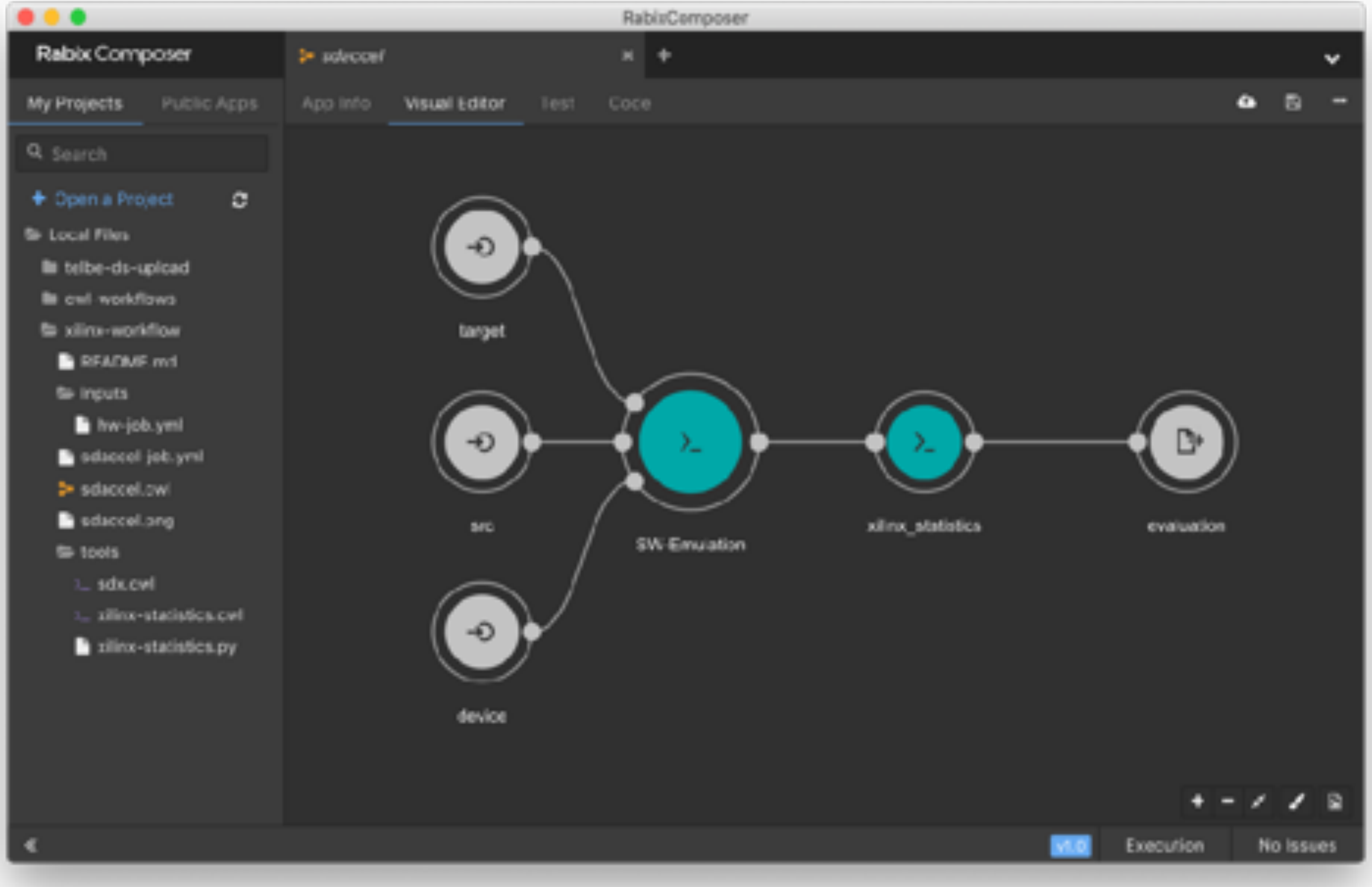
— The execution of scientific workflows must be:

- Comprehensibly
- Archivable,
- Reusable,
- Reproducible,
- Publishable.

— Based on an open standard: **OWL** or **WDL**

— Our evaluation is still in progress:

- **Reana** from Cern,
- Model-based approach from University **Turin**,
- **Knime** as stand-alone Product.



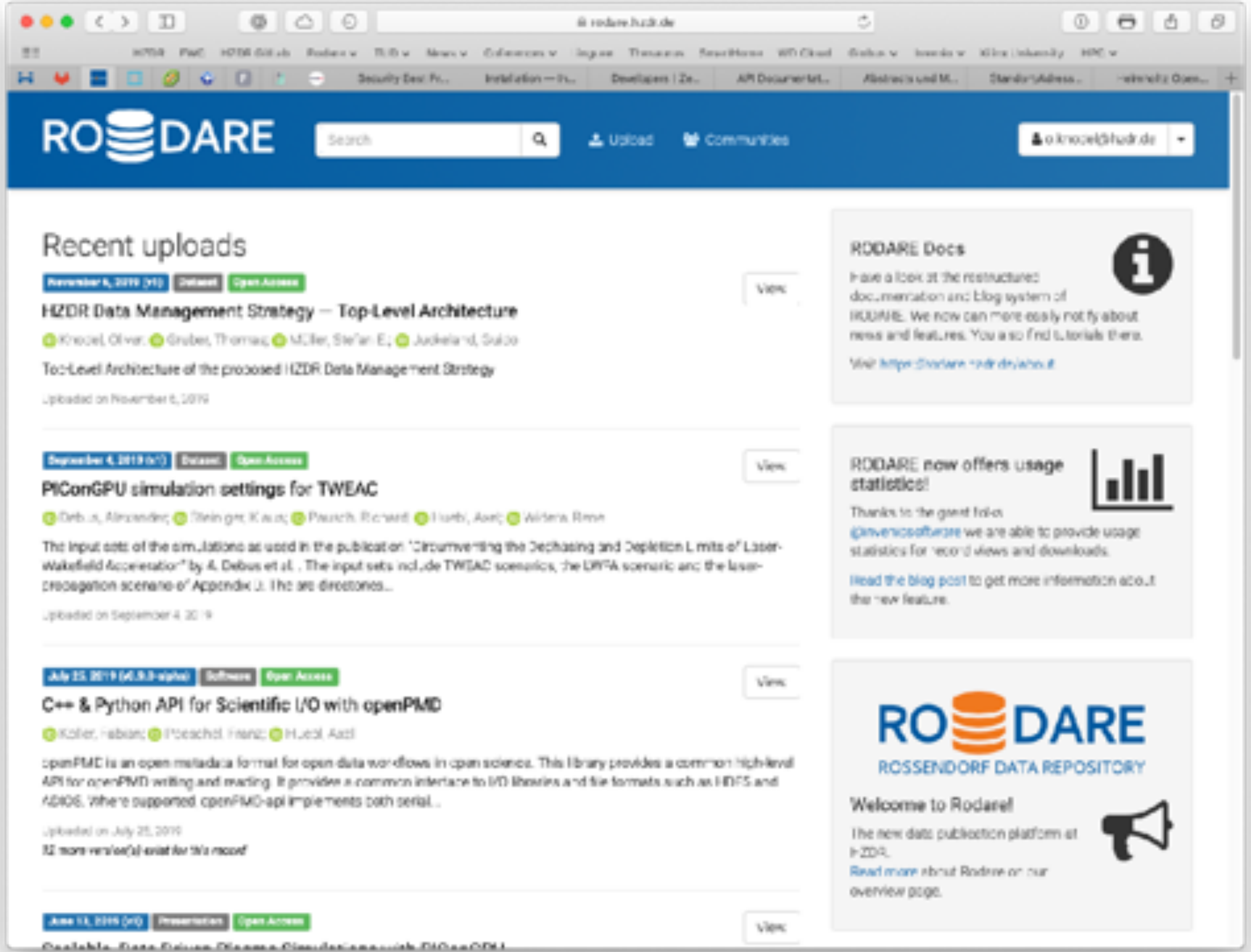
RODARE - The HZDR Publication Platform

— Based on **Invenio** from CERN

— Highly modular and proven

- Own contributions:
- Shibboleth authenticator
- SFTP File Browser/Uploader
- Bittorrent Downloader
- GitLab-Integration

— HZDR is Part of CERN Community Collaboration Project



Recent uploads

November 6, 2019 (v1) Dataset Open Access

HZDR Data Management Strategy — Top-Level Architecture

Knodel, Oliver; Gruber, Thomas; Müller, Stefan E.; Juckeland, Guido

Top-Level Architecture of the proposed HZDR Data Management Strategy

Uploaded on November 6, 2019

September 4, 2019 (v1) Dataset Open Access

PIConGPU simulation settings for TWEAC

Debus, Alexander; Steiniger, Klaus; Pausch, Richard; Huebl, Axel; Widera, Rene

The input sets of the simulations as used in the publication "Circumventing the Dephasing and Depletion Limits of Laser-Wakefield Acceleration" by A. Debus et al. . The input sets include TWEAC scenarios, the LWFA scenario and the laser-propagation scenario of Appendix D. The are directories...

Uploaded on September 4, 2019

July 25, 2019 (v0.9.0-alpha) Software Open Access

C++ & Python API for Scientific I/O with openPMD

Koller, Fabian; Poeschel, Franz; Huebl, Axel

openPMD is an open metadata format for open data workflows in open science. This library provides a common high-level API for openPMD writing and reading. It provides a common interface to I/O libraries and file formats such as HDF5 and ADIOS. Where supported, openPMD-api implements both serial...

Uploaded on July 25, 2019

32 more version(s) exist for this record

June 13, 2019 (v1) Presentation Open Access

Scalable, Data-Driven Plasma Simulations with PIConGPU

RODARE Docs

Have a look at the restructured documentation and blog system of RODARE. We now can more easily notify about news and features. You also find tutorials there.

Visit https://rodare.hzdr.de/about

RODARE now offers usage statistics!

Thanks to the great folks @inveniosoftware we are able to provide usage statistics for record views and downloads.

Read the blog post to get more information about the new feature.

RODARE
ROSSENDORF DATA REPOSITORY

Welcome to Rodare!

The new data publication platform at HZDR.
Read more about Rodare on our overview page.

IT-Hosted Services for Collaborative Work

- The **Department of Information Services and Computing** supports all the institutes as well as external users with a wide range of IT-services:

GitLab

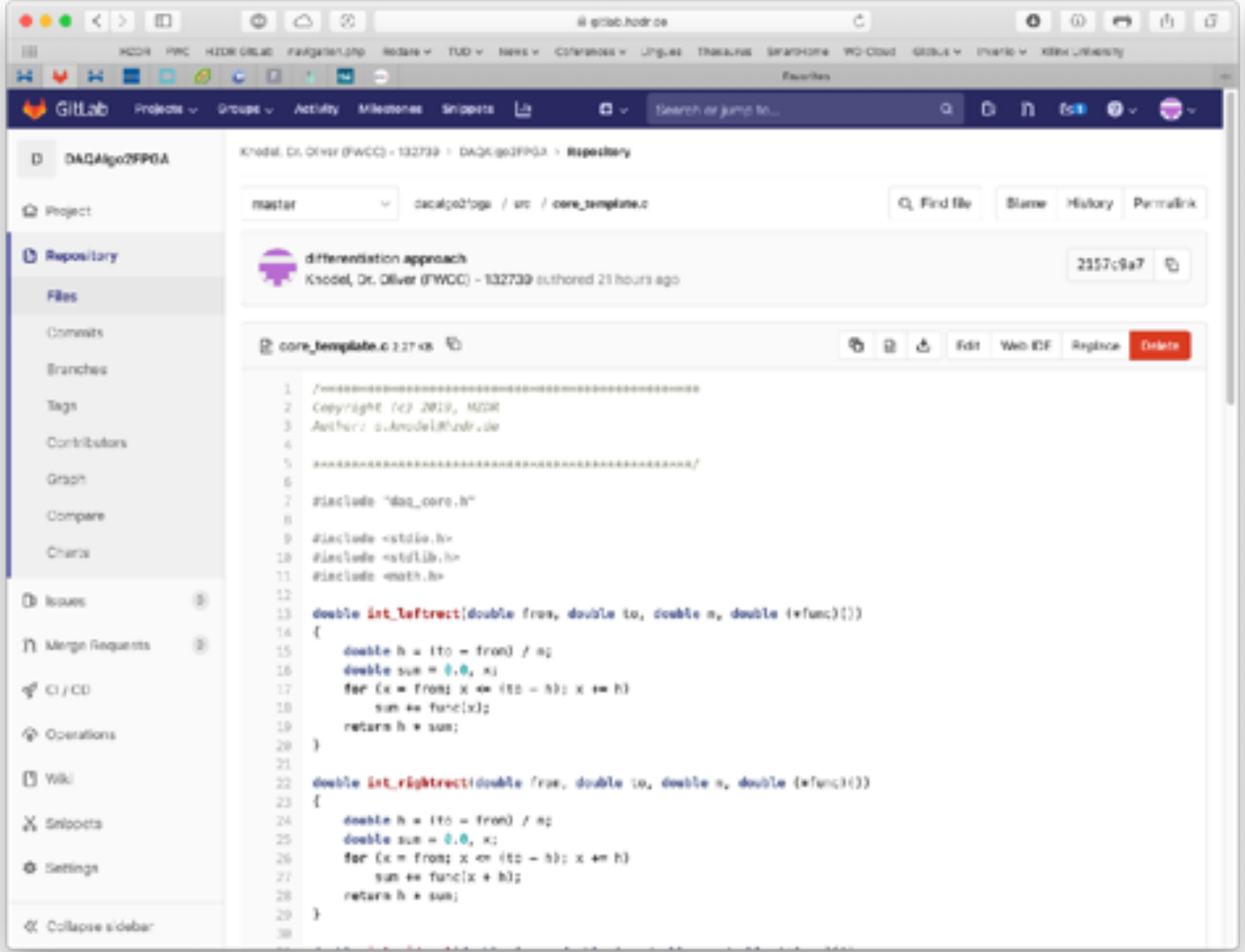
ownCloud

jupyter

ShareLaTeX

REDMINE flexible project management

Mattermost



differentiation approach
Knodel, Dr. Oliver (FWCC) - 132739 authored 21 hours ago

core_template.c 2.27 KB

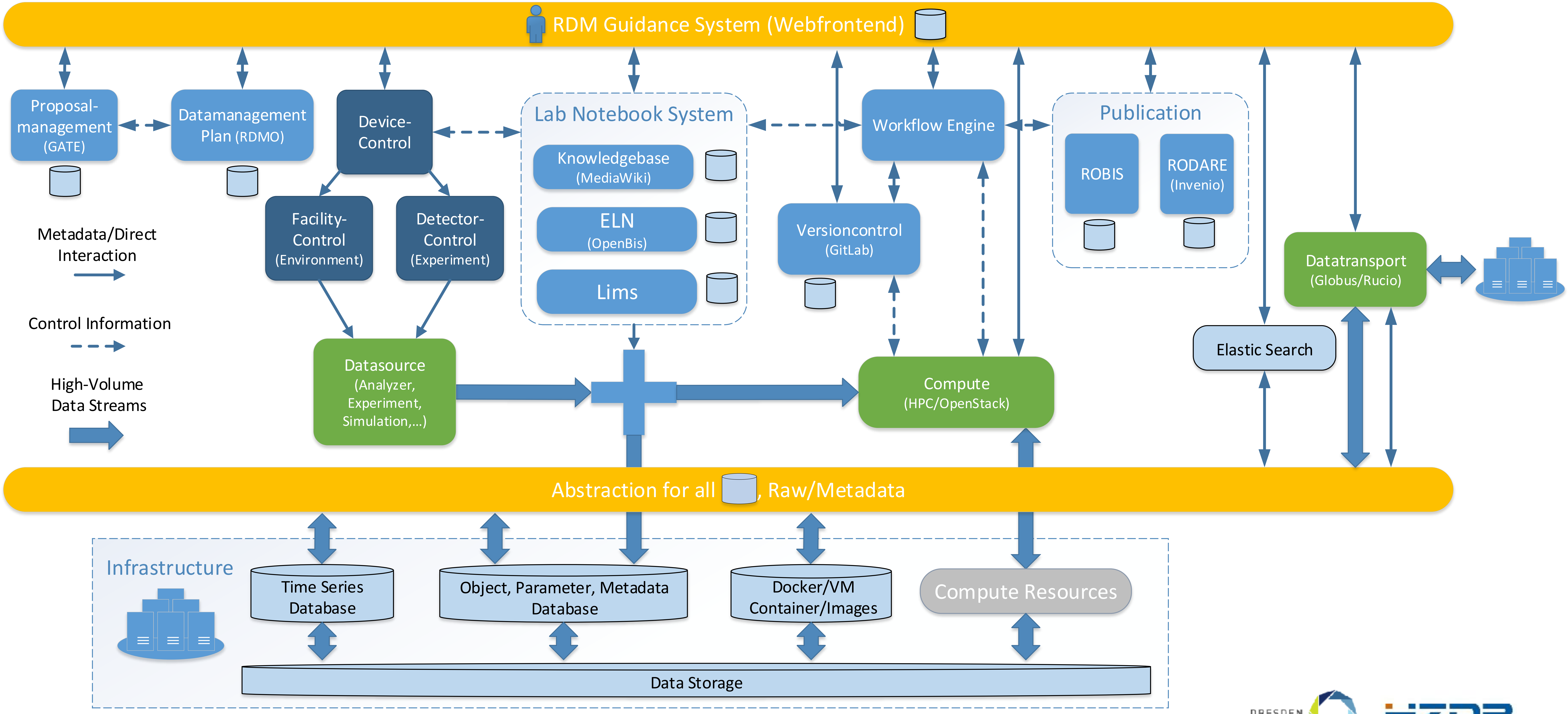
```
/*****************************************************
Copyright (c) 2019, HZDR
Author: o.knodel@hzdr.de

*****************************************************/

#include "daq_core.h"

#include <stdio.h>
#include <stdlib.h>
#include <math.h>

double int_leftrect(double from, double to, double n, double (*func)())
{
    double h = (to - from) / n;
    double sum = 0.0, x;
    for (x = from; x <= (to - h); x += h)
        sum += func(x);
    return h * sum;
}

double int_rightrect(double from, double to, double n, double (*func)())
{
    double h = (to - from) / n;
    double sum = 0.0, x;
    for (x = from; x <= (to - h); x += h)
        sum += func(x + h);
    return h * sum;
}
```

DRESDEN concept

HZDR

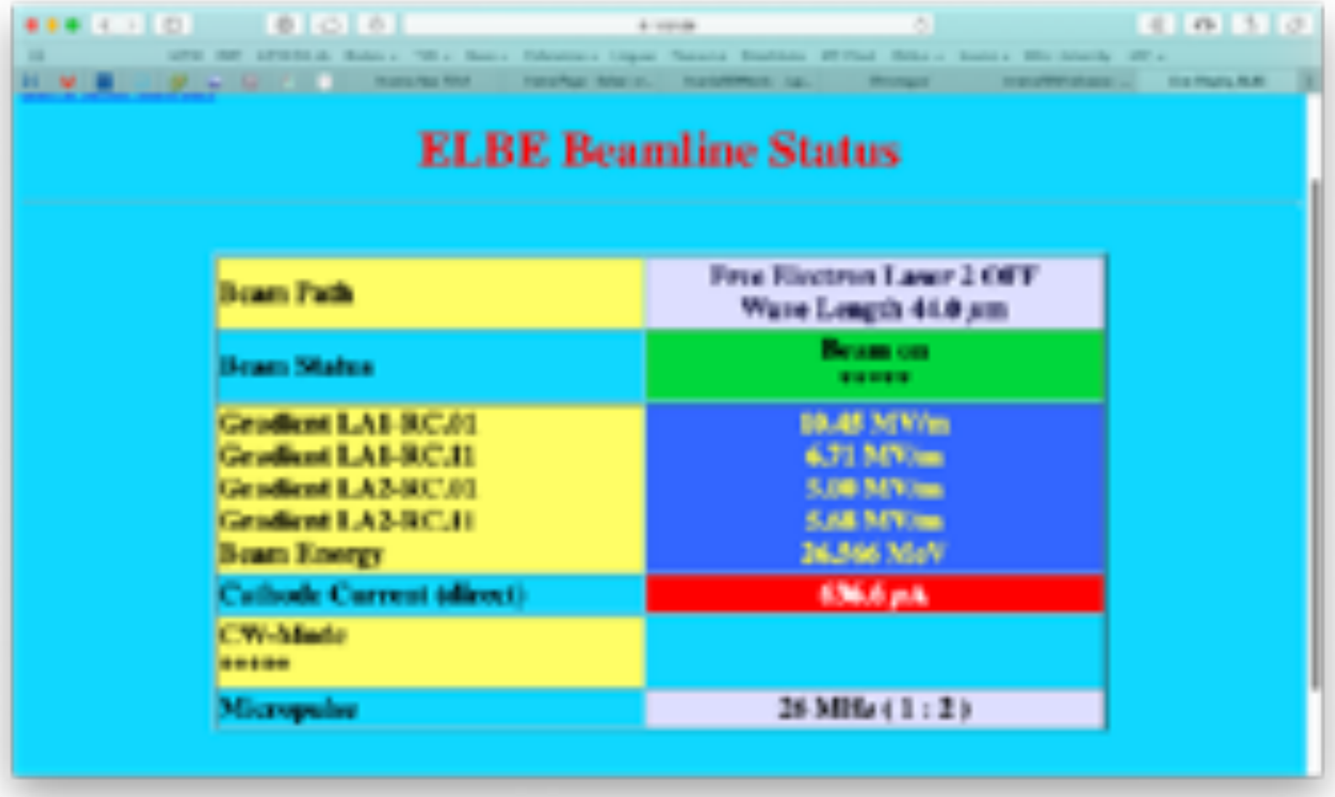
DOI 10.14278/rodare.194

HZDR Data Management Strategy — Top Level Architecture


RDM Guidance System (Webfrontend)
Proposal-management (GATE)
Datamanagement Plan (RDMO)
Device-Control
Facility-Control (Environment)
Detector-Control (Experiment)
Lab Notebook System
Knowledgebase (MediaWiki)
ELN (OpenBis)
Lims
Workflow Engine
Versioncontrol (GitLab)
Publication
ROBIS
RODARE (Invenio)
Datatransport (Globus/Rucio)
Elastic Search
Metadata/Direct Interaction
Control Information
High-Volume Data Streams
Datasource (Analyzer, Experiment, Simulation,...)
Compute (HPC/OpenStack)
Abstraction for all , Raw/Metadata
Infrastructure
Time Series Database
Object, Parameter, Metadata Database
Docker/VM Container/Images
Compute Resources
Data Storage

DRESDEN concept HZDR

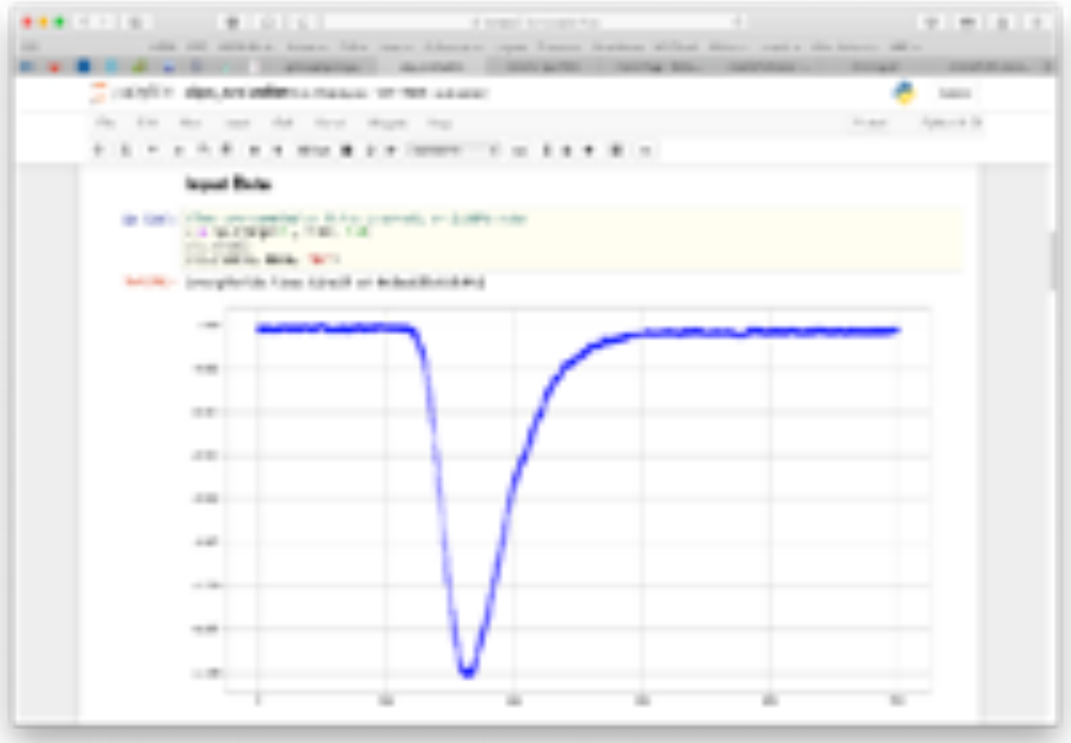
HZDR Data Management Strategy — Possible Dataflow



ELBE Beamline Status

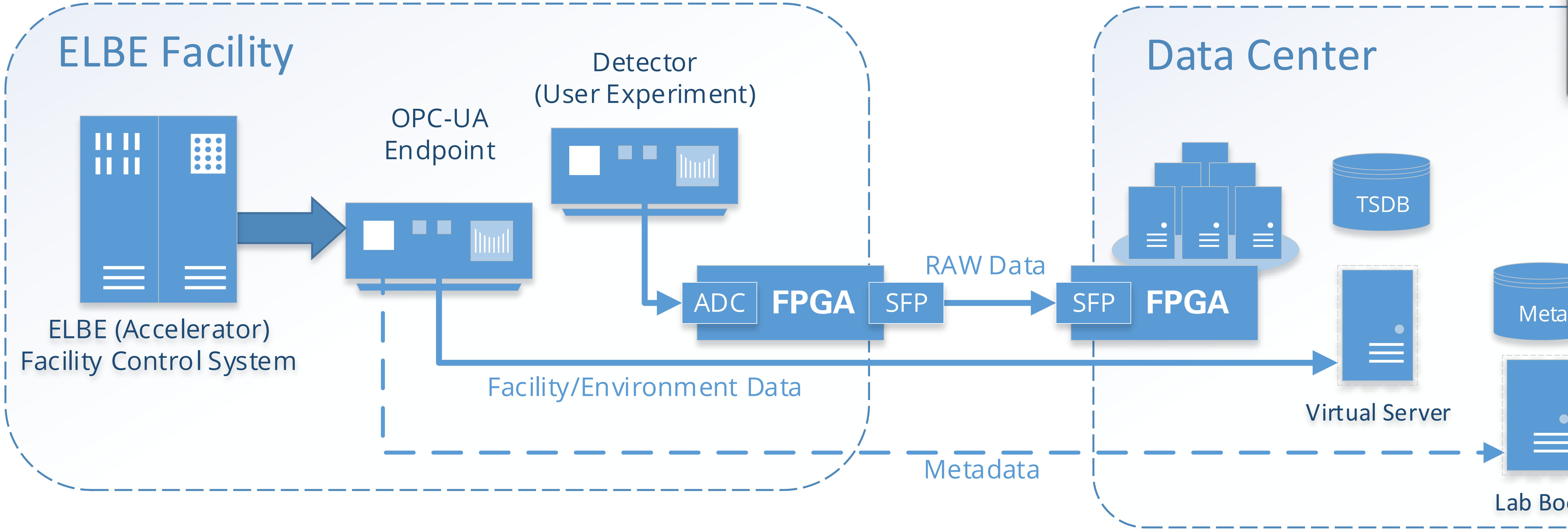
Beam Path	Free Electron Laser 2 OFF Wave Length 44.0 µm
Beam Status	Beam on *****
Gradient LA1-RC.01 Gradient LA1-RC.11 Gradient LA2-RC.01 Gradient LA2-RC.11 Beam Energy	10.45 MV/m 6.71 MV/m 5.00 MV/m 5.68 MV/m 26.566 MeV
Cathode Current (direct)	636.6 µA
CW-Mode *****	
Micropulse	26 MHz (1 : 2)

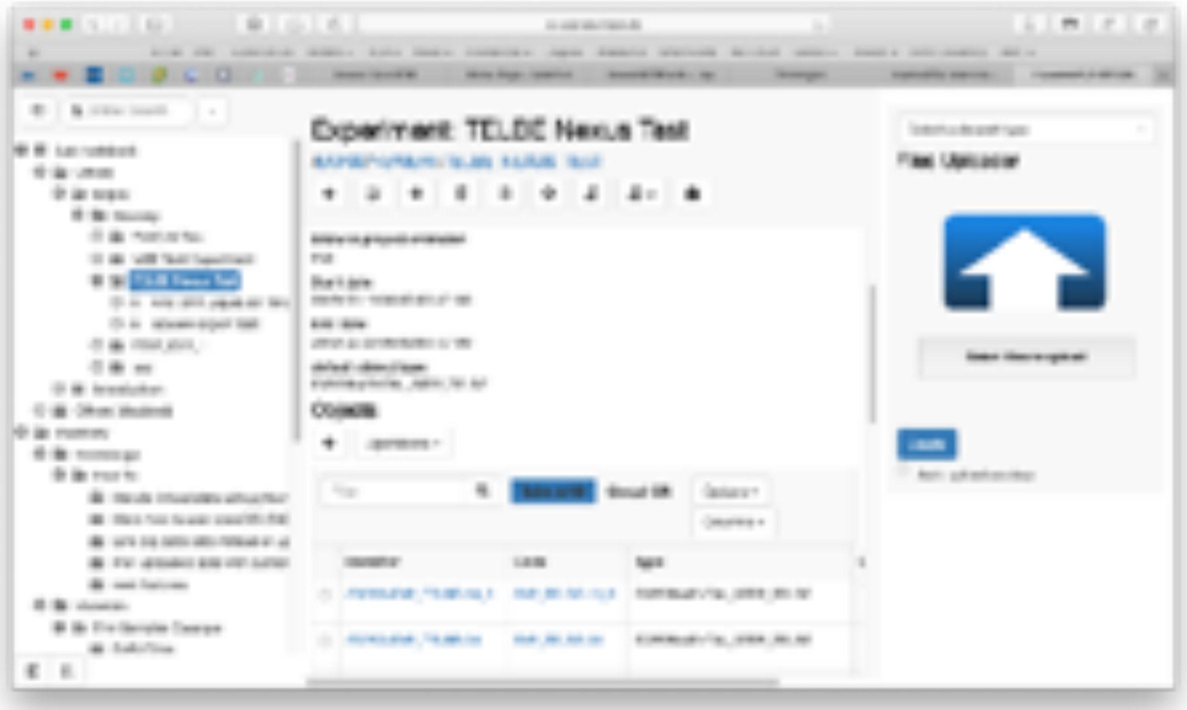
Live Beamline Monitor



RAW Data

Time-Series Data

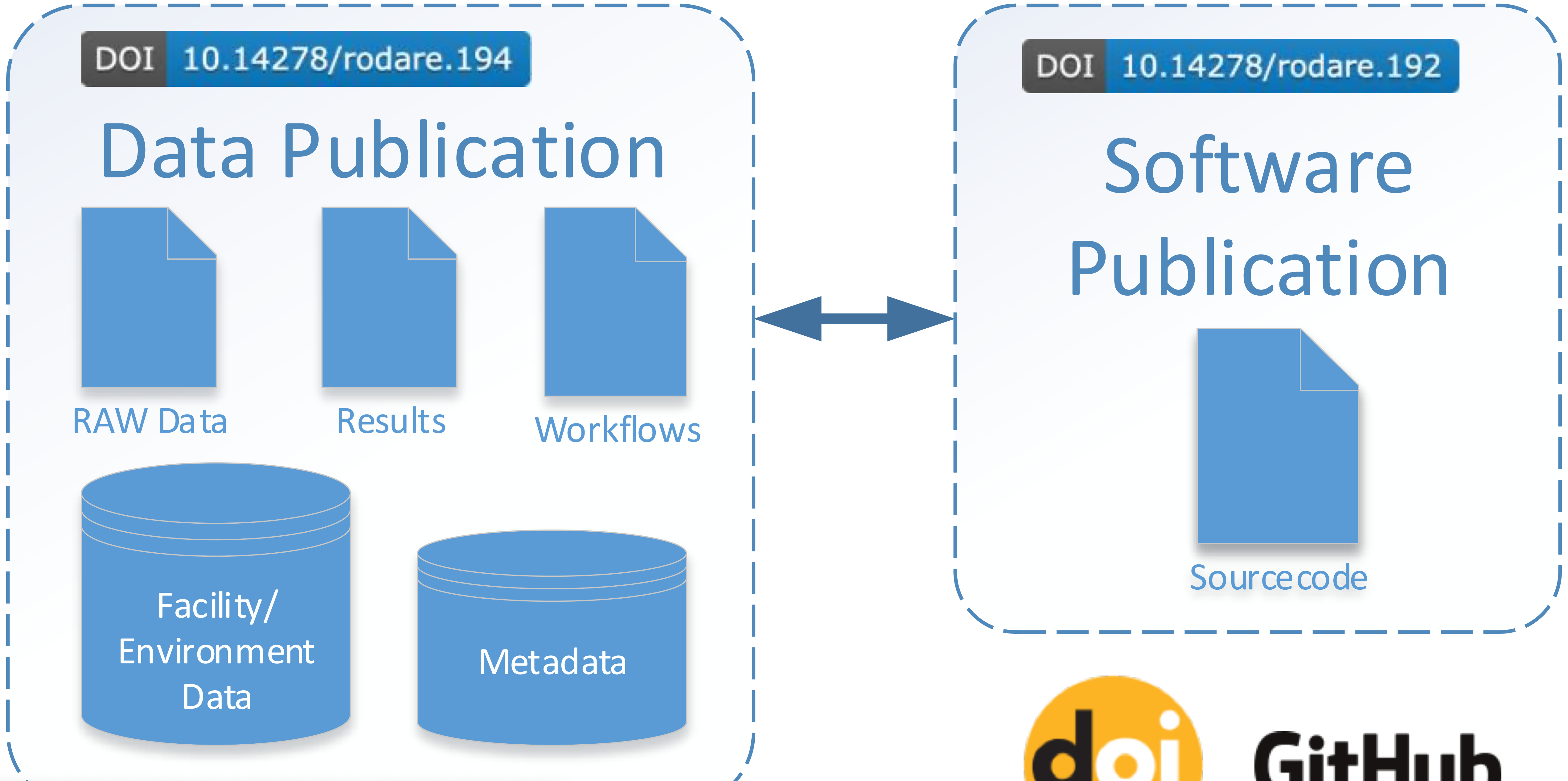


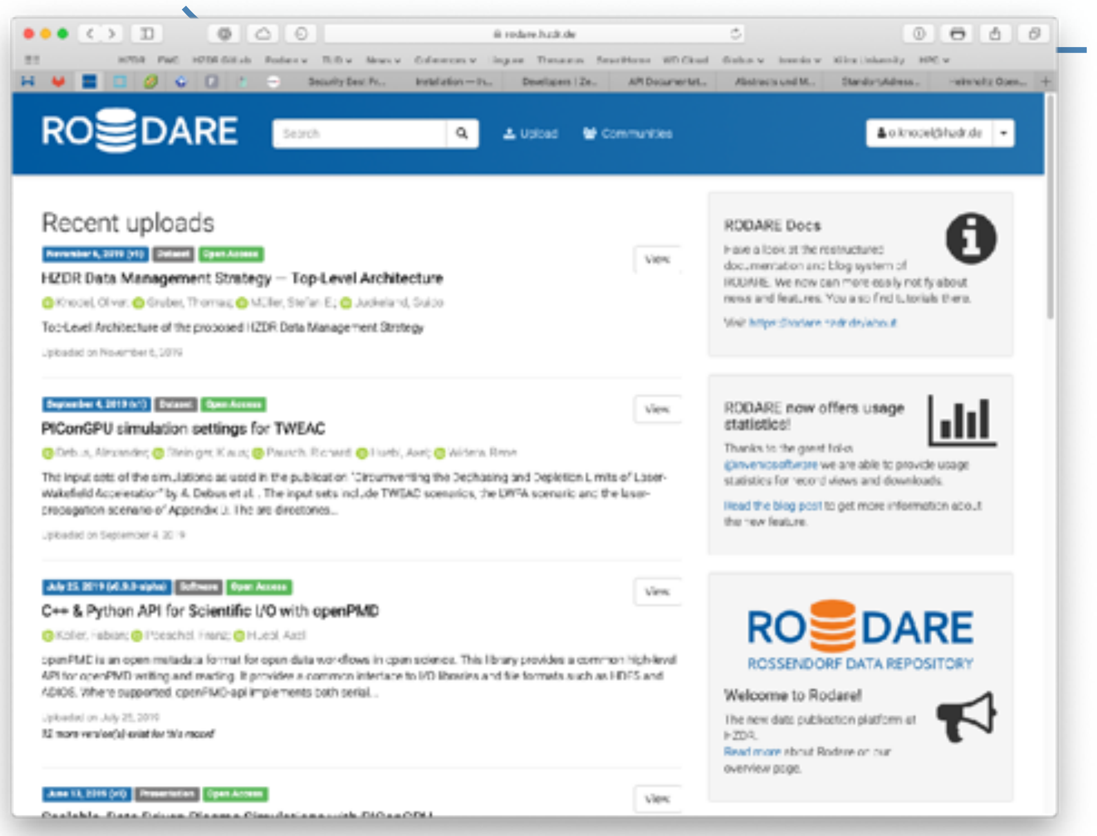


Structured (Meta)data

HZDR Data Management Strategy — Data Publication

- Data/Meta-Data standards are the key of a usable data publication (e.g. NeXus, CERIF, …).
- **Metadata:** the who, what, when, where, why, how of your research.
- All data generated during the experiment:
  - RAW and
  - Facility Data,
  - Results/Analysis,
  - Workflows and
  - Metadata from the Lab Book.





Recent uploads

November 6, 2019 (v1) Dataset Open Access
HZDR Data Management Strategy — Top-Level Architecture
Knodel, Oliver; Gruber, Thomas; Müller, Stefan E.; Juckeland, Guido
Top-Level Architecture of the proposed HZDR Data Management Strategy
Uploaded on November 6, 2019

September 4, 2019 (v1) Dataset Open Access
PIConGPU simulation settings for TWEAC
Debus, Alexander; Steiniger, Klaus; Pausch, Richard; Huebl, Axel; Widera, Rene
The input sets of the simulations as used in the publication "Circumventing the Dephasing and Depletion Limits of Laser-Wakefield Acceleration" by A. Debus et al. . The input sets include TWEAC scenarios, the DWFA scenario and the laser-propagation scenario of Appendix D. The are directories...
Uploaded on September 4, 2019

July 25, 2019 (v0.9.0-alpha) Software Open Access
C++ & Python API for Scientific I/O with openPMD
Koller, Fabian; Poeschel, Franz; Huebl, Axel
openPMD is an open metadata format for open data workflows in open science. This library provides a common high-level API for openPMD writing and reading. It provides a common interface to I/O libraries and file formats such as HDF5 and ADIOS. Where supported, openPMD-api implements both serial...
Uploaded on July 25, 2019
12 more version(s) exist for this record

June 13, 2019 (v1) Presentation Open Access
Scalable, Data-Driven Plasma Simulations with PIConGPU

RODARE Docs
Have a look at the restructured documentation and blog system of RODARE. We now can more easily notify about news and features. You also find tutorials there.
Visit https://rodare.hzdr.de/about

RODARE now offers usage statistics!
Thanks to the great folks @inveniosoftware we are able to provide usage statistics for record views and downloads.
Read the blog post to get more information about the new feature.

RODARE
ROSSENDORF DATA REPOSITORY
Welcome to Rodare!
The new data publication platform at HZDR.
Read more about Rodare on our overview page.

doi GitHub ORCID Connecting Research and Researchers

DRESDEN concept HZDR