



© HZDR/Detlev Müller

Research Data Management at HZDR

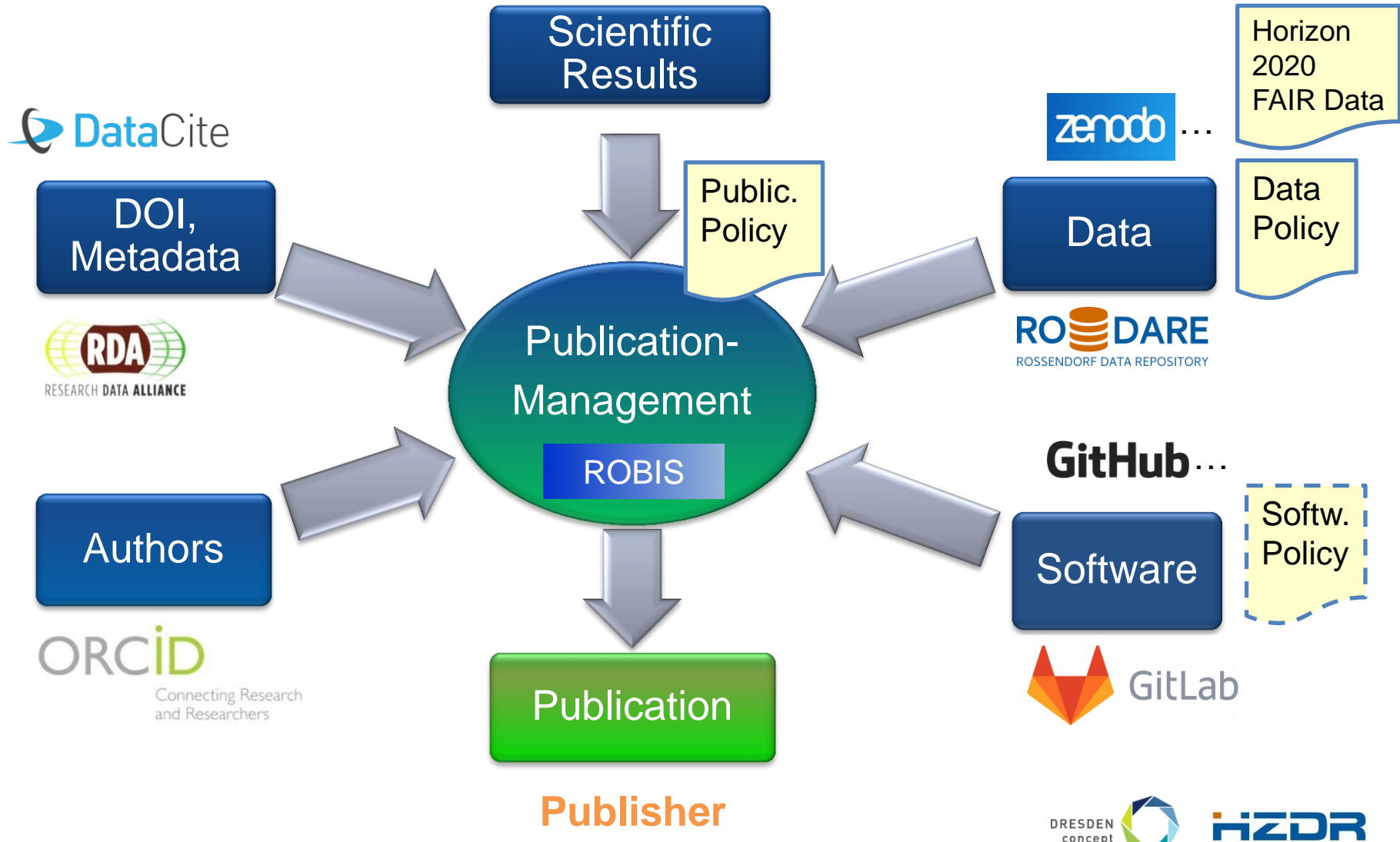


hzdr



HELMHOLTZ
ZENTRUM DRESDEN
ROSSENDORF

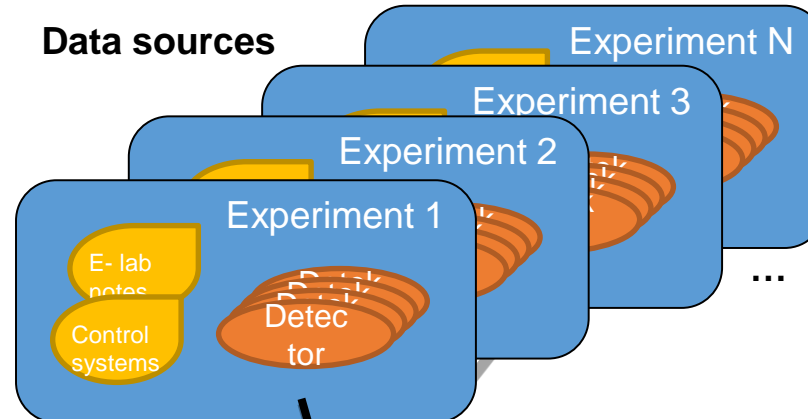
Publication Components



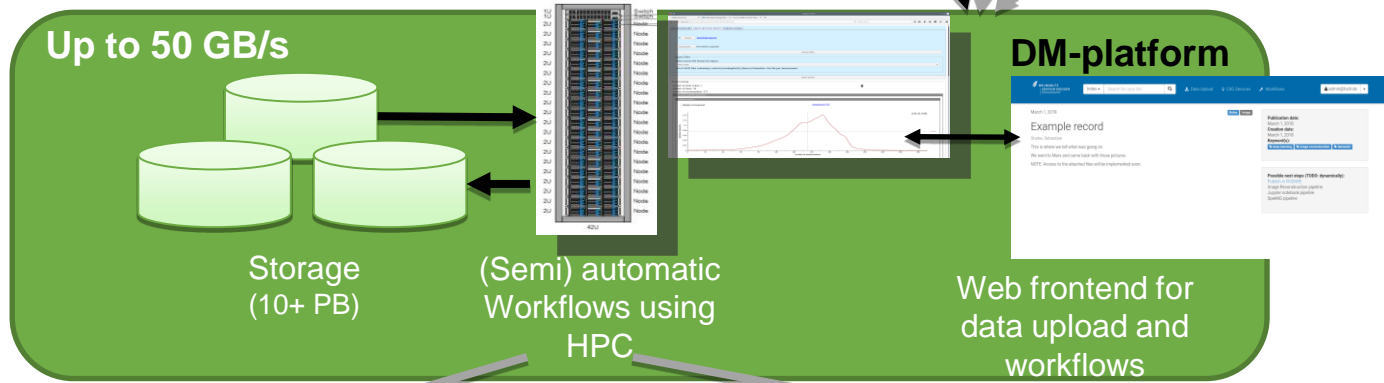
The Concept

Basic DM Concept

Data sources



- 8 kHz CT
- 1,5 MHz Cameras
- HPC Simulations
- ...



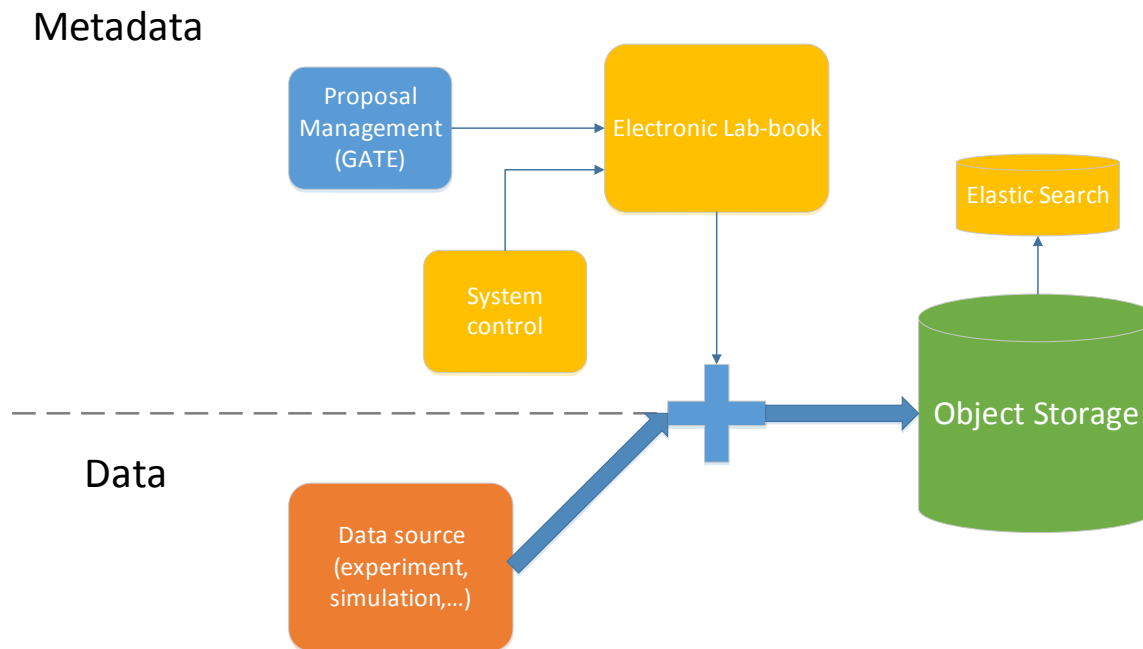
Long term archive



Data publication system RODARE

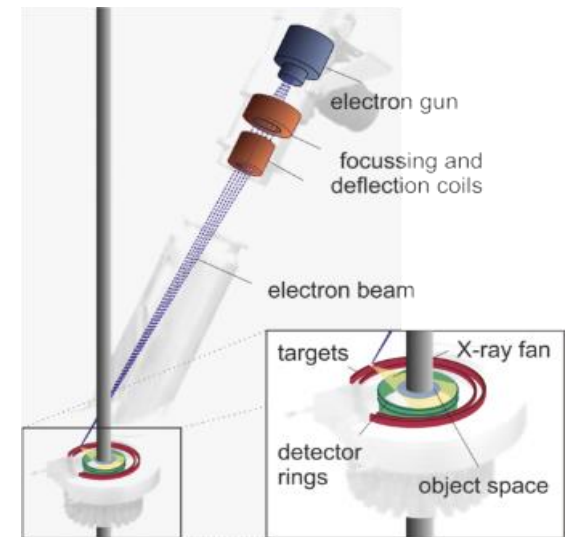
Data Ingest

- 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)

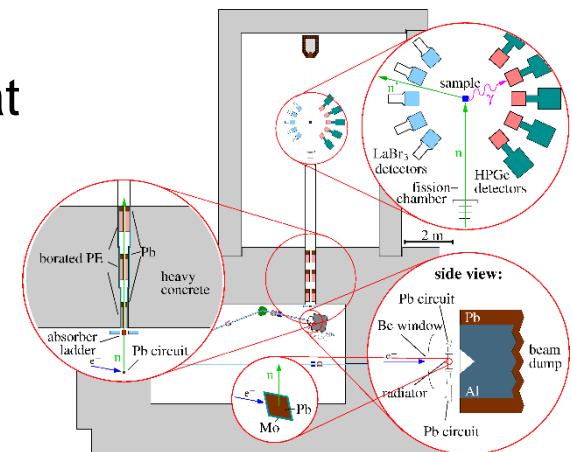


How to deal with heterogeneous metadata schemes?

- 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)



(Fischer et al., 2008)



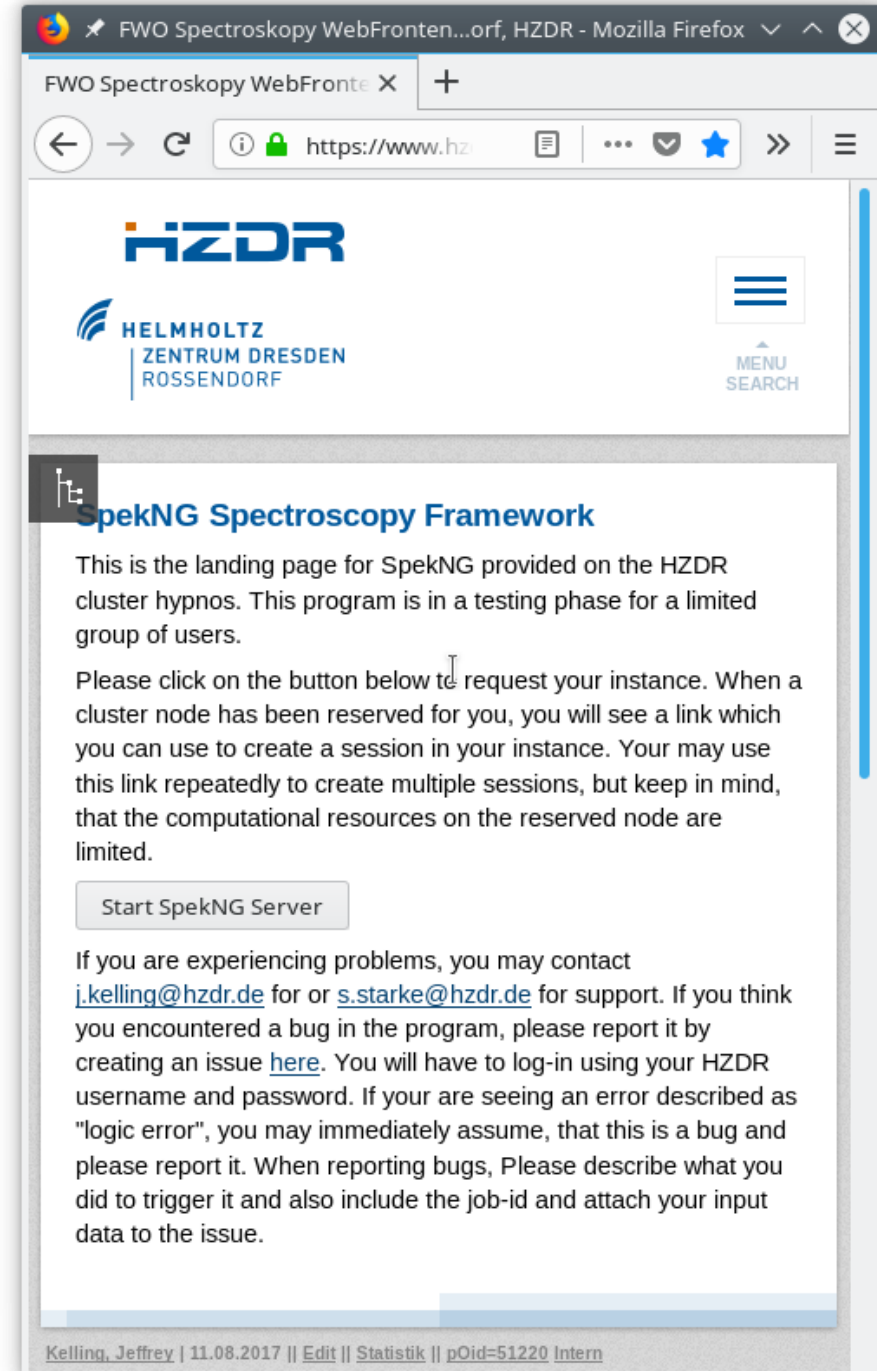
HZDR Data Management

- 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



The screenshot shows a Mozilla Firefox browser window with the URL <https://www.hzdr.de>. The page header features the HZDR logo (Helmholtz Zentrum Dresden Rosendorf) and a navigation menu. The main content area is titled "SpekNG Spectroscopy Framework" and contains the following text:

This is the landing page for SpekNG provided on the HZDR cluster hypnos. This program is in a testing phase for a limited group of users.

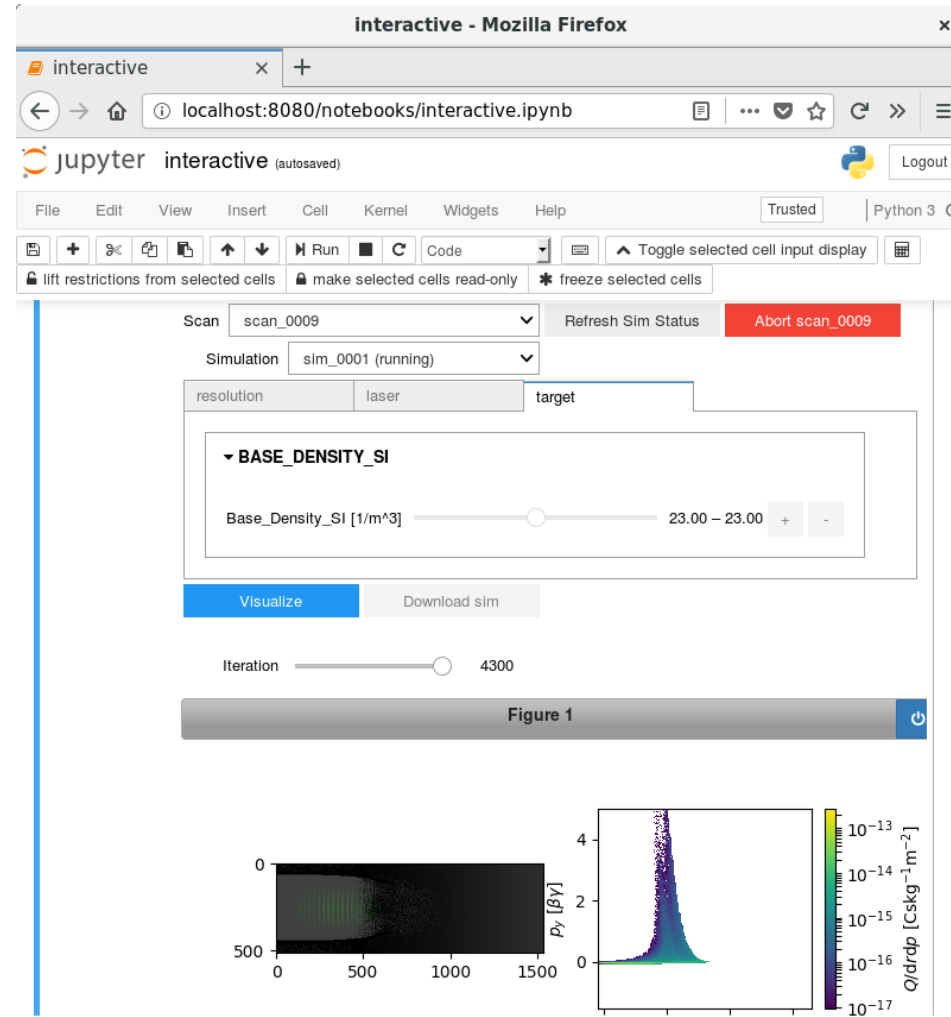
Please click on the button below to request your instance. When a cluster node has been reserved for you, you will see a link which you can use to create a session in your instance. You may use this link repeatedly to create multiple sessions, but keep in mind, that the computational resources on the reserved node are limited.

If you are experiencing problems, you may contact j.kelling@hzdr.de for or s.starke@hzdr.de for support. If you think you encountered a bug in the program, please report it by creating an issue [here](#). You will have to log-in using your HZDR username and password. If you are seeing an error described as "logic error", you may immediately assume, that this is a bug and please report it. When reporting bugs, Please describe what you did to trigger it and also include the job-id and attach your input data to the issue.

At the bottom of the page, there is a footer with the text: "Kelling, Jeffrey | 11.08.2017 || Edit || Statistik || pOid=51220 Intern".

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, ...)



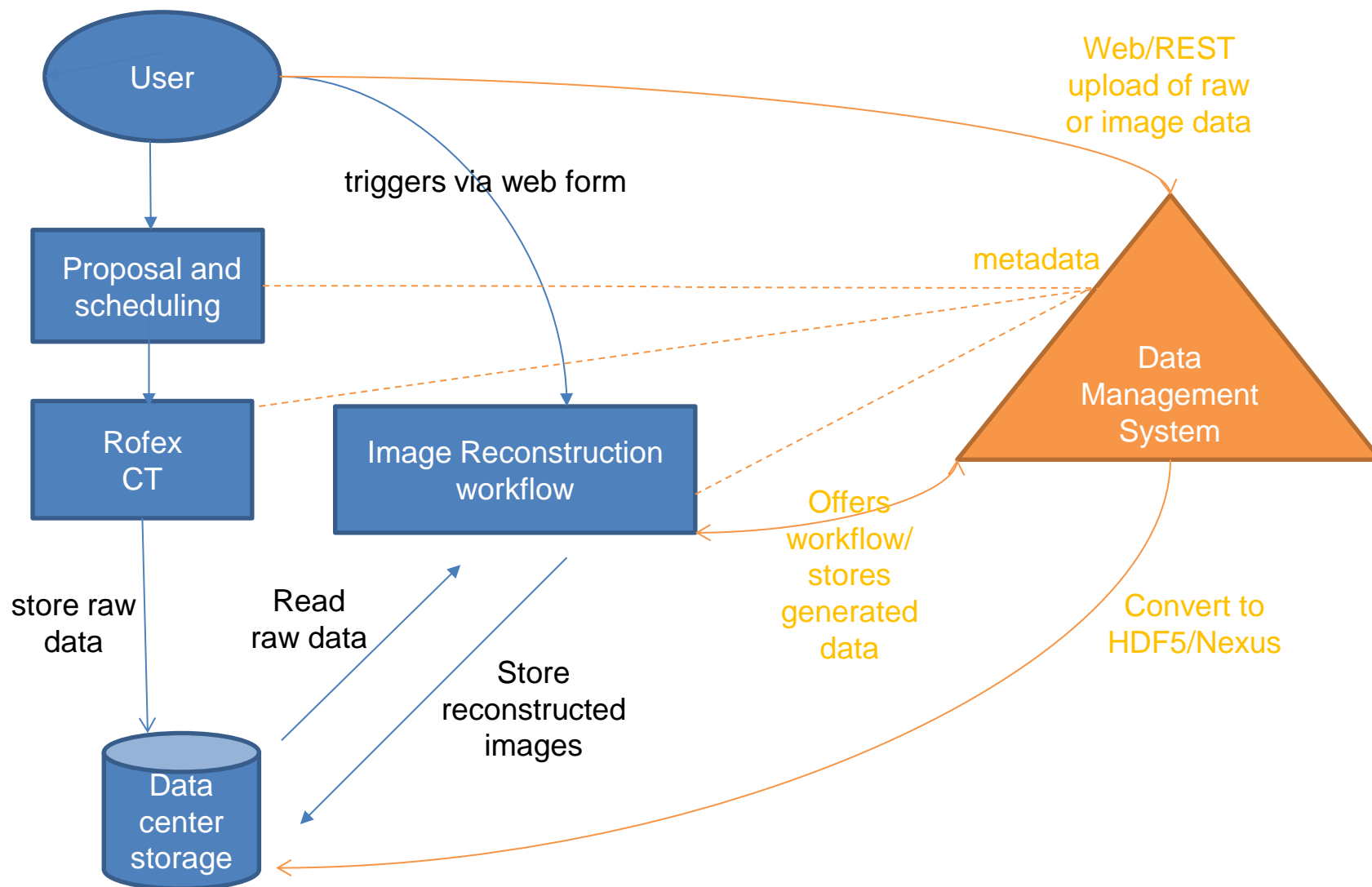
HZDR data publication system: RODARE

- 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

The screenshot displays the RODARE web interface. At the top, there is a blue navigation bar with the RODARE logo, a search bar, and links for 'Upload' and 'Communities'. A 'Log in' button is located in the top right corner. Below the navigation bar, a light blue banner contains the text: 'Not in production yet! DO NOT USE! All records will be removed.' The main content area is divided into two columns. The left column, titled 'Recent uploads', lists two datasets: 'Test: Dataset 1' (uploaded January 18, 2018) and 'Test Datenpublikation' (uploaded January 12, 2018). Each entry includes a 'View' button and a green 'Open Access' badge. The right column features a 'Welcome to Rodare!' message with a megaphone icon, stating that it is the new data publication platform at HZDR. Below this, a section titled 'RODARE' lists a bullet point: 'Open Access to HZDR research data. — publish your HZDR research output on the HZDR data publication platform Rodare.'

Looking at an Example: ROFEX CT

Data flow for ROFEX: Now and future

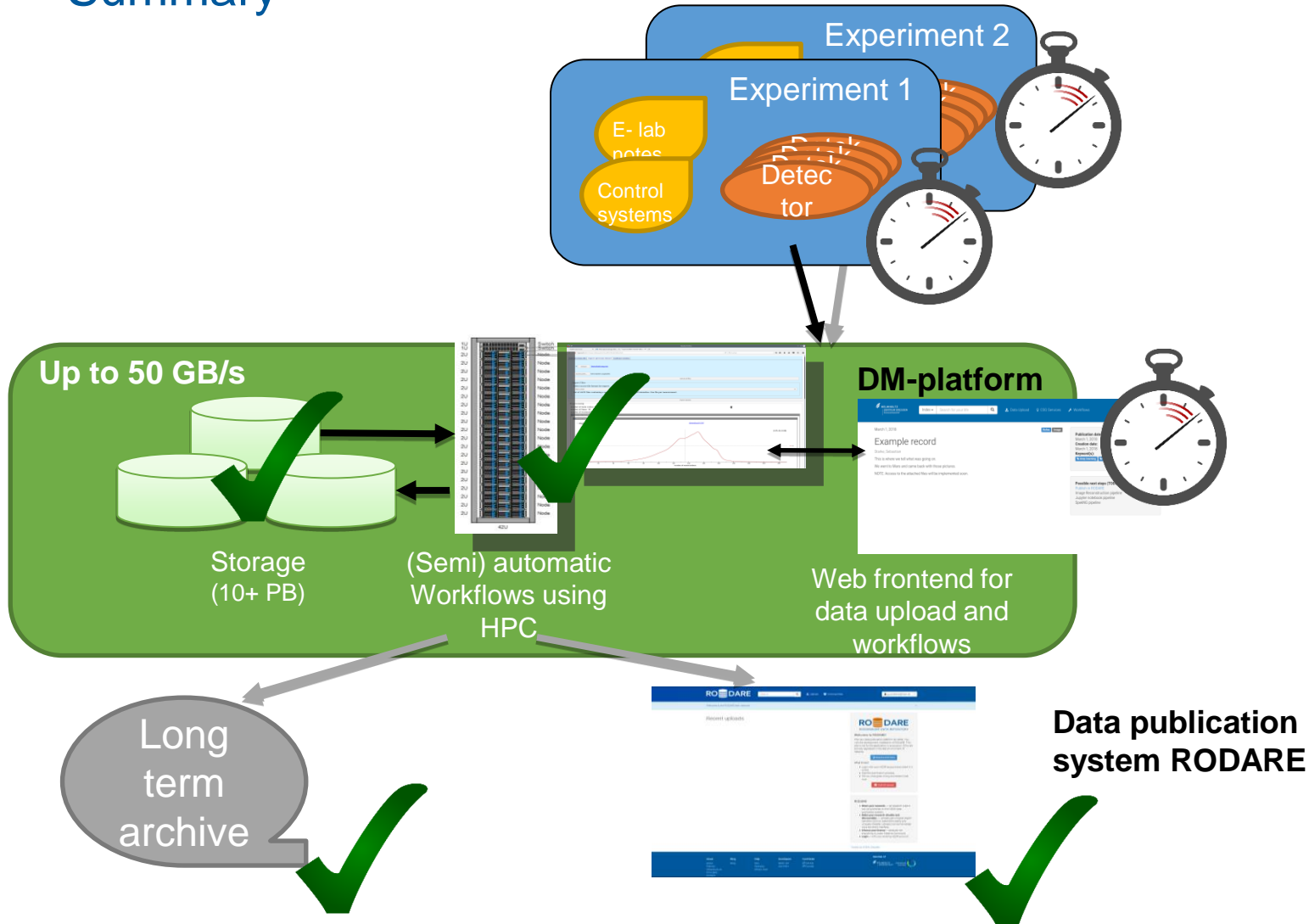


Data management and analytics platform: vision

- 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

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).