Open Data for DESY, HIFIS, NFDI and EOSC

Bundling portals for DESY, HIFIS, NFDI and their pilot node in EOSC Beyond

Tim Wetzel, Patrick Fuhrmann, Uwe Jandt, Paul Millar, Sophie Servan, Franz Rhee, Peter van der Reest, Regina Hinzmann, Noel Barth, Johannes Reppin, Christian Voss, Linus Pithan, Anton Barty, …

IBERGRID 2024, Porto, 29th October 2024

HIFIS

DAPHNE 4NFDI

In cooperation with

nfdi

eosc | BEYOND

EOSC Beyond receives funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101131875.

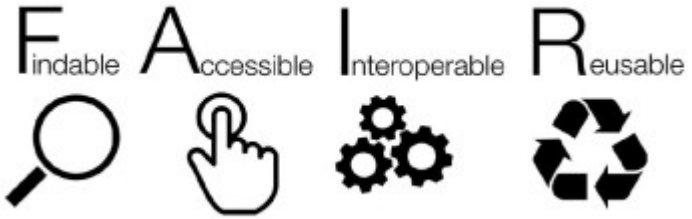
DESY.

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

IBERGRID · fct Fundação para a Ciência e a Tecnologia · Gobierno de España Ministerio de Ciencia, Innovación y Universidades · LIP Laboratório de Instrumentação e Física Experimental de Partículas · INCD Infraestrutura Nacional de Computação Distribuída · CSIC Consejo Superior de Investigaciones Científicas · European Commission · FM UP · HPC Now! Do IT Now Alliance

Open and FAIR data for Photon Science

The motivation for a prototype system



FAIR data is becoming the standard

- Open and/or FAIR data demanded by funding agencies and journals
  - Public money = public data (embargo periods may apply)
  - Supplemental data for publications
- Reproducibility is key
- More sustainable (re-)use of results obtained from laboriuos experiments and enables AI/ML training

Let there be light - starting with Photon Science

- As one of the largest photon science laboratories in Europe, DESY will start providing a standardized way to host Open and FAIR data for her scientists

Towards a blueprint for HIFIS, NFDI, EOSC and the community

- After successful initial operations with DESY photon science, the portal will be opened as a HIFIS service
- We also hope to create a blueprint for OpenData portals that will be shared openly

DESY Photon Science setup

A high-level view of the world

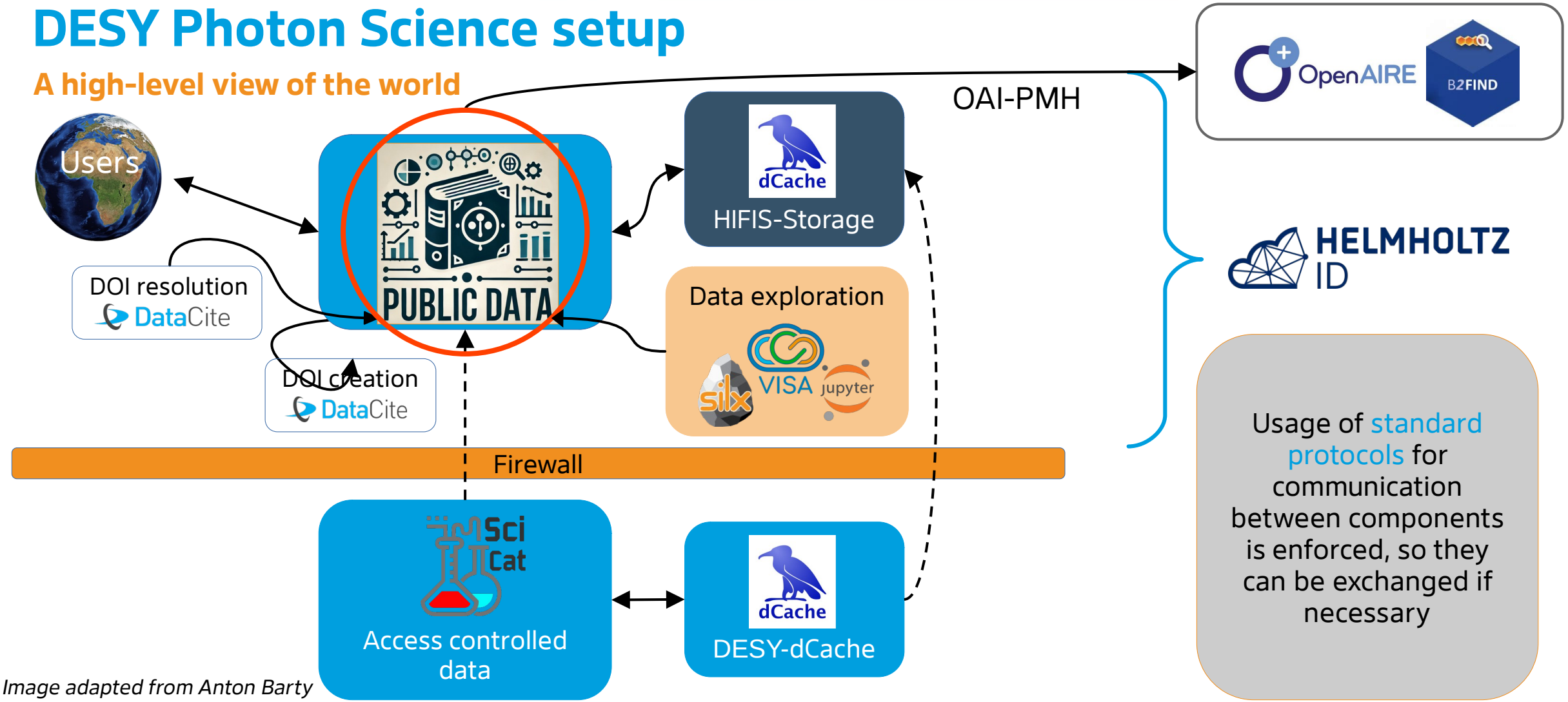


Image adapted from Anton Barty

The minimum viable system for DESY.

Essential components with federated access (authenticated & non-authenticated)

Long term storage (dCache via hifis-storage.desy.de**)**

- accessible via standard protocols (https, NFS, WebDAV)

Metadata Catalogue with

- mandatory core metadata fields
- optional domain specific metadata fields
- OAI-PMH protocol for data harvesting of core metadata by high level catalogues

1st phase

DOI Minting Service

- In cooperation with our library, technical prototype in working state

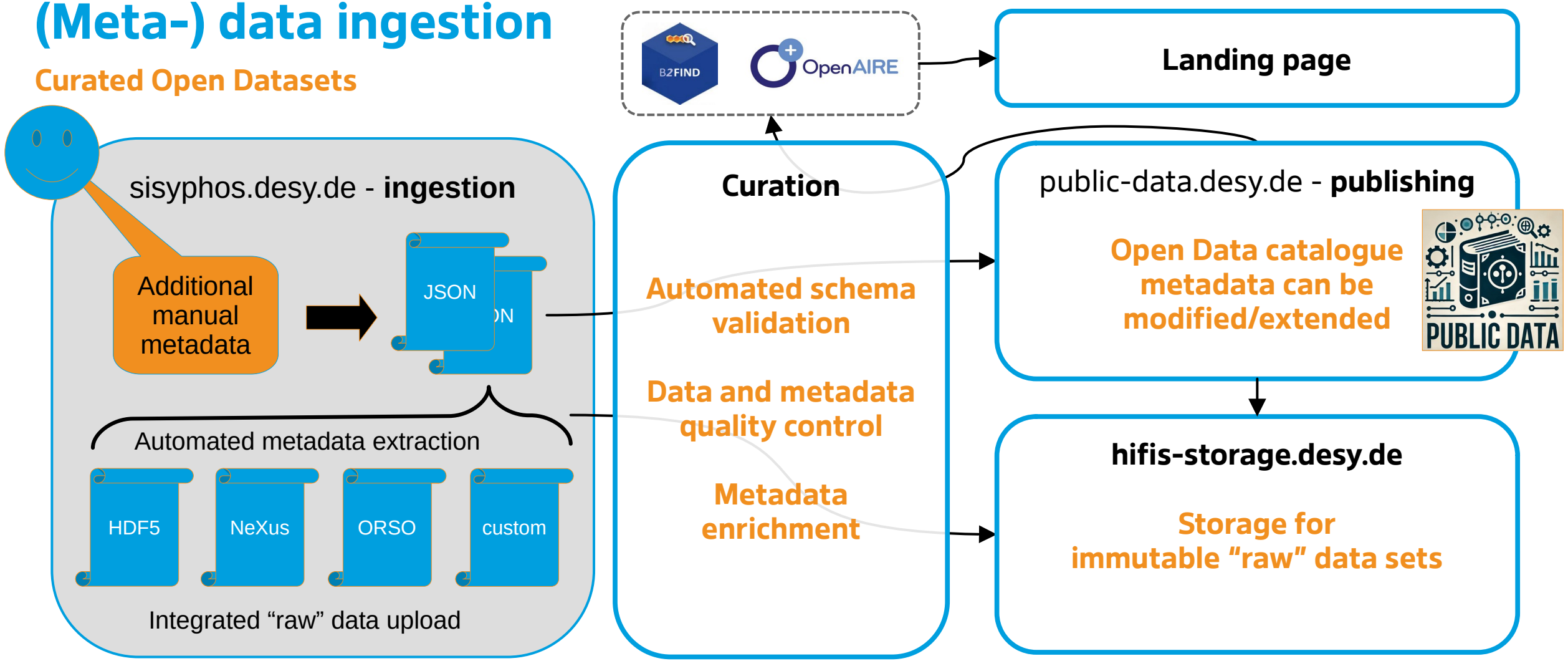
Open Science (Virtual Research) infrastructure

- VISA portal, currently working on it together with other synchrotron facilities in Europe under an MoU

2nd phase

(Meta-) data ingestion

Curated Open Datasets



Importance of proper metadata definitions

Consensus and standards are key

Mandatory **core metadata** fields	Defined in prior activities and by responsible reference bodies e.g. DublinCore, DataCite v4.4
Optional **domain-specific metadata** fields	To be provided by the community e.g. former PaNOSC/ExPaNDS, Daphne4NFDI, Photon Science Community
Additional metadata fields	Experiment/Beamline/Facility-specific metadata might be needed

Special challenge for open data:

Heterogeneous origin of data sets from different experiments with different specific metadata need to be mapped into the same catalogue

→ Metadata **input and verification** need to be handled properly in the publication process

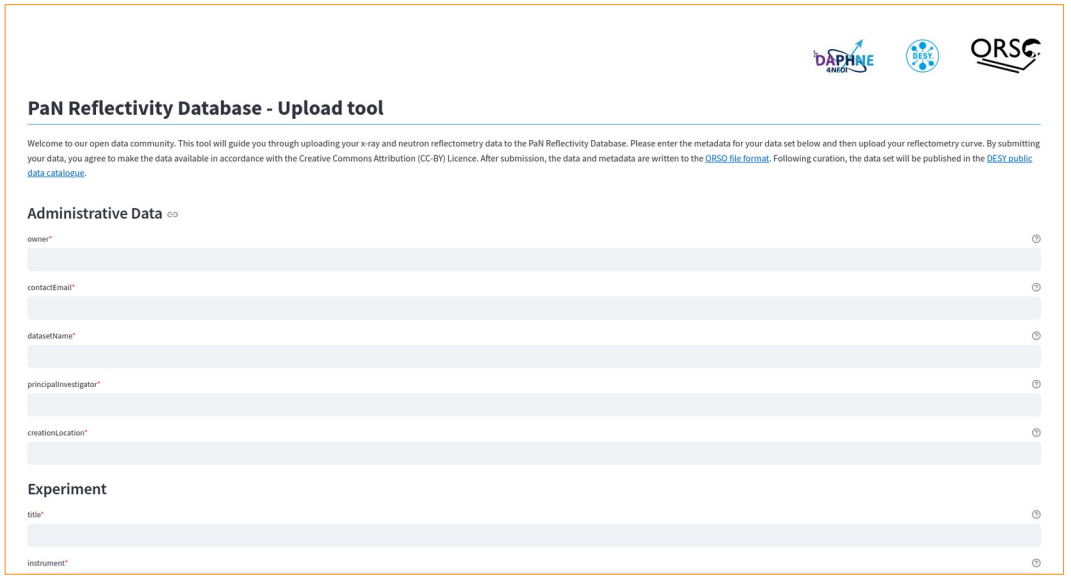
(Meta-) data ingestion

(Meta)data ingestion and quality verification – sisyphos.desy.de

- Prototype
  - Built with Streamlit as application framework and LinkML to specify metadata schemata
  - Schema definition through YAML documents
  - Schemata built from "classes" and "slots" → allows for inheritance and mixins to create custom modular schemata (base enhanced by community/experiment specifics)
  - Tooling for introspection, validation, format conversion, crosswalks …

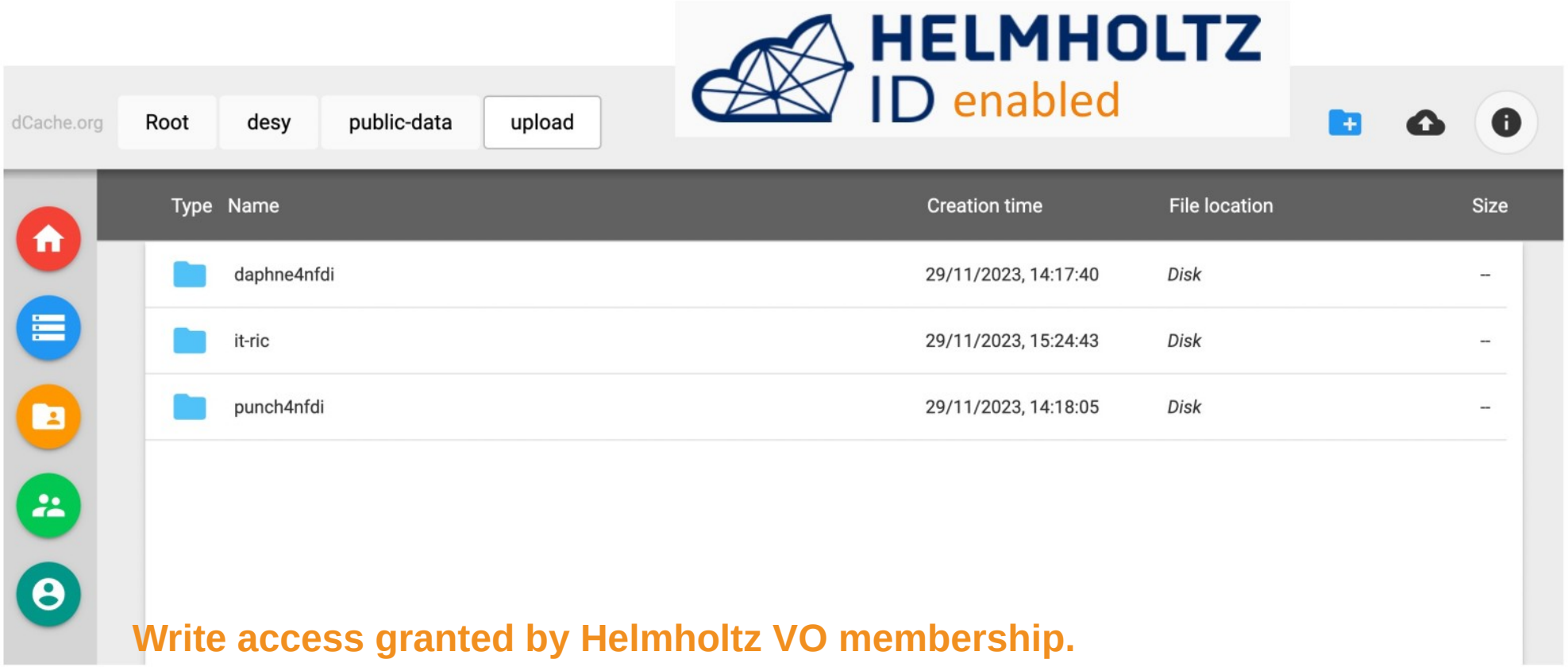
- Starting for the X-Ray reflectivity community within DAPHNE4NFDI

- If you are interested in details:
  - https://gitlab.desy.de/ric/opendata-metadata/
  - sisyphos.desy.de
  - Let me know so I can get you into contact with my colleagues



hifis-storage.desy.de

The “drop box” and and final storage space for Open Data



Write access granted by Helmholtz VO membership.

public-data.desy.de

The metadata catalog!

PUBLIC DATA

Search | Clear

PID

Text Search

Location

Group

Type

Keywords

Start Date – End Date

+ Add Condition

Items per page: 25 | 1 – 2 of 2

Name	Source Folder	Start Time	Type
Reflectometry curves (XRR and NR) and corresponding fits for machine learning	...do.6497438	2024-01-25 Thu 18:34	raw
spain	.../nfs	2023-12-18 Mon 06:27	derived

General Information

Name	Reflectometry curves (XRR and NR) and corresponding fits for machine learning
Description	This is a compiled dataset of raw X-ray reflectivity (XRR, reflectometry) measurements together with corresponding fit parameters, intentionally published to use as training or test data for machine learning models. (The authors aim to include NR data in further versions of this dataset and plan to include other substrates and materials for XRR. Contributions welcome!)
PID	undefined/10242df2-3868-42cb-bcb2-81c2c44533ec
Type	raw
Creation Time	2024-01-25 18:34
Keywords	

Creator Information

Owner	Linus Pithan
Principal Investigator	linus.pithan@desy.de
Contact Email	linus.pithan@desy.de
Owner Group	fsec
Access Groups	

File Information

Source Folder	/desy/public-data/upload/daphne4nfdi/10.5281_zenodo.6497438

Scientific Metadata

Search

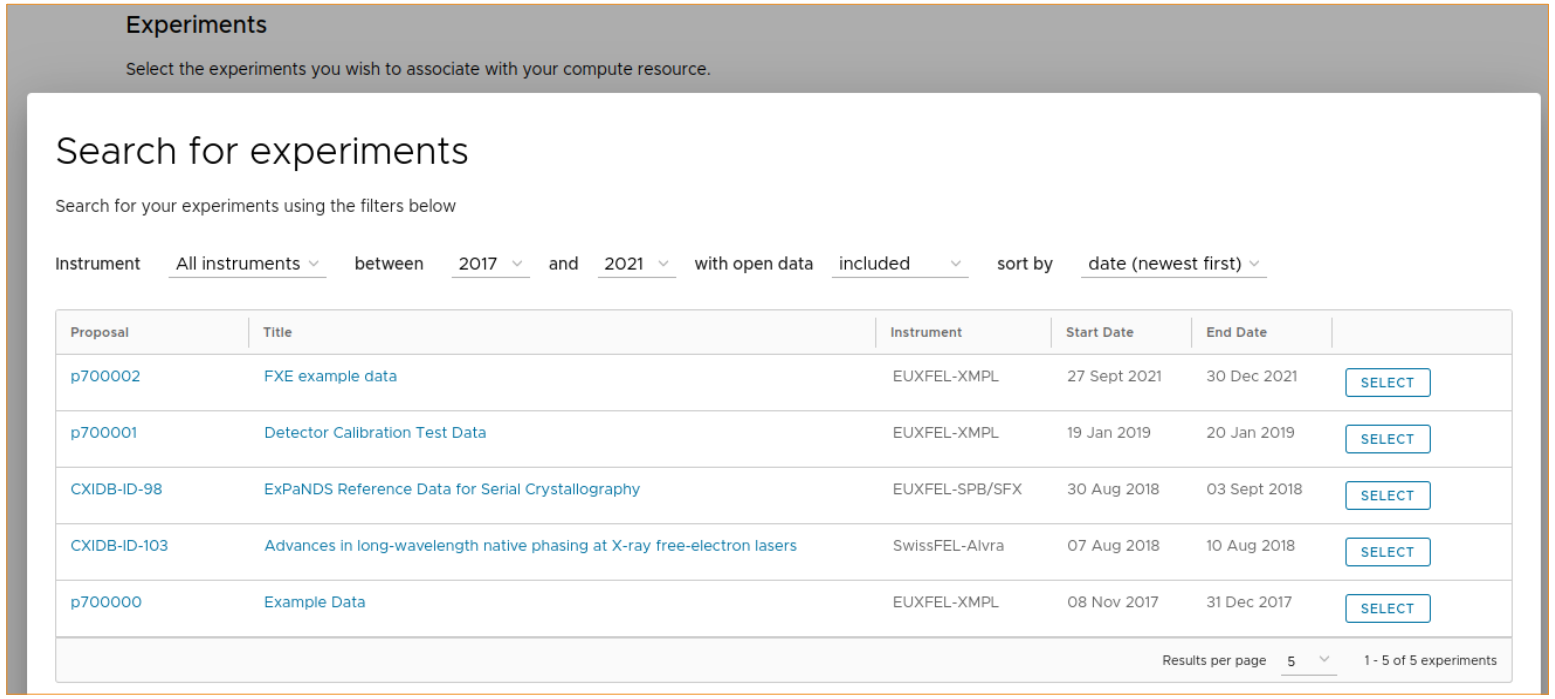
DIP_1

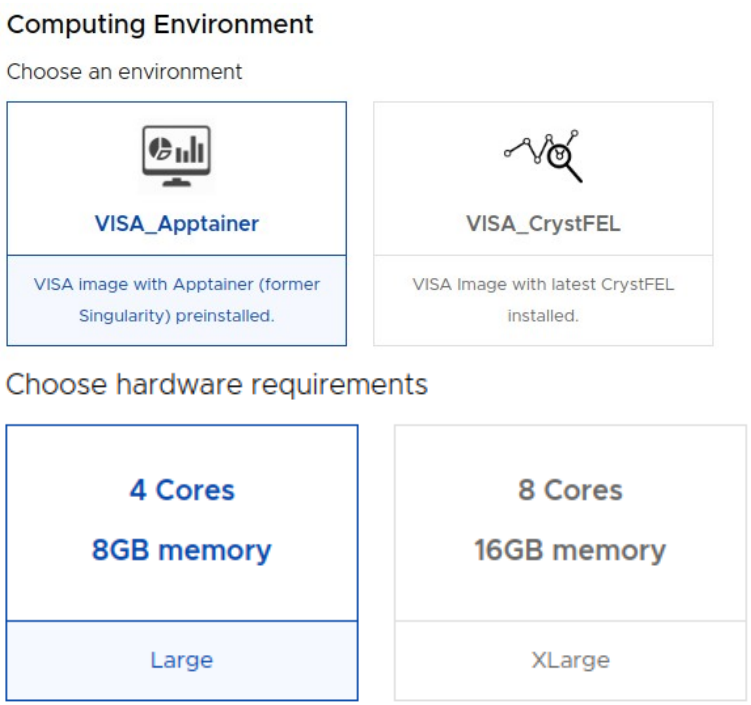
Experimentalists	Kowark, Stefan
Layer_CAS	188-94-3
Layer_formula	C32H16
Layer_material	Diindenoperylene
Substrate_temperature	303 (K)
instrument	ESRF, ID10b
	0.15 (1/Ang)
	2005

Path	Size
calc_xrr.py	2 KB
conda_env.yml	7 KB
prepare_plot.py	4 KB
README.html	6 MB
README.ipynb	9 MB
requirements.txt	76 B
xrr_dataset.h5	254 KB

visa.desy.de

Select a dataset to spawn a virtual machine



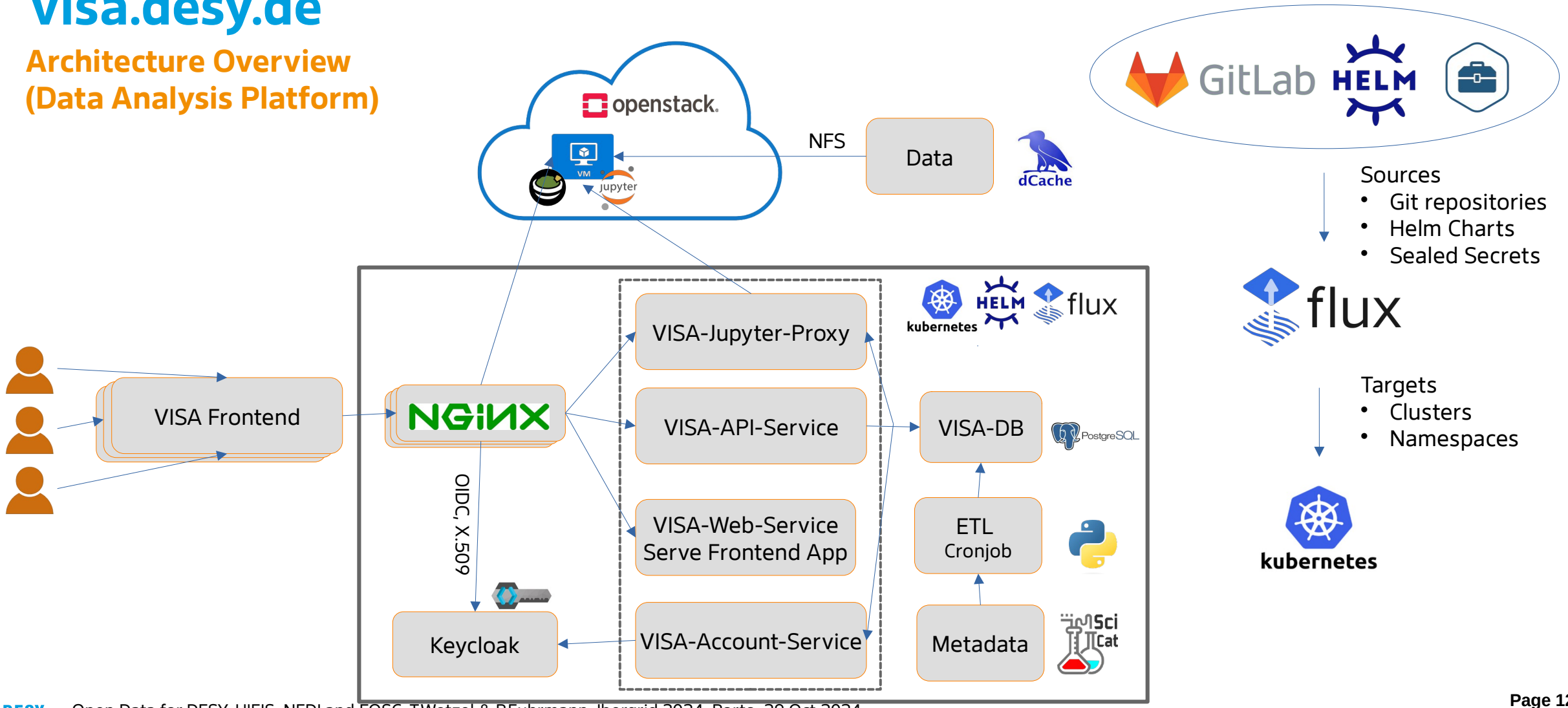


VISA database currently populated with example datasets.

Open Data to be integrated during 2024 via automated data export from public-data.desy.de

visa.desy.de

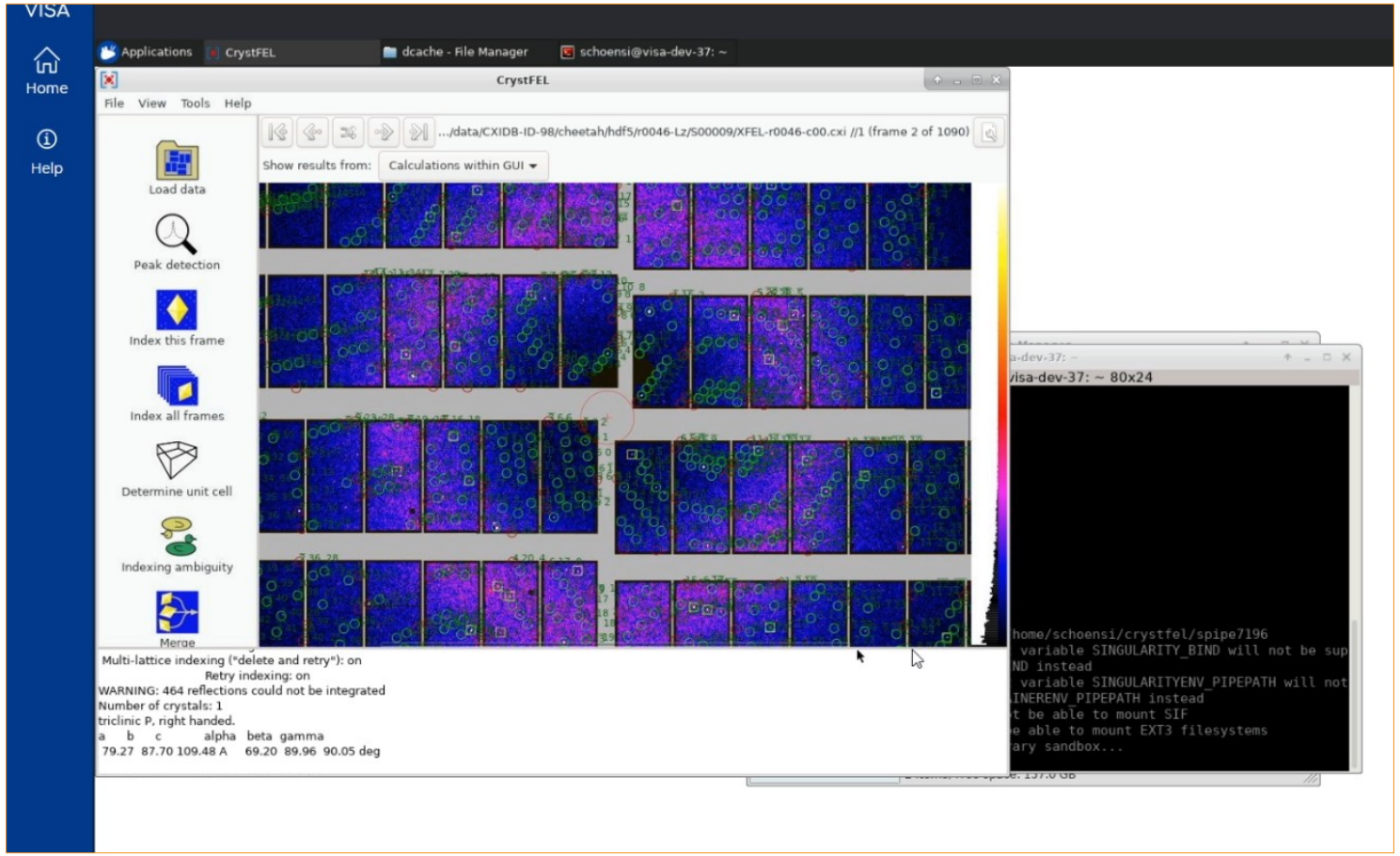
Architecture Overview (Data Analysis Platform)



visa.desy.de

Work via remote desktop connection with graphical interfaces

Using CrystFEL Docker Images to run Singularity Container and work with Crystfel 10 Graphical Interface.



Example provided by Silvan Schön (DESY/FS-SC)

Thank you!

Questions?

Contact

DESY. Deutsches Elektronen-Synchrotron

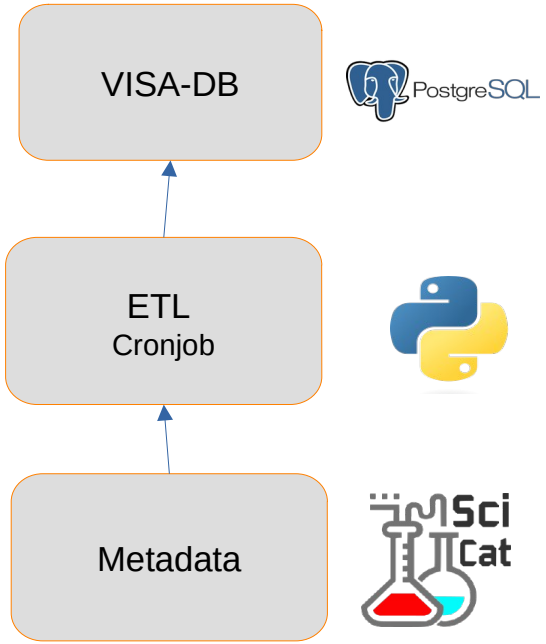
www.desy.de

Tim Wetzel, Patrick Fuhrmann
IT-RIC (Research & Innovation in Scientific Computing)
tim.wetzel@desy.de, patrick.fuhrmann@desy.de

Backup slides

visa.desy.de

Metadata import via custom ETL process



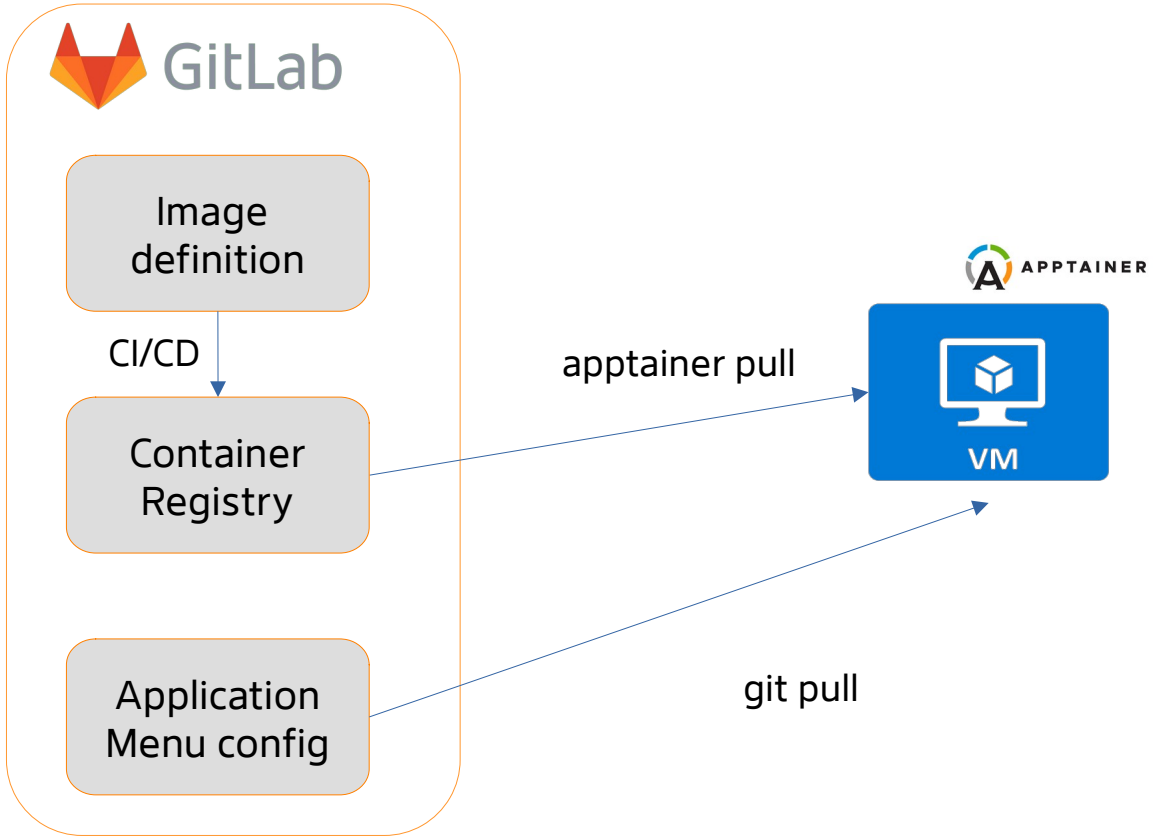
- Python ETL script
- Customizable depending on the metadata source (catalogue API format, authN/Z, ...)
- Can be run once for static data or as a cronjob for dynamic data
- Event-based execution would be nice to have (e.g. webhooks)

- Metadata import
  - Experimental specifications
  - Dataset status (embargoed or public)
  - User access rights
  - Storage paths

- Database backup

visa.desy.de

Analysis software provisioning via Apptainer images



- Software in Apptainer images
  - Many applications already available as Apptainer image from HPC workflows
- Built from .def file in CI/CD pipeline
- Image publicly available in Gitlab registry
- Pulled on application startup

- Application menu entries defined separately in git repository
- Seamless integration into the OS applications
- Menu entries updated from menu config by cronjob pulls the repository regularly
- Seamless updates to the menu by admins