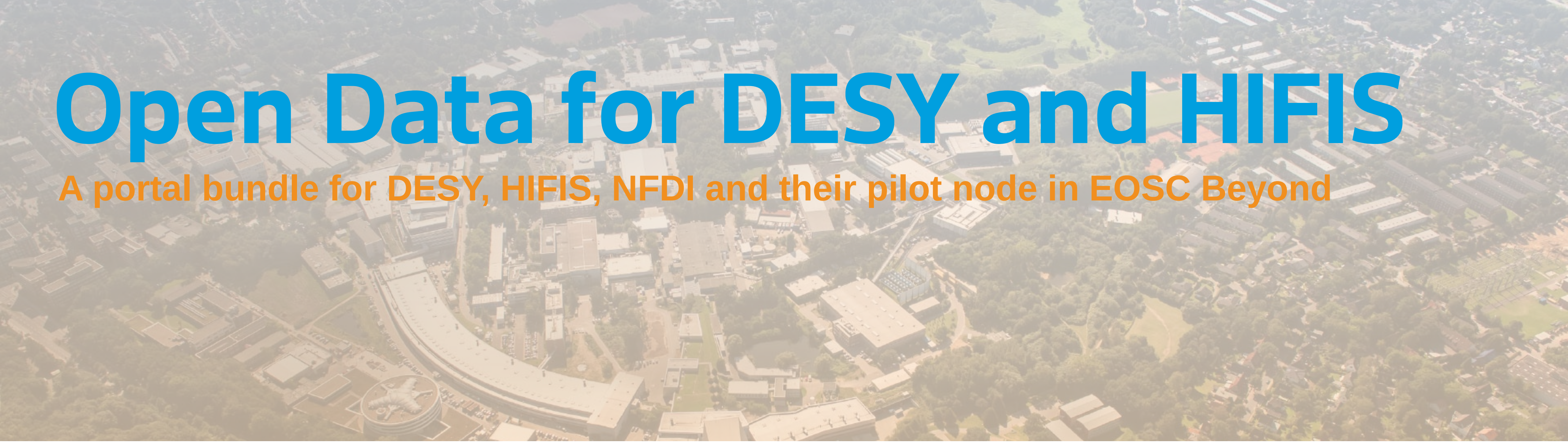


Open Data for DESY and HIFIS

A portal bundle 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, …

EGI2024, Lecce, 03rd October 2024

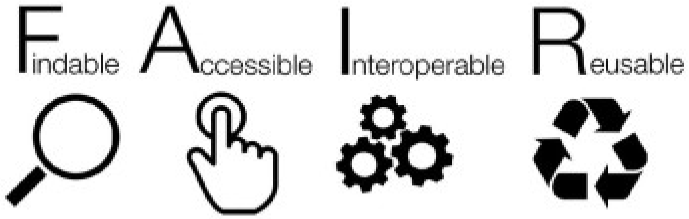
In cooperation with

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

Open and FAIR data for Photon Science

The motivation for a prototype system

FAIR data will become the standard



- Funding bodies and journals demand data to be open and/or FAIR
  - Public money = public data (after embargo period)
  - Supplemental data for publications
- Combatting the reproducibility crisis in science
- Reusability makes for a more sustainable (re-)use of results obtained from costly and laboriuos experiments and enables AI/ML workflows

Starting with Photon Science

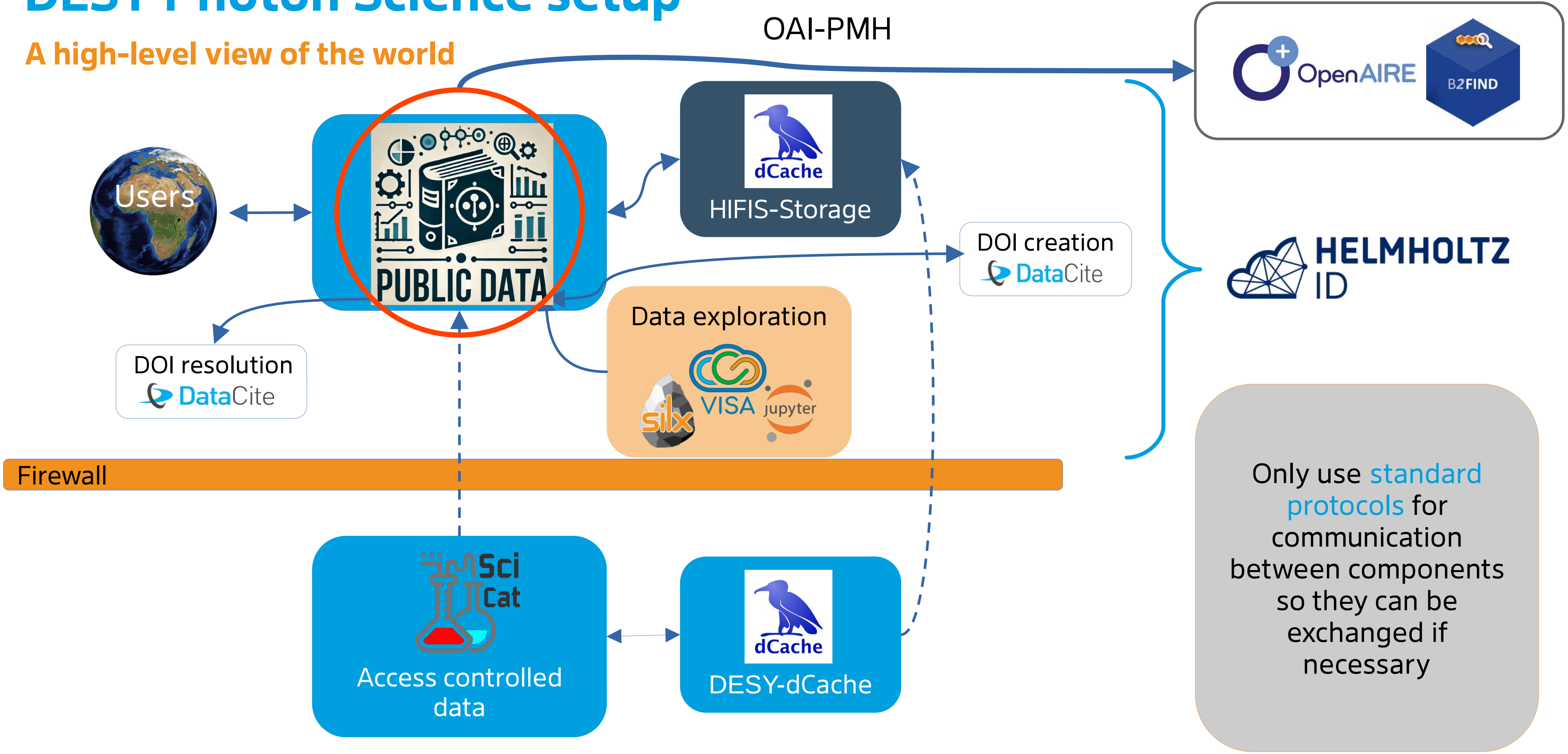
- As one of the largest photon science laboratories in Europe, DESY will start providing a standardized way to host Open and FAIR data for her scientists

Towards a blueprint for HIFIS, NFDI 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 in the community that will be shared openly

DESY Photon Science setup

A high-level view of the world



Only use standard protocols for communication between components so they can be exchanged if necessary

Image adapted from Anton Barty's slide

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

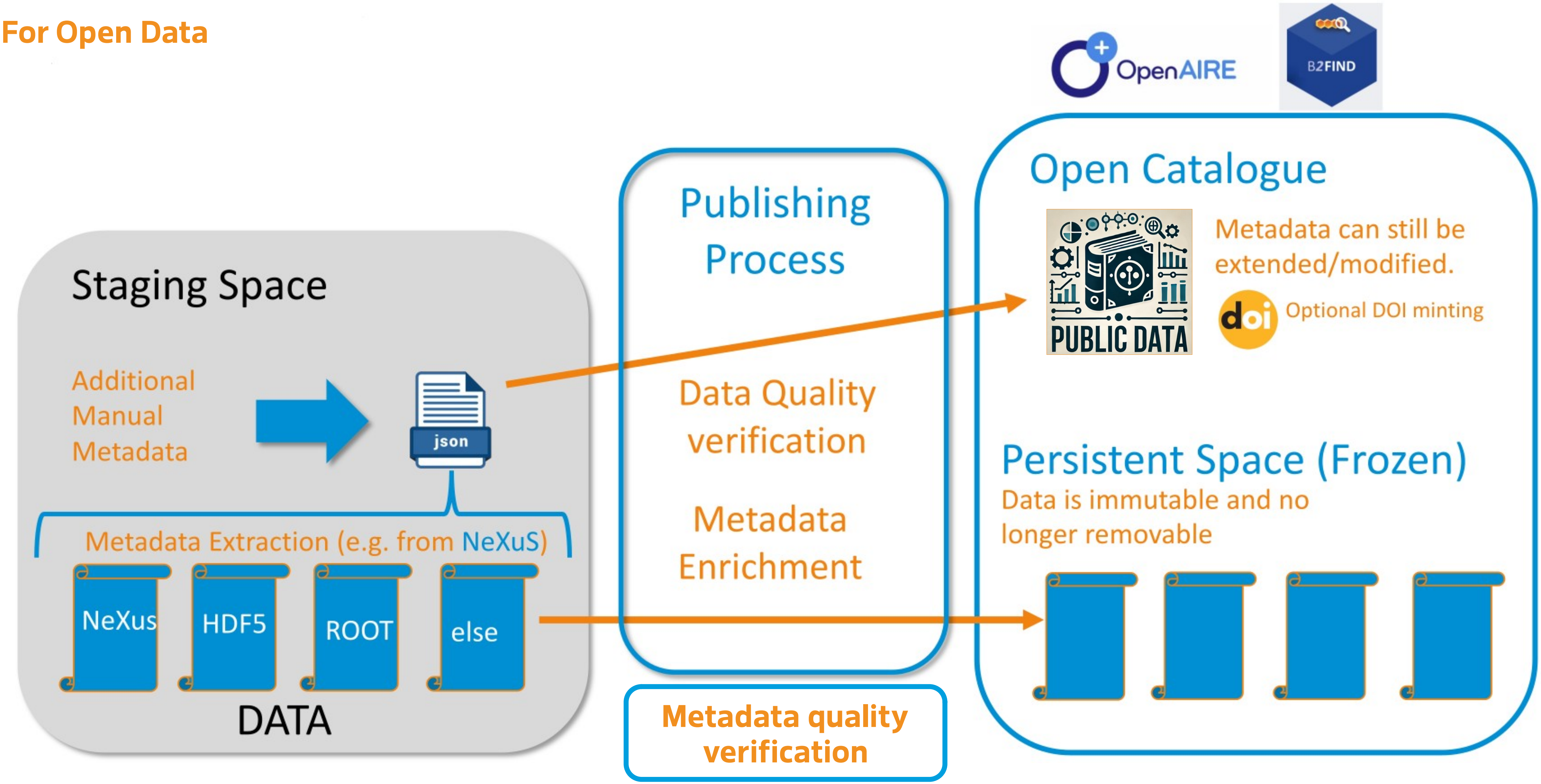
Open Science (Virtual Research) infrastructure

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

2nd phase

Our envisioned data ingestion process

For Open Data



Staging Space

Additional Manual Metadata

Metadata Extraction (e.g. from NeXuS)

DATA

Publishing Process

Data Quality verification

Metadata Enrichment

Metadata quality verification

Open Catalogue

Metadata can still be extended/modified.

Optional DOI minting

Persistent Space (Frozen)

Data is immutable and no longer removable

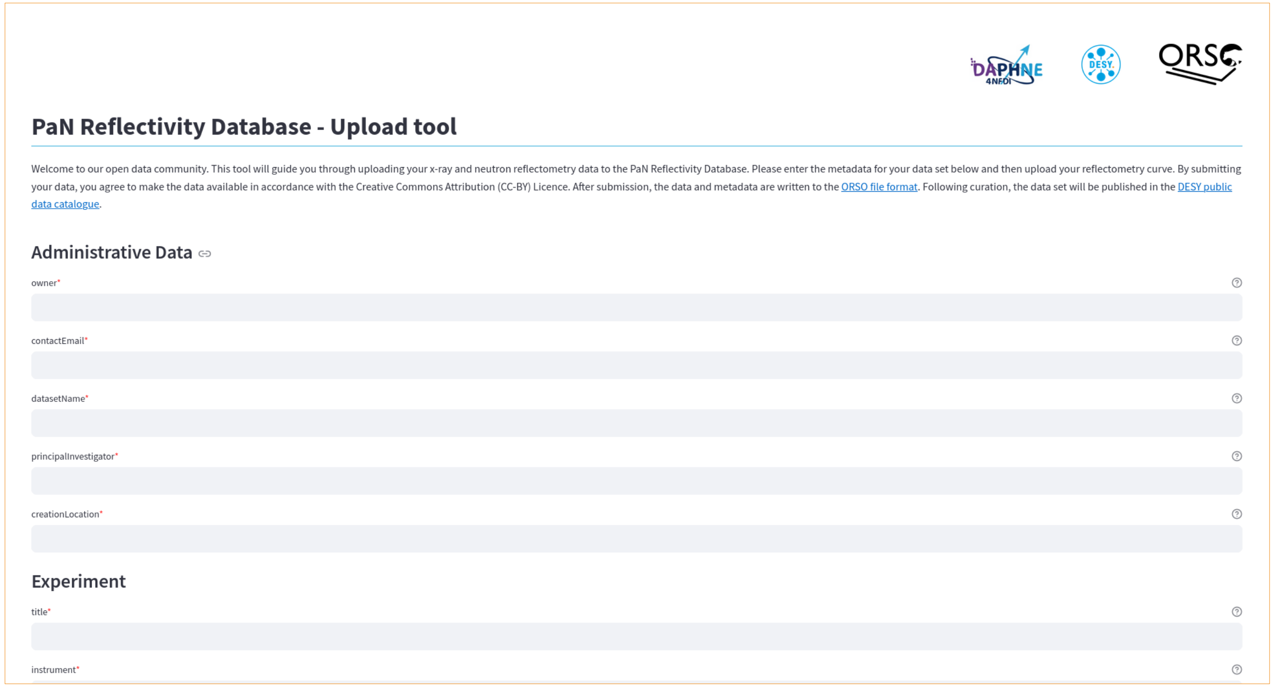
Our envisioned data ingestion process

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

- First installment with LinkML and Streamlit
  -
  - Metadata schema description via YAML documents setting standards that metadata has to conform to
  - Data description in terms of "classes" and "slots", allowing inheritance and mixins for creating custom types
  - 60+ different open-source tools to work with schemata for introspection, validation, format conversion, …

- 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



PaN Reflectivity Database - Upload tool

Welcome to our open data community. This tool will guide you through uploading your x-ray and neutron reflectometry data to the PaN Reflectivity Database. Please enter the metadata for your data set below and then upload your reflectometry curve. By submitting your data, you agree to make the data available in accordance with the Creative Commons Attribution (CC-BY) Licence. After submission, the data and metadata are written to the ORSO file format. Following curation, the data set will be published in the DESY public data catalogue.

Administrative Data

owner*

contactEmail*

datasetName*

principalInvestigator*

creationLocation*

Experiment

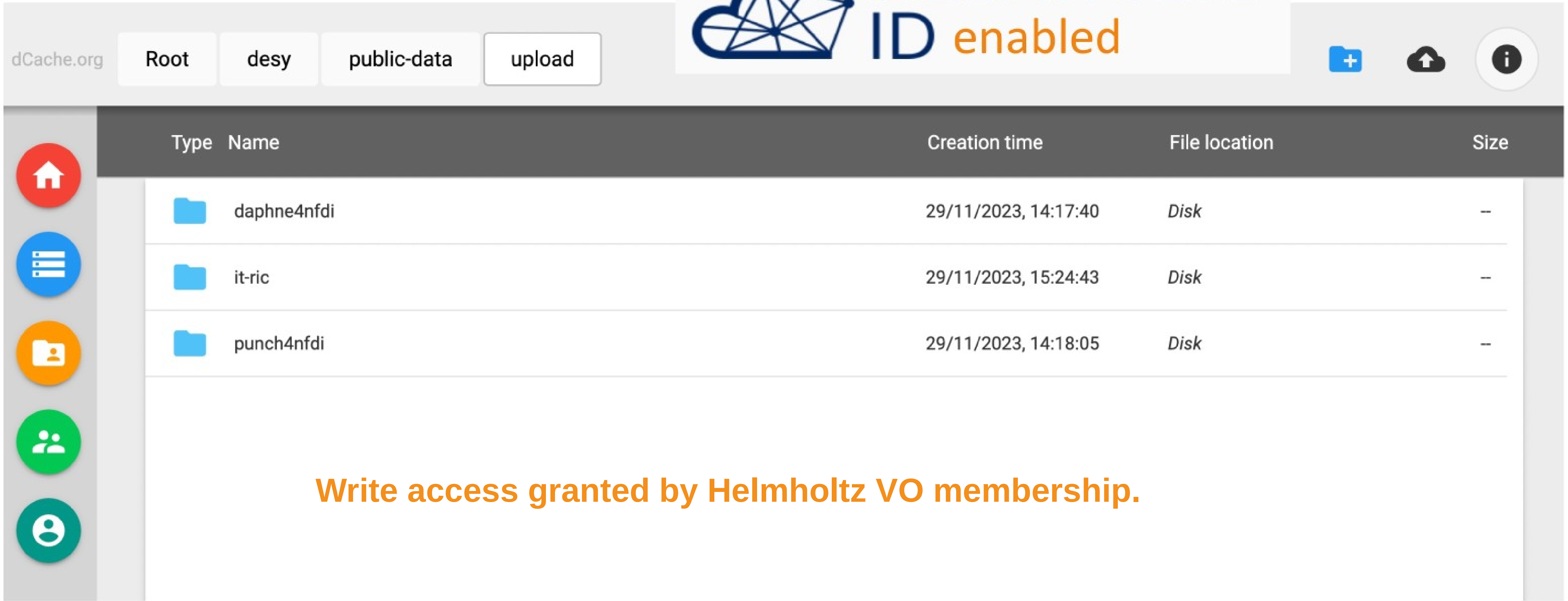
title*

instrument*

hifis-storage.desy.de

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

HELMHOLTZ ID enabled



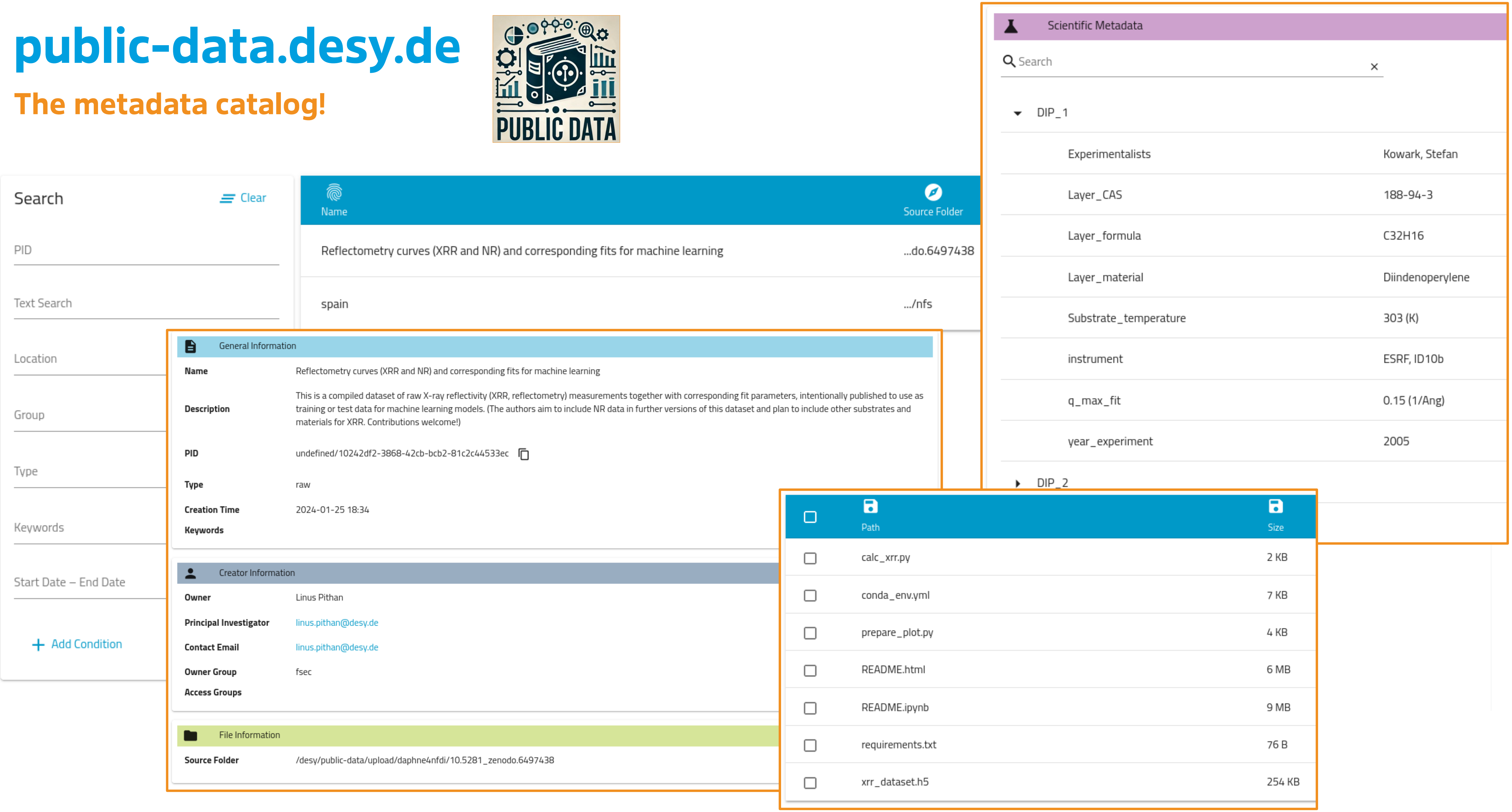
dCache.org Root desy public-data upload

Type	Name	Creation time	File location	Size
📁	daphne4nfdi	29/11/2023, 14:17:40	*Disk*	--
📁	it-ric	29/11/2023, 15:24:43	*Disk*	--
📁	punch4nfdi	29/11/2023, 14:18:05	*Disk*	--

Write access granted by Helmholtz VO membership.

public-data.desy.de

The metadata catalog!



Search
Clear

PID

Text Search

Location

Group

Type

Keywords

Start Date – End Date

+ Add Condition

Name	Source Folder
Reflectometry curves (XRR and NR) and corresponding fits for machine learning	...do.6497438
spain	.../nfs

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

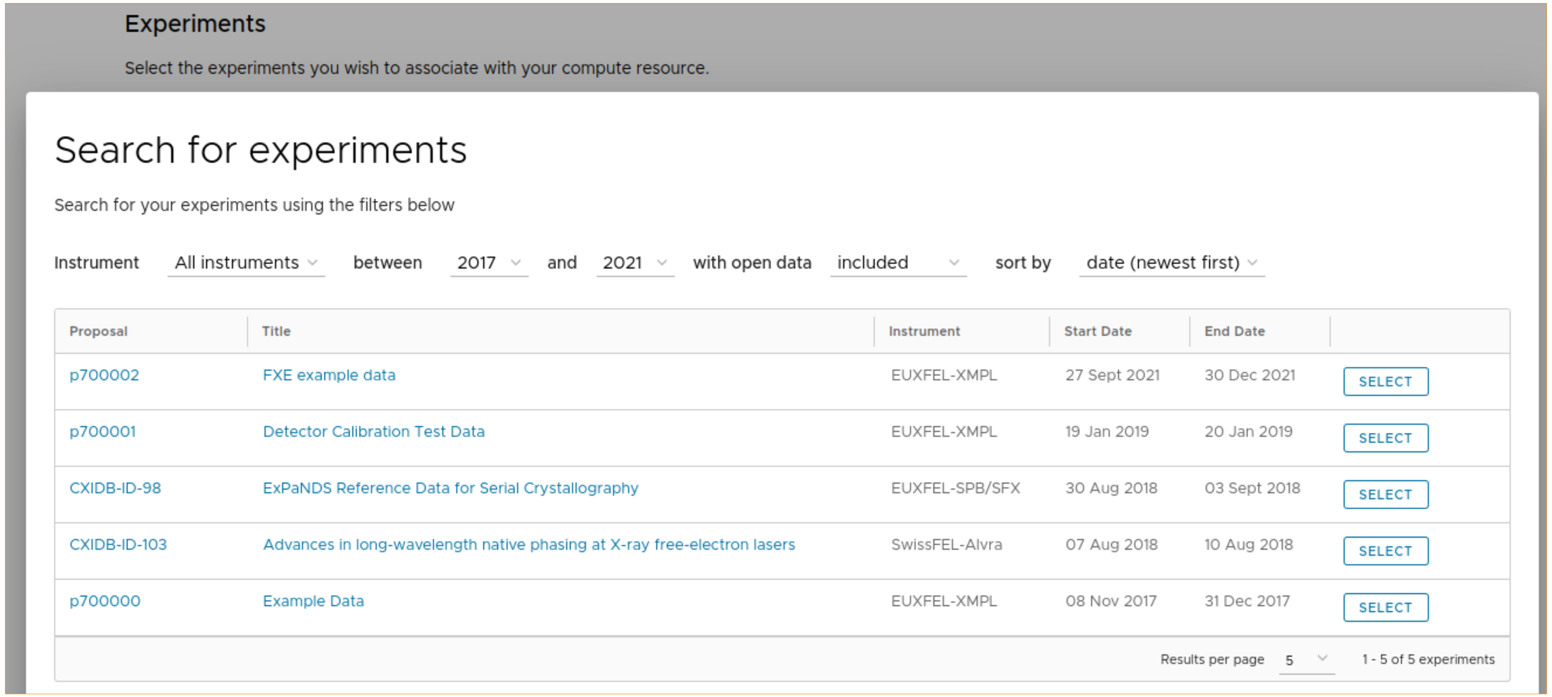
Experimentalists	Kowark, Stefan
Layer_CAS	188-94-3
Layer_formula	C32H16
Layer_material	Diindenoperylene
Substrate_temperature	303 (K)
instrument	ESRF, ID10b
q_max_fit	0.15 (1/Ang)
year_experiment	2005

▸ DIP_2

□	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



Experiments

Select the experiments you wish to associate with your compute resource.

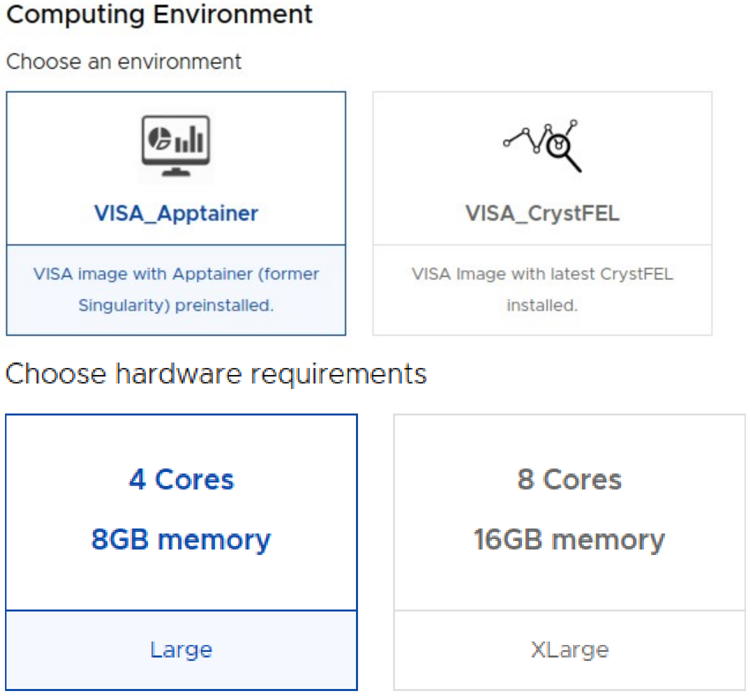
Search for experiments

Search for your experiments using the filters below

Instrument: All instruments; between 2017 and 2021; with open data: included; sort by: date (newest first)

Proposal	Title	Instrument	Start Date	End Date	
p700002	FXE example data	EUXFEL-XMPL	27 Sept 2021	30 Dec 2021	SELECT
p700001	Detector Calibration Test Data	EUXFEL-XMPL	19 Jan 2019	20 Jan 2019	SELECT
CXIDB-ID-98	ExPaNDS Reference Data for Serial Crystallography	EUXFEL-SPB/SFX	30 Aug 2018	03 Sept 2018	SELECT
CXIDB-ID-103	Advances in long-wavelength native phasing at X-ray free-electron lasers	SwissFEL-Alvra	07 Aug 2018	10 Aug 2018	SELECT
p700000	Example Data	EUXFEL-XMPL	08 Nov 2017	31 Dec 2017	SELECT

Results per page: 5; 1 - 5 of 5 experiments



Computing Environment

Choose an environment

- VISA_Apptainer: VISA image with Apptainer (former Singularity) preinstalled.
- VISA_CrystFEL: VISA Image with latest CrystFEL installed.

Choose hardware requirements

- 4 Cores, 8GB memory: Large
- 8 Cores, 16GB memory: XLarge

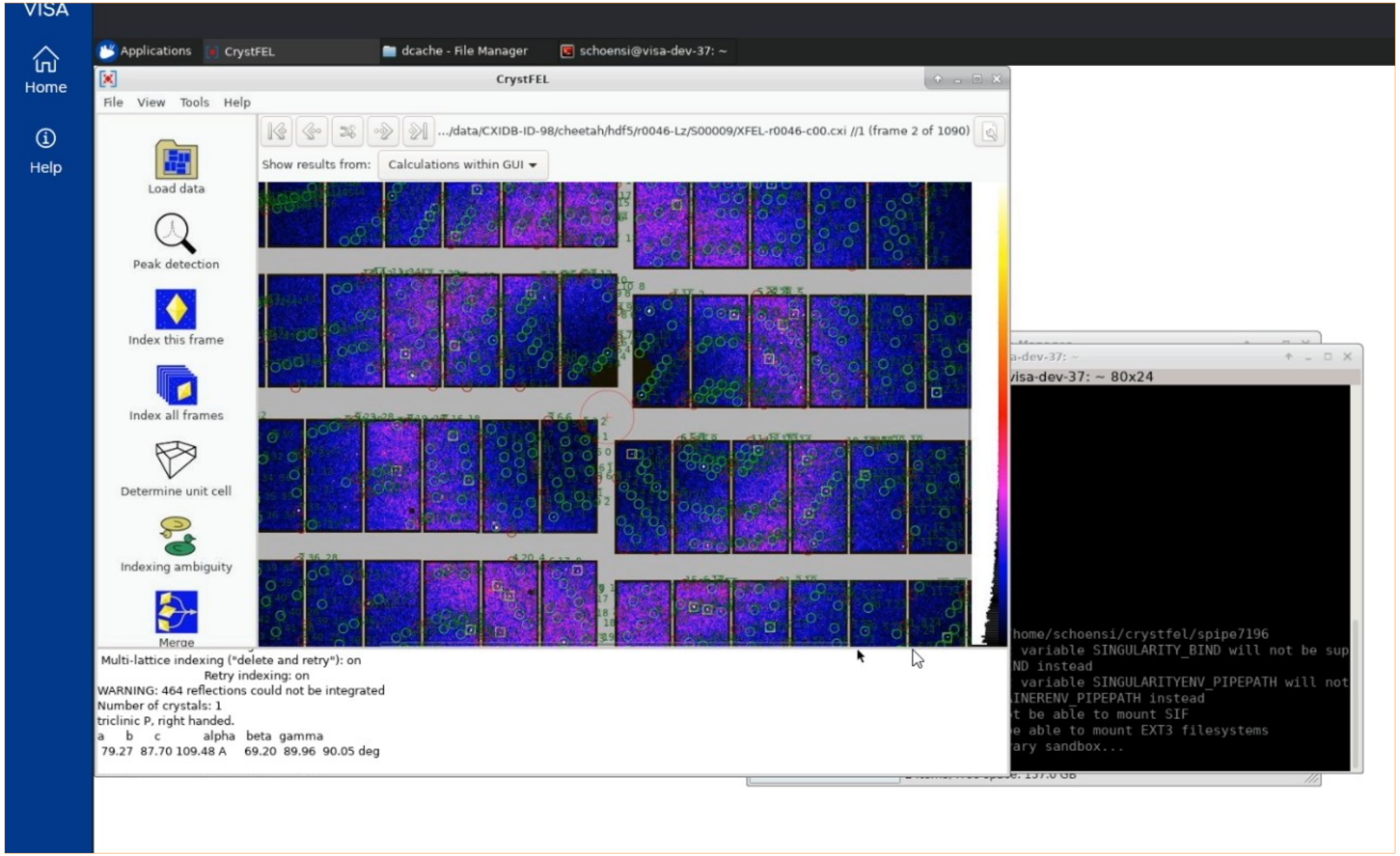
VISA database currently populated with example datasets.

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

visa.desy.de

Work via remote desktop connection with graphical interfaces

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



.../data/CXIDB-ID-98/cheetah/hdf5/r0046-Lz/S00009/XFEL-r0046-c00.cxi //1 (frame 2 of 1090)

Show results from: Calculations within GUI

```
Multi-lattice indexing ("delete and retry"): on
                     Retry indexing: on
WARNING: 464 reflections could not be integrated
Number of crystals: 1
triclinic P, right handed.
a      b      c            alpha   beta  gamma
 79.27  87.70 109.48 A     69.20  89.96  90.05 deg
```

```
home/schoensi/crystfel/spipe7196
 variable SINGULARITY_BIND will not be sup
ND instead
 variable SINGULARITYENV_PIPEPATH will not
INERENV_PIPEPATH instead
t be able to mount SIF
e able to mount EXT3 filesystems
ary sandbox...
```

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

CrystFEL

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

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

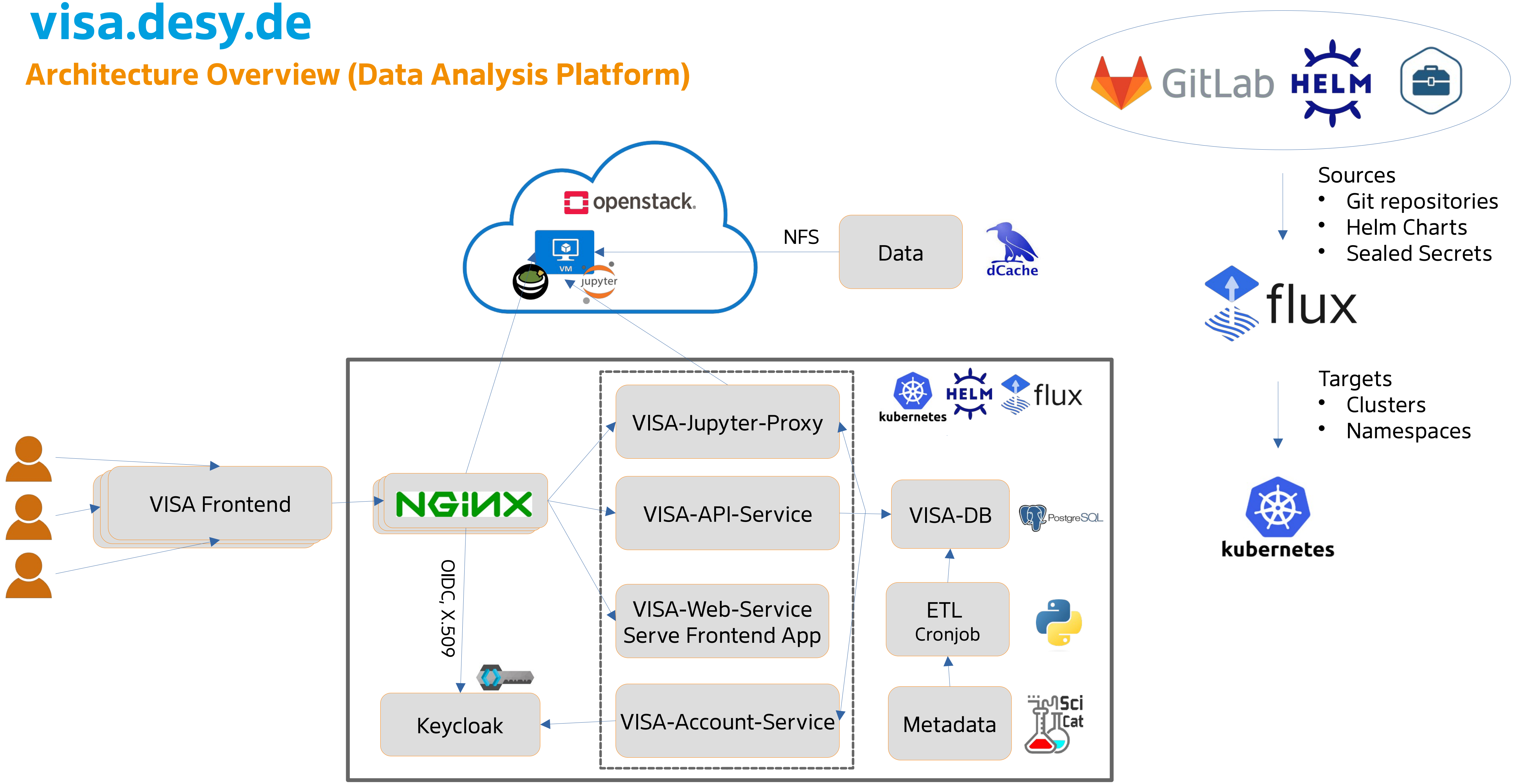
Special challenge for open data:

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

visa.desy.de

Architecture Overview (Data Analysis Platform)

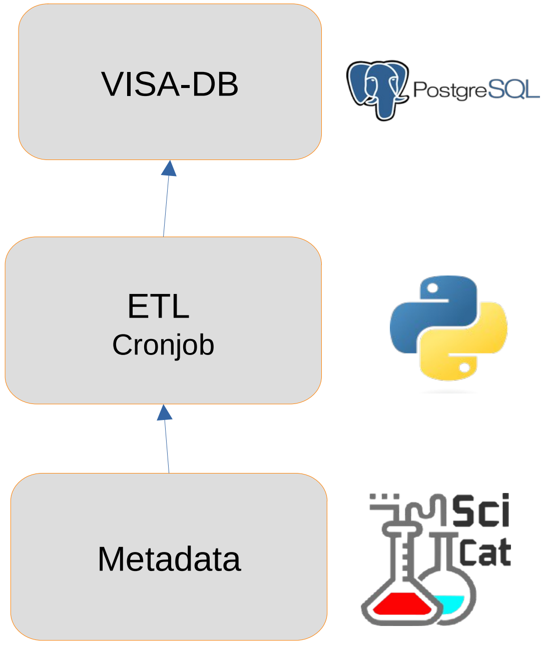


Sources
- Git repositories
- Helm Charts
- Sealed Secrets

Targets
- Clusters
- Namespaces

visa.desy.de

Metadata import via custom ETL process



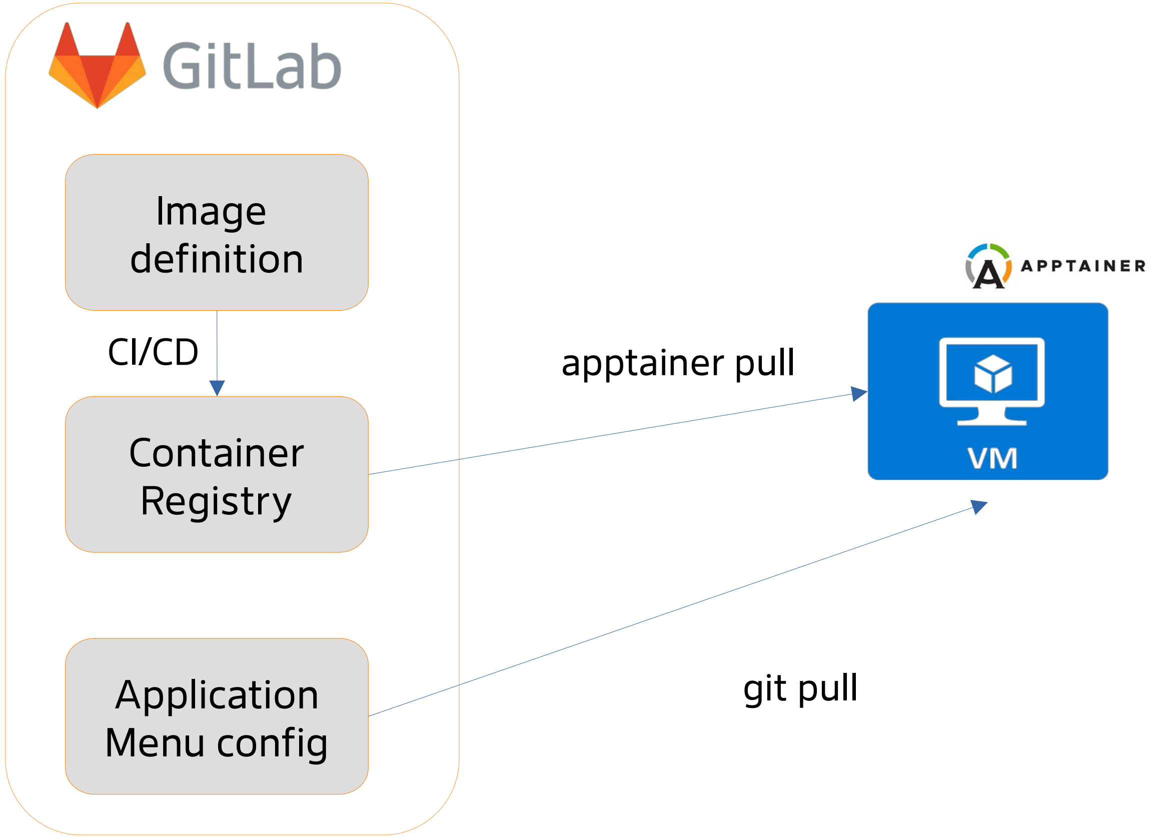
- Python script
- Customizable depending on the metadata source (catalogue API format, authN/Z, ...)
- Can be run once for static data or as a cronjob for dynamic data
- Event-based execution would be nice to have (e.g. webhooks)

- 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