HELMHOLTZ METADATA COLLABORATION

HZB Helmholtz Zentrum Berlin

FAIR data curation applied to HZB infrastructure

NeXus files, Icat server, python routines for data ingestion

www.helmholtz-metadata.de

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

Agenda

- Overview of data staging workflow at HZB
- NeXus as a community standard for metadata schema and file format
- Icat services
- Use case EMIL
- Development directions and discussion

Data management solutions at the HZB

Consultancy:

- Development in collaboration with researchers of ad hoc solutions for the data curation
- Curation on data access and publication services
- Awareness on Data staging workflows and data – metadata community standards

HZB

Rolf Krahl	rolf.krahl@helmholtz-berlin.de
Heike Görzig	heike.goerzig@helmholtz-berlin.de

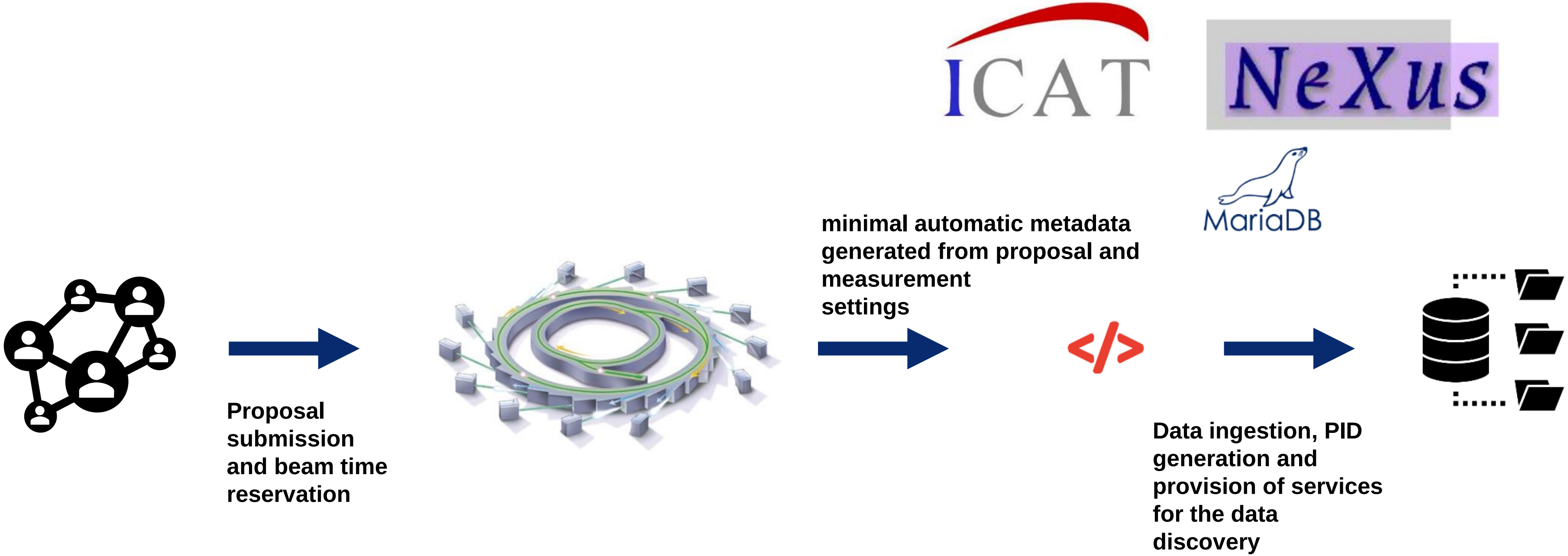
Hub Matter

Gerrit Günther	gerrit.guenther@helmholtz-berlin.de
Markus Kubin	markus.kubin@helmholtz-berlin.de
Oonagh Mannix	oonagh.mannix@helmholtz-berlin.de

FAIR guidelines ⟹ FAIR Tools ⟹ FAIR Experiments

[1]

Data management workflow at HZB facilities: Storage and processing

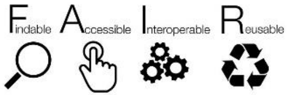


Proposal submission and beam time reservation

minimal automatic metadata generated from proposal and measurement settings

Data ingestion, PID generation and provision of services for the data discovery

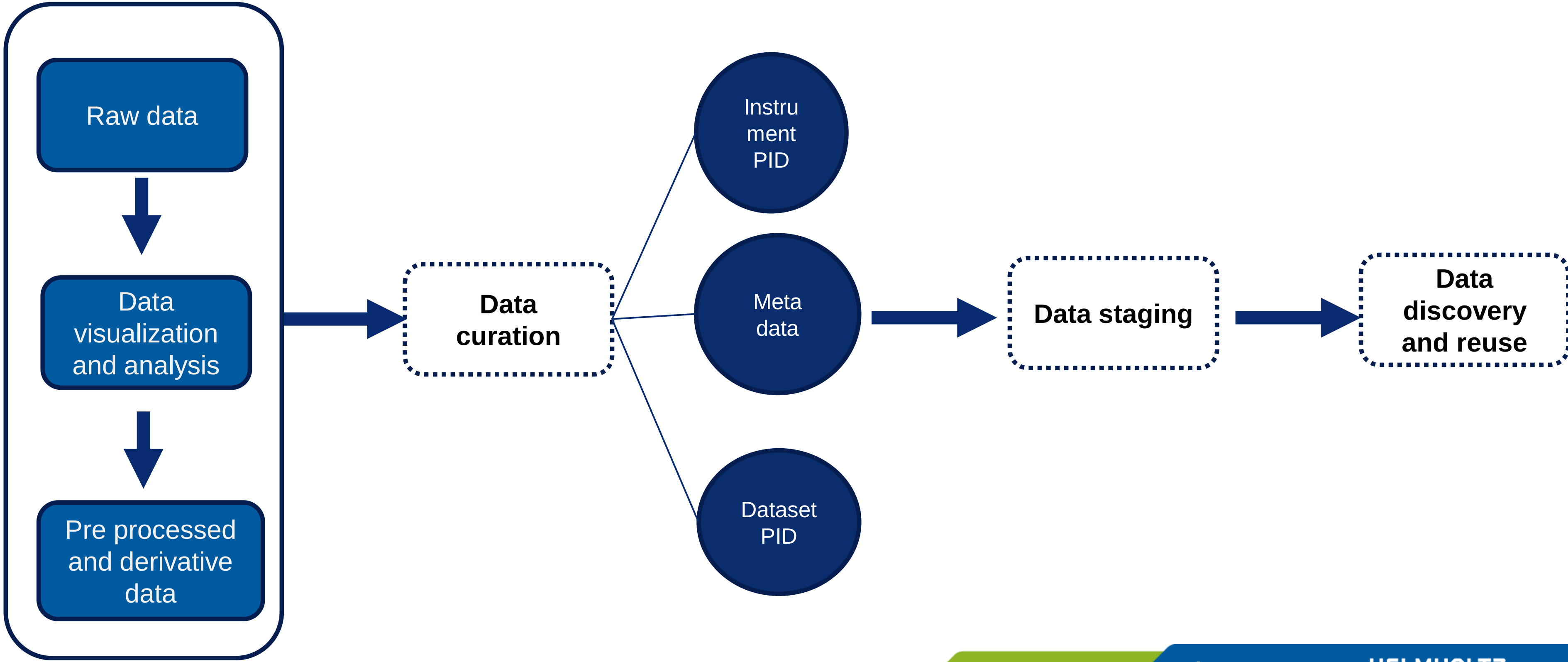
Which services are running to implement FAIR data access?



[3]

Pan Data life cycle at HZB

From data acquisition to data repository



Harvesting the metadata

At the measurement station

- **Automatize the metadata capture** with fully integrated tools in the measurement station system
- Identify among the machine metadata **relevant metadata** and **map** to measurement description- metadata schema
- **Optimize** the metadata harvesting **avoiding interference** with data acquisition/measurements

This tools integration is the focus of the collaboration Hub Matter, FDM-HZB and your group.

Entry points for metadata acquisition at beamline

Before the experiment

Proposal

- Proposal number ID
- Project description and meas. background
- Method
- Instrumentation
- Sample
- Scientific team
- Duration
- Instrument metadata

Facility

Beam line , Detectors

During the experiment

Facility logs

- Beam parameters
- Motor position
- Instrument configuration
- Detector settings
- Sample settings

Users log book

Parameter settings

changes respect proposal

sample description

Run logs

After the experiment

Analysis & derivative data

- Description of pre-process.
- Extraction procedure description
- External link to codes
- Software
- Workflow
- plots

Publication

Subset or full

quality parameter

reference to journal publ.

DOI

[4]

Services for the data description

Before the experiment

GATE

- Project description,
- authors, People metadata
- sample description,
- time frame, embargo

During the experiment

Elog notebook

- Measurements parameter, screenshot, free text notes
- Instruments logs
- Control system logs

After the experiment

Nexus-python routines, metadata server icat

- Description of pre-processing
- Instrument description
- External link to codes and software
- Licence
- Ancillary information

Guidelines for PaN metadata acquisition

Recommendations for key elements of metadata by ExPands working group on FAIR data in the PaN community:

- PI and CoII
- Requested instrument
- Sample description
- Facility
- Proposal ID
- Experiment description
- Experimental team /Experiment time
- Data format /raw data or derivative
- Acquisition and pre processing software
- File ID
- Creator
- Publisher
- Release date and use Licence

- Preservation description
- Representation information
- Instrument parameters
- Analysis software
- File generator system
- Lab Notebook should be attached
- Calibration information
- Data relationship
- File identifiers
- PIDs
- Workflow

Expands.eu

[25]

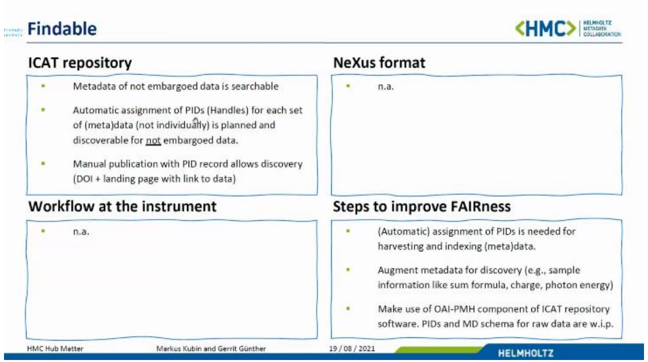
Indexing, search and retrieval of dataset, metadata at HZB

Tools: elog book+ Nexus writer + ICAT

- **Linking** publication and datasets
- Nexus for rich metadata (from proposal and from measurements setting)
- **Providing unique identification** for data sets
- **Restricted access** to the research data under HZB **data policy**
- **Access to data and metadata** via searchable online catalogue (ICAT)

Services maintenance and tools development
FDM+Hub Matter

Current implementation of FAIR data



Findable

ICAT repository

- Metadata of not embargoed data is searchable
- Automatic assignment of PIDs (Handles) for each set of (meta)data (not individually) is planned and discoverable for not embargoed data.
- Manual publication with PID record allows discovery (DOI + landing page with link to data)

NeXus format

- n.a.

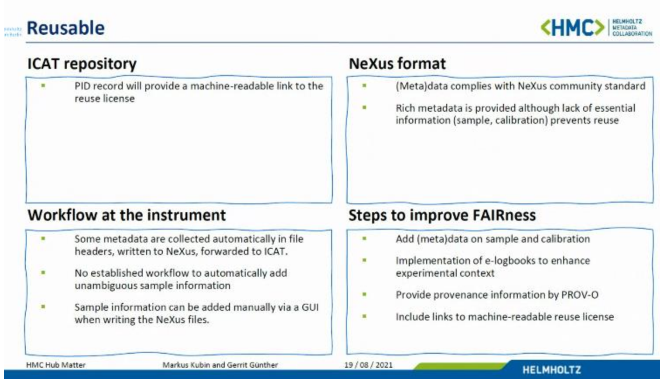
Workflow at the instrument

- n.a.

Steps to improve FAIRness

- (Automatic) assignment of PIDs is needed for harvesting and indexing (meta)data.
- Augment metadata for discovery (e.g., sample information like sum formula, charge, photon energy)
- Make use of OAI-PMH component of ICAT repository software. PIDs and MD schema for raw data are w.i.p.

HMC Hub Matter Markus Kubin and Gerrit Günther 19 / 08 / 2021



Reusable

ICAT repository

- PID record will provide a machine-readable link to the reuse license

NeXus format

- (Meta)data complies with NeXus community standard
- Rich metadata is provided although lack of essential information (sample, calibration) prevents reuse

Workflow at the instrument

- Some metadata are collected automatically in file headers, written to NeXus, forwarded to ICAT.
- No established workflow to automatically add unambiguous sample information
- Sample information can be added manually via a GUI when writing the NeXus files.

Steps to improve FAIRness

- Add (meta)data on sample and calibration
- Implementation of e-logbooks to enhance experimental context
- Provide provenance information by PROV-O
- Include links to machine-readable reuse license

HMC Hub Matter Markus Kubin and Gerrit Günther 19 / 08 / 2021

[5]

Instrument PID

- Similar to dataset DOI or ORCID but identifying instruments
- PaN facilities have a complex of measuring stations
- Internal manufactured stations and not standard instrument
- Versioning to track the evolution of the measuring unit
- Interlinking PIDs at each stage in the process to provide traceability
- **Further questions ?**

Rolf Krahl, HZB is the reference person for this development

[6]

SISSY I @EMIL PIDs

Instrument DOI:

- https://doi.org/10.5442/ni000018

Instrument landing page

- https://www.helmholtz-berlin.de/pubbin/igama_output?modus=einzel&sprache=en&gid=1978

Publication details

- https://commons.datacite.org/doi.org/10.5442/ni000018

PID interlinking improves / facilitate the data discovery

Sample ID to sample tracking

- Logs of all the operations (sample preparation) on the sample embed the sample history.
- This log file is preserved and attached to the metadata

PID interlinking improves / facilitate the data discovery

- Potential assignement of a PID ?

Considerations : sample charactr are changing while treated, measurements and sample creation can be really distinguished? What is the sample preservation after the measurements ?

NeXus Format

Metadata and data format: Nexus

- Nexus for rich and easy to access metadata.

  Why NeXus ? https://www.nexusformat.org/

- HDF5/NeXus used as institutional standard at neutron, x-ray and muon facilities
- Each facility diversify the dictionary limiting the immediate re-usability.
- NeXus files may help to improve the situation.
- HDF5 format and a tree structure for metadata representative of the complexity of PaN data
- Built in Vocabulary for research community interoperability
- Geometry of the beamline, sample stages, orientation and description of detectors, exposure time, beamline calibration info, scan description

Hierarchy in Nexus

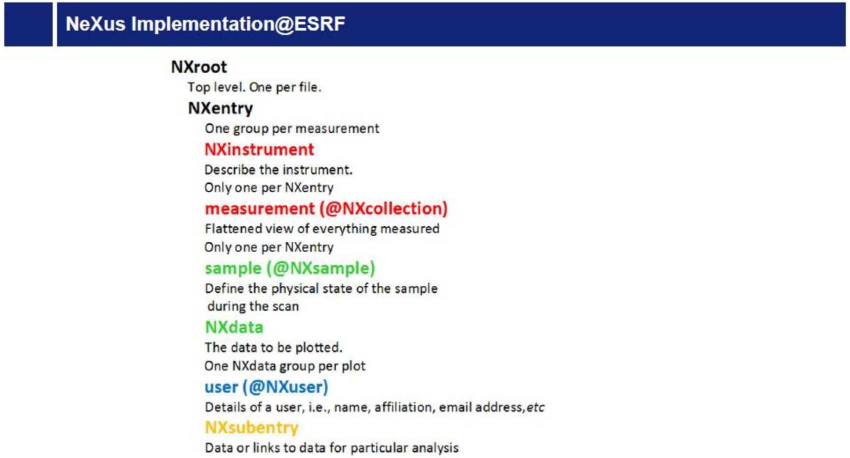
Classes (dictionary)

Groups

Levels

Attribute

MultiD array and scalars



NeXus Implementation@ESRF

NXroot
Top level. One per file.
NXentry
One group per measurement
NXinstrument
Describe the instrument.
Only one per NXentry
measurement (@NXcollection)
Flattened view of everything measured
Only one per NXentry
sample (@NXsample)
Define the physical state of the sample during the scan
NXdata
The data to be plotted.
One NXdata group per plot
user (@NXuser)
Details of a user, i.e., name, affiliation, email address, *etc*
NXsubentry
Data or links to data for particular analysis

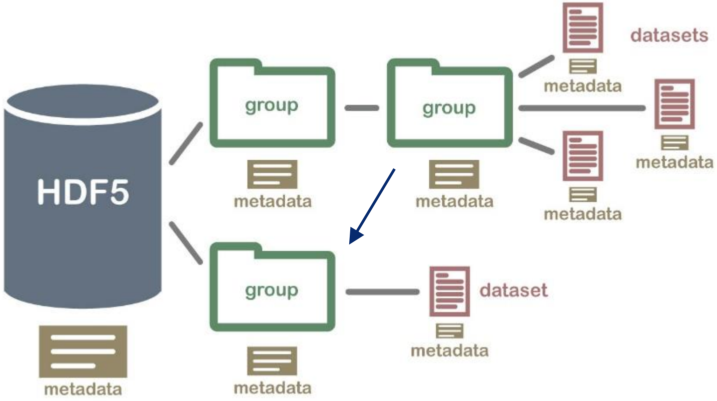
- To store multiple related data set create more entries

NeXus structure allows links and pointing to data stored in other parts of the group

[7]

HDF5 Format

- HDF5 format and a tree structure for metadata representative of the complexity of PaN data
- Allows chunked storage and slices reading
- Metadata can be attached
- The I/O can be faster than contiguous data files
- Compression
- Can be prefixed or open database size
- Heterogeneous database with links
- Platform independent
- Suitable for massive databases with a datatype and dataspace definition per dataset

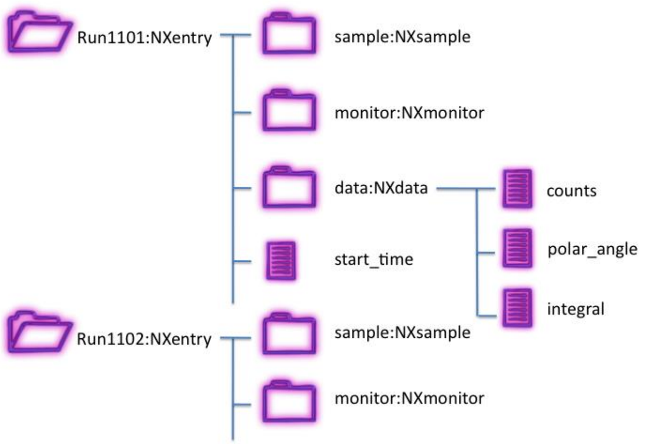


HDF5 structure allows links and pointing to data stored in other parts of the group

[8]

Nexus Format

- New applications can be defined
- A rich set of tools for verification and file validation (e.g. nxvalidate) and easy data extraction
- Process data must contain Nxprocess group
- Nxentry is mandatory root element



List of attributes

- http://definition.nexusformat.org/nxdl/3.1/
- https://github.com/nexusformat/definitions/blob/main/nxdlTypes.xsd
- Classes definitions in xml– mapping

Schema of the Nexus classes and attributes

- http://definition.nexusformat.org/nxdl/nxdl.xsd
- NXDL files must adhere to the specifications of the NeXus XML Schema, as defined in *nxdl.xsd* and *nxdlTypes.xsd*.

```
xmllint --noout --schema nxdl.xsd base_classes/NXentry.nxdl.xml
base_classes/NXentry.nxdl.xml validates
```

- Expands (expands.eu), Panosc (panosc.eu/) , CERIC-ERIC EOSC work on tools and standards for the interoperability of the PaN community data and connection to the European Open Science Cloud

[9]

Data analysis with Nexus files

- https://manual.nexusformat.org/utilities.html#data-analysis

- A number of Python routines to process X-ray photons emission data in hdf5
- XRF spectroscopy PyMCA
- https://gitlab.elettra.eu/panosc/xrffitvis/
- Xrayutilities for conversions spec hdf5

- IGOR pro can upload HDF5 www.wavemetric.com
- ORIGIN lab (+HDF5Browser App)
- DAWN
- Matlab
- Spec2hdf5 available tools (silx.org)
- Spec2nexus (https://spec2nexus.readthedocs.io)
- PyMCA (http://pymca.sourceforge.net/)
- NeXpy (http://nexpy.github.io/nexpy/)

[10]

HELMHOLTZ METADATA COLLABORATION

NeXus structure and playground

www.helmholtz-metadata.de

NeXus Writer and Icat data ingestion

- Ingestion workflow
- Access to the tools by virtual machine

- 1. data collection
- 2. identification of the instrument dictionary
- 3. sample data info collection
- 4. parameters attribution
- 5. local data saving
- 6. icat repository ingestion

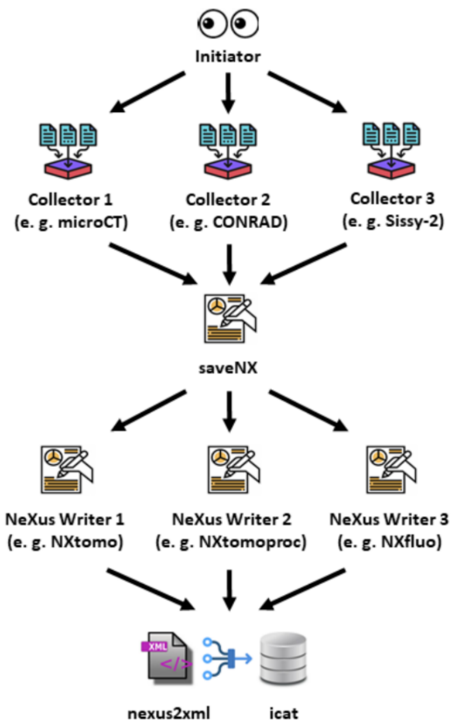
Contact person: Gerrit Günther, HMC

https://gitlab.helmholtz-berlin.de/jaf/nexuswriter/-/blob/master/nexusCore/icatRepo/nexus2xml.py

Nexus application for describing XAS

https://github.com/nexusformat/definitions/blob/main/applications/NXxas.nxdl.xml

- Sketch for the ingestion


Initiator
Collector 1 (e. g. microCT)
Collector 2 (e. g. CONRAD)
Collector 3 (e. g. Sissy-2)
saveNX
NeXus Writer 1 (e. g. NXtomo)
NeXus Writer 2 (e. g. NXtomoproc)
NeXus Writer 3 (e. g. NXfluo)
XML
nexus2xml
icat

Initiator: The entry point that starts a collector. There are various initiators ranging from command-line interface to GUI.

Collector: An experiment specific module that collects data from different sources and assigns values to a python dictionary {...}

saveNX: A distributor that starts the appropriate NeXus Writer according to the NeXus definition schema of an entry; it may start different writer routines for a file.

NeXus Writer: A module that reads the python dictionary {...} and writes its content to the NeXus file according to a specific NeXus schema.

nexus2xml: An interface to HZB's icat to satisfy its demands: structures the produced files in folders, reads their content and writ a summary of searchable terms to a xml file before starting the inge

Test the tools: NeXus writer and icat ingestion

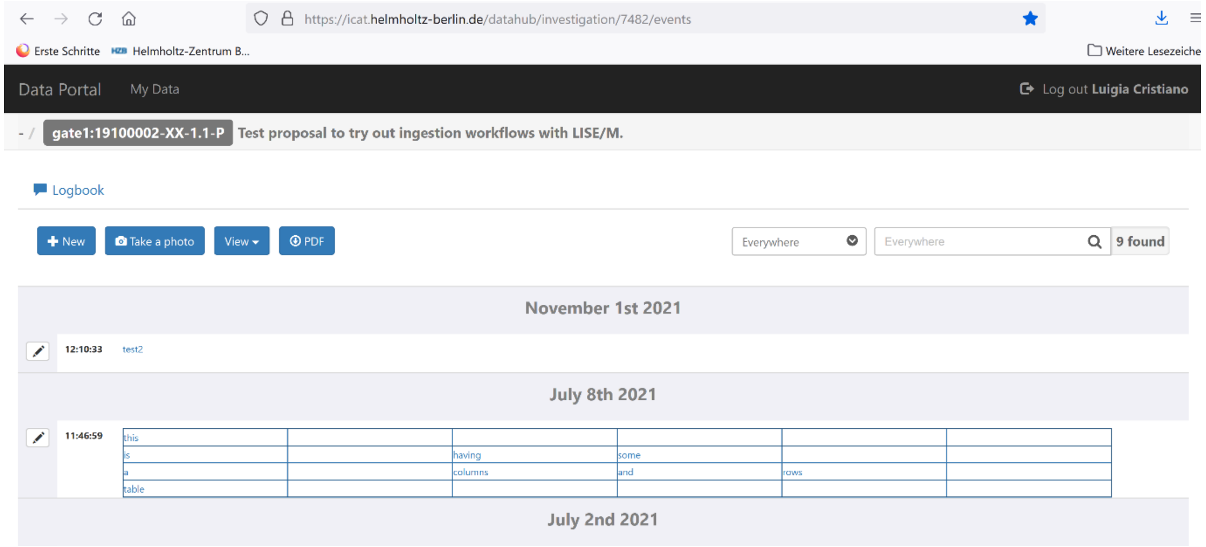
User portals : Elog book

- Notebook to optimize the data reuse with annotations, protocols
- The data in the notebook are under embargo
- Software esrf notebook
- **Web interface**

https://icat.helmholtz-berlin.de/datahub/

Time stamped in chrono o seq order

can be used by users or software



- / gate1:19100002-XX-1.1-P Test proposal to try out ingestion workflows with LISE/M.

November 1st 2021

12:10:33 test2

July 8th 2021

11:46:59

this					
is		having	some		
a		columns	and	rows	
table					

July 2nd 2021

- Contact person: Rolf Krahl

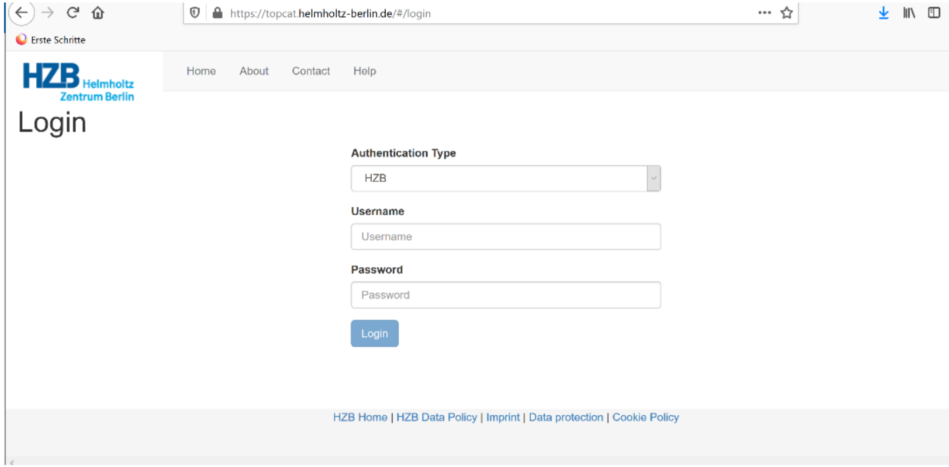
- Umbrella ID ; https://www.umbrellaid.org/
- Keycloak log in services are in implementation https://www.keycloak.org/

[5][11]

User portals: access to Icat metadata repository /Data catalogue

- Provide access to the data and is a DBMS developed at PaN facilities icatproject.org
- The data request means data extracted from tape and allocated on disk
- The software uses a **schema** or metadata, a web interface to the database
- **Dataservice** for data upload and download
- **Web interface** to search the data (topcat)

https://topcat.helmholtz-berlin.de/#/login


https://topcat.helmholtz-berlin.de/#/login
Erste Schritte
HZB Helmholtz Zentrum Berlin
Home
About
Contact
Help
Login
HZB Home | HZB Data Policy | Imprint | Data protection | Cookie Policy

Login

Authentication Type: HZB

Username: Username

Password: Password

- **Further questions ?**

  Rolf Krahl

[12]

EMIL implementation

In development ! Will is giving you technical details on this collaborative work

Read out routine of the **sample database** using the time of the measurement to extract the corresponding sample history.

A **manual** integration vs **automatic** readout

is possible

Notes, comments and images can be attached,

Sample information from proposal is also possible

Information from elogbook can also be attached

Python routines to ingest the data do not interfere with the measurements workflow

Reference to Guenther et al. ICALEPCS 21

[14]

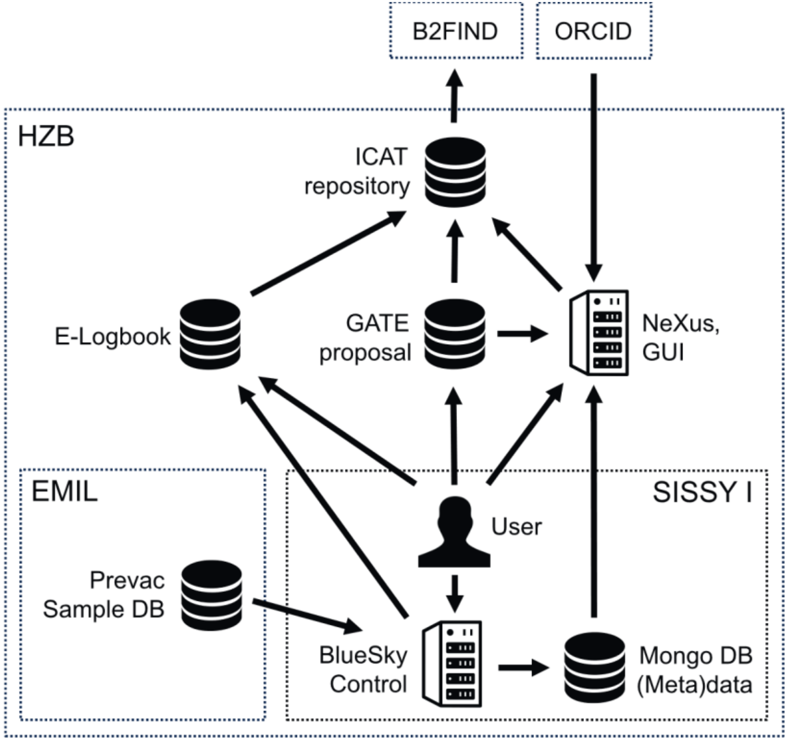


Figure 2: Visualization of the data infrastructure connecting the SISSY I instrument with services inside and outside the HZB

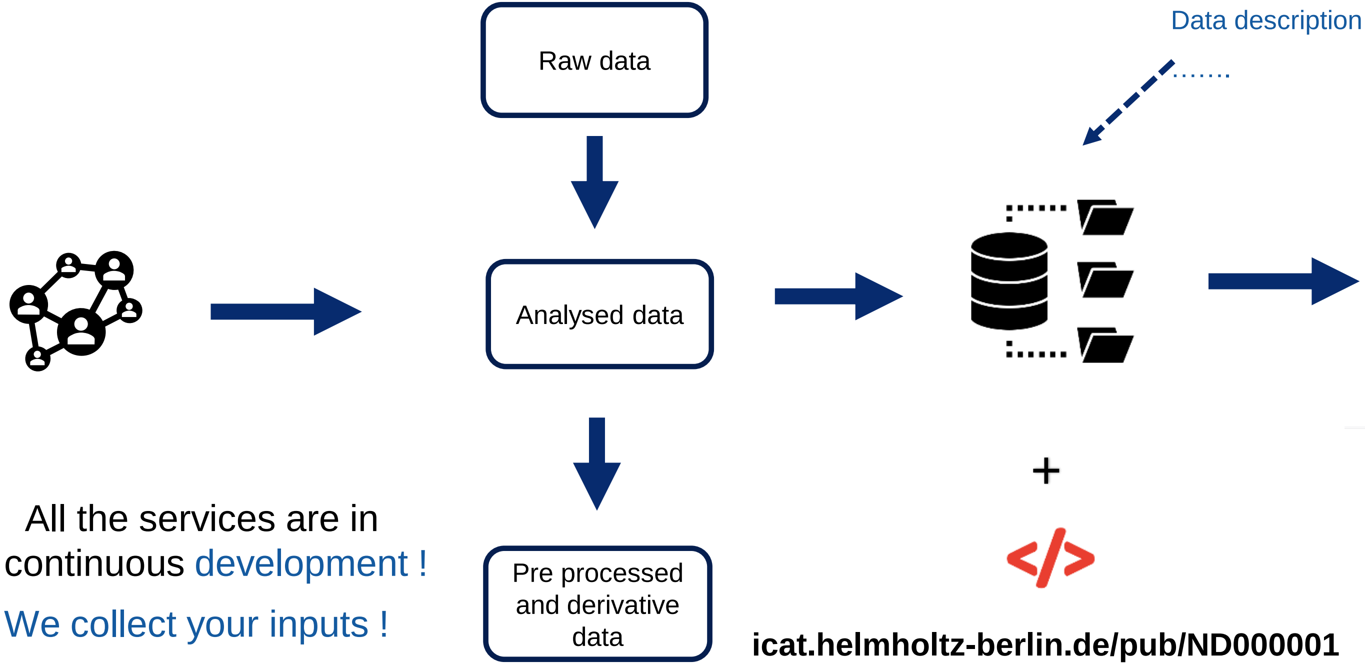
Data publication at HZB

…Quality control, direct upload of xml files, overview of the available publications

Data description template

…….

Findable Accessible Interoperable Reusable

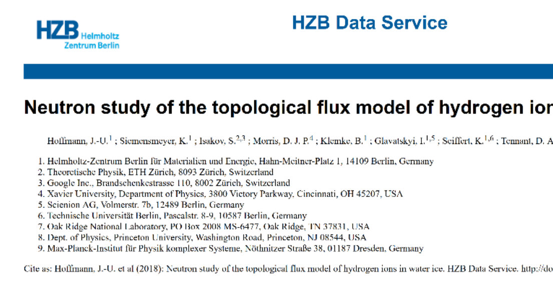


All the services are in continuous development !

We collect your inputs !

icat.helmholtz-berlin.de/pub/ND000001
https://doi.org/10.5442/ND000001

+


HZB Helmholtz Zentrum Berlin

HZB Data Service

Neutron study of the topological flux model of hydrogen ior

Hoffmann, J.-U.[1] ; Siemensmeyer, K.[1] ; Isakov, S.[2,3] ; Morris, D. J. P.[4] ; Klemke, B.[1] ; Glavatskyi, I.[1,5] ; Seiffert, K.[1,6] ; Tennant, D. A

1. Helmholtz-Zentrum Berlin für Materialien und Energie, Hahn-Meitner-Platz 1, 14109 Berlin, Germany
2. Theoretische Physik, ETH Zürich, 8093 Zürich, Switzerland
3. Google Inc., Brandschenkestrasse 110, 8002 Zürich, Switzerland
4. Xavier University, Department of Physics, 3800 Victory Parkway, Cincinnati, OH 45207, USA
5. Scienion AG, Volmerstr. 7b, 12489 Berlin, Germany
6. Technische Universität Berlin, Pascalstr. 8-9, 10587 Berlin, Germany
7. Oak Ridge National Laboratory, PO Box 2008 MS-6477, Oak Ridge, TN 37831, USA
8. Dept. of Physics, Princeton University, Washington Road, Princeton, NJ 08544, USA
9. Max-Planck-Institut für Physik komplexer Systeme, Nöthnitzer Straße 38, 01187 Dresden, Germany

Cite as: Hoffmann, J.-U. et al (2018): Neutron study of the topological flux model of hydrogen ions in water ice. HZB Data Service. http://do

[16]

Development directions

- **Federated service** for data discovery

(~20 PaN research facilities involved)

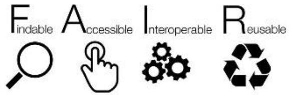
Interlinking PIDs at each stage in the data process to provide traceability

EOSC data catalogues services and analysis services

- Development of **research group- community dictionary** to implement as standard.

- Define **application** for a general representation of the beamlines measurement and specific for each end station or research group
- **Enable elogbook access to external data authors**
- Actions: gather ideas, test new implementation, optimizing and updating the tools

[17]



- Further **tailoring of tools**
- Test the workflow in action

- HZB in ExPaNDS and Pansoc that work to bridge the national facilities into EOSC and OpenAire

Initiatives focusing on the definition of ontologies and vocabularies for experimental techniques

- External users: remote desktop applications_ VISA, FastX and Jupyter notebook providing a recipe for data analysis

Material sources

- 10. https://zenodo.org/record/4424770#.YXKNwedCRzp
- 2. I. Boscaro.Clarke, F. Cesmat, K. Roarty. 2020. Expands vision and roadmap. 10.5281/zenodo.4424770
- 1. B.Matthews, Expands symposium for librarians, 2021
- 6. Rolf Krahl Berlin 2018; Persistent Identification of Instruments WG

  www.rd-alliance.org/groups/persisten-identification-instrument-wg
- Rolf Krahl workshop for Research data management at HZB, 2019

https://www.helmholtz-berlin.de/media/media/spezial/events/datenmanagement/5-hzb-icat.pdf

17. Collins et al., 2021. ExPaNDS ontologies v1.0. 10.5281/zenodo.4806026

HELMHOLTZ
METADATA
COLLABORATION

luigia.cristiano@helmholtz-berlin.de
heike.goerzig@helmholtz-berlin.de
gerrit.guenther@helmholtz-berlin.de
rolf.krahl@helmholtz-berlin.de
markus.kubin@helmholtz-berlin.de
oonagh.mannix@helmholtz-berlin.de

Helmholtz-metadata.de/en/pages-helpdesk
Hmc-matter@helmholtz-berlin.de

Thanks.

HMC Hub Matter

Development strategies

There were great efforts put in the last two decades in the research community to elaborate a common standard for high data-rate macromolecular crystallography (HDRMX). This agreed “Gold Standard” builds on the NeXus/HDF5 NXmx application definition and the International Union of Crystallography (IUCr) imgCIF/CBF dictionary, and it is compatible with major data-processing programmes and pipelines. Here we demonstrate the EuXFEL data packed into a NeXus file, which is fully compliant with the Gold Standard by design, since it is built directly from HDRMX NeXus definitions. We use open-source software developed both by community (cctbx) and in-house (extra-data).

Small files generated t beamline better to be encapsulated in larger HDF5 for the ingestions in tapes.
In the framework of national and international initiatives work on homogeneous solutions that take nevertheless in account the diversity of each infrastructure

Need of have granted access to the data from the external