

Using ICAT for Research Data Management at HZB

Rolf Krahl

Shared Metadata and Data Formats for Big-Data Driven Materials
Science: A NOMAD-FAIRDI workshop, Berlin, July 2019

HZB operates large scale scientific facilities, including the neutron source BER II and the synchrotron radiation source BESSY II. It offers access to these facilities to the scientific community.

Research data management at HZB is motivated by:

- Good Scientific Practice: scientific data used as the basis for a publication should be kept for at least ten years in the institution of their origin.
- Open Access: data obtained from research using public funding should be openly accessible.
- FAIR Data: scientific data should be Findable, Accessible, Interoperable, and Reusable.
- Research data management at HZB is considered to be part of the service that we deliver to our users.

We consider following classes of data:

Raw Data Data collected at HZB's instruments as a result of a measurement.

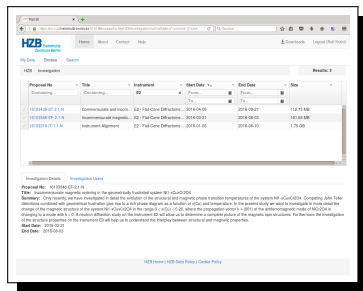
Results Data created by the user from the analysis or interpretation of raw data.

Metadata Data pertaining to other data, describing them and putting them into a context. Metadata may be associated to raw data or to results.

- HZB Data Policy regulates management of scientific data from public research at HZB's large-scale facilities.
- Acceptance of the Data Policy is a condition of the award of beamtime.
- Distinguish raw data, results, and metadata.
- Raw data and associated metadata will be curated and stored by HZB for at least ten years.
- Raw data and associated metadata are placed in the public domain (Creative Commons CC0 Dedication).
- Access to raw data and associated metadata is restricted to their creators for an embargo period of five years. After that, they become openly accessible.
- Results and associated metadata may be stored with the raw data. They will not be curated by HZB. They may be made openly accessible upon request of their creators.

ICAT Metadata Catalogue

- Access to the data is provided by the ICAT metadata catalogue.
- Search for the user's own and for public data.
- A request to download data automatically triggers the staging of that data from tape to disk. The data may be downloaded after the staging is complete.
- ICAT is developed as free software in cooperation with other Photon and Neutron sources (STFC, DLS, ISIS, ESRF, HZB).

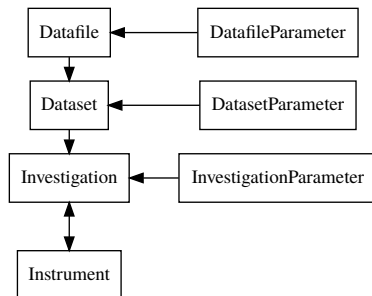


The screenshot displays the ICAT Metadata Catalogue web interface. At the top, there is a navigation bar with the HZB logo and links for Home, About, Contact, and Help. Below this, there are tabs for 'My Data', 'Dataset', and 'Search'. The main content area shows a table of investigations with columns for 'Proposed No.', 'Title', 'Instrument', 'Start Date', 'End Date', and 'Size'. The table contains three rows of data:

Proposed No.	Title	Instrument	Start Date	End Date	Size
10-00148-01-2-1-4	Coarse-tuning and scans	S2 - Flat-Cone Diffraction	2018-04-08	2018-08-01	112.11 MB
10-00148-01-2-1-6	Instrumentwide magnets	S2 - Flat-Cone Diffraction	2018-02-01	2018-08-02	162.93 MB
10-00148-01-1-1-6	Instrument Alignment	S2 - Flat-Cone Diffraction	2018-01-05	2018-08-10	1.71 GB

Below the table, there are links for 'Investigation Details' and 'Investigation Links'. The 'Investigation Details' section shows the 'Proposed No.' 10-00148-01-2-1-4 and a title. The 'Investigation Links' section contains a paragraph of text describing the investigation, mentioning the use of the instrument S2 and the goal of understanding the coupling between structural and magnetic properties. At the bottom of the page, there is a footer with the text 'HZB Home | HZB Data Policy | Contact Policy'.

- Central elements for organizing the data in ICAT are:
Investigation ← Dataset ← Datafile.
- Correspondence:
 - Investigation $\hat{=}$ Proposal,
 - Dataset $\hat{=}$ Measurement,
 - Datafile $\hat{=}$ File.
- DatasetParameter allows storing physical metadata of the measurement in a simple keyword/value schema.



We distinguish two classes of metadata:

Administrative metadata

- Proposal, title, abstract, user, access rights.
- Get imported from the user office portal GATE beforehand.
- Relevant for the control of internal ICAT workflows.

Physical metadata

- Parameter of the measurement, sample etc.
- Will be collected (preferably) automatically.
- Will be stored in the datafiles along with the raw data before ingestion into ICAT.
- A selection of the metadata may be additionally stored in ICAT as `DatasetParameter`, if they are relevant for the search of data.

Steps in the life cycle for scientific data at HZB:

- 1 User submits a proposal in GATE.
- 2 Proposal gets accepted.
- 3 Administrative metadata get imported from GATE into ICAT.
(For data not related to a proposal, create an Investigation in ICAT instead.)
- 4 User comes to HZB and performs experiments. Data is collected and curated at the instrument.
- 5 Data are ingested from the instrument into ICAT.
- 6 User have exclusive access to their data.
- 7 User may optionally upload results from data analysis into ICAT.
- 8 After expiring of the embargo period, the raw data and associated metadata become openly accessible. The user may decide to make also the results openly accessible.

- Curation of raw data should be integrated into the measurement workflow and should be automated as much as possible.
- Curation includes the choice of suitable datafile formats and storing the metadata of the measurement along with the raw data in these files.
- Aim at describing the instrument and the measurement technique as good as possible to allow third party to understand the data.
- Collect all relevant parameters of the measurement in the metadata.
- Data curation workflow must be defined under the lead of the instrument scientist with assistance from IT.
- File format: recommend NeXus as a particular useful datafile format, but need also to respect standards of the respective scientific community.

- An ICAT component implementing the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is currently under development at HZB.
- It will allow the automatic harvest of metadata from the ICAT by other data collections.
- OAI-PMH could be a path to access HZB experimental data by NOMAD.
- This would require the specification of the metadata items of interest and the formulation of a suitable (even private) metadata standard.

Data Publications

- As an additional service to our users, we will offer to make data publications.
- It will require an additional step of collecting and editing the bibliographic metadata for the dataset.
- The data publication will receive a DOI and is fully citable.





The screenshot shows a web browser displaying the 'HZB Data Service' page. The page title is 'Supplement to: Machine learning classification for field distributions of photonic modes'. Below the title, it lists the authors: 'Berk, Cole¹, Becker, Christian²' and the affiliation: '1. Humboldt-Universität Berlin, 2. Leibniz-Institut für Photonische Technologien, Albert-Ludwigs-Str. 11, 11149 Berlin, Germany'. The page content is divided into sections: 'Abstract', 'Keywords', 'Dates', and 'Metadata'. The 'Abstract' section contains a detailed description of the photonic structures and the machine learning classification process. The 'Keywords' section lists terms like 'Photonic crystals', 'Laser modes', 'Machine learning', etc. The 'Dates' section shows a list of dates from 2019 to 2020. The 'Metadata' section shows a list of metadata items, including 'arXiv:1908.08111v1 [physics.optics]' and 'doi:10.1101/2019.08.11.261828'.

Current status:

- The storage systems are available.
- ICAT operates in test production. Mostly ready for production.
- Two BER II instruments (E2 and E9) register routinely their data.
- Implementation at the first BESSY II station (NanoclusterTrap) is in preparation and will begin shortly.

Implementation:

- For each instrument, we need to hook up the data curation and the data ingestion to ICAT into the measurement workflow.
- This requires consideration of the prerequisites and the workflows at the instrument individually.
- We will proceed instrument by instrument.

-  **Helmholtz-Zentrum Berlin für Materialien und Energie.**
HZB Data Policy, Version 1.1, 2017.
[https://www.helmholtz-berlin.de/pubbin/vademecumdatei?did=326.](https://www.helmholtz-berlin.de/pubbin/vademecumdatei?did=326)
-  **The ICAT Project.**
ICAT Project Home.
[https://icatproject.org/.](https://icatproject.org/)
-  **The ICAT Project.**
GitHub Source Repository.
[https://github.com/icatproject.](https://github.com/icatproject)
-  **Open Archives Initiative.**
The Open Archives Initiative Protocol for Metadata Harvesting,
version 2.0, 2002.
[https://www.openarchives.org/pmh/.](https://www.openarchives.org/pmh/)