Datasets in Operando Experiments

Rolf Krahl

ICAT F2F Meeting, 10 February 2026, Abingdon, UK

Structure of Data in HZB ICAT

Data in HZB ICAT is structured in Investigations and Datasets:

Investigation

The Investigation sets the scientific context. It describes *why* the data has been created. For most of the data at HZB, the Investigation corresponds to a proposal in the user portal. Access permissions are managed at the level of the Investigation.

Dataset

The Dataset sets the context of the creation of the data. It describes *how* the data has been created. For data collected during an experiment, a Dataset usually corresponds to an individual measurement, experiment or simulation. Physical parameters of the measurement are stored as Dataset parameters. In general, Datasets are the smallest pieces of data to be referenced.

Structure of Data in HZB ICAT (cont.)

- Investigation and Datasets automatically get a PID.
- Investigations are disseminated as *Collection*, Datasets as *Dataset* in DataCite metadata.
- We use `ids.server` with two level storage and `storageUnit=dataset`, so whole Datasets are stored as a ZIP in archive.

Operando Experiments

- Operando experiments become more and more important at HZB: long running experiments with many intermittent beamline measurements.
- Examples:
  - A catalysis experiment runs long time. Many XAS measurements are taken during the experiment in order observe the chemical reaction while it happens. (Model experiment in ROCK-IT.)
  - Batteries are cycled, e.g. continuously charged and discharged for multiple hours. Tomography measurements are used to observe and understand the progress of the degradation of the electrodes during the cycles.
- Question: what is "the measurement" here? Is it the whole operando experiment? Or is it each individual beamline measurement?
- How should we store the data? As one big dataset for the whole thing or each beamline measurement as an individual dataset?

Option: One Big Dataset

- We could store the whole thing in one single big dataset.
- We would still spit it into separate files, one master HDF5 file for the outer operando experiment, multiple NeXus files for the individual beamline measurements.
- Use HDF5 file links from the master to the individual NeXus files to put everything neatly together into one hierarchy.
- Pros:
  - The context of the whole experiment remains intact.
  - Inter HDF5 file links are more likely to remain intact, if the whole thing is downloaded together.
- Cons:
  - Could not store the parameters of the beamline measurements as dataset parameters, could not search (as easily) for the individual beamline experiments.
  - Could not reference the individual beamline measurement by its PID.
  - At least in the battery case, the data volume for the whole thing will become prohibitively huge to be treated as one dataset.

Option: Separate Datasets

- Store the data in multiple datasets: one dataset for the outer operando experiment, an individual dataset for each beamline measurement.
- Pros:
  - We could store the beamline measurement with their dataset parameters as usual. They could individually be searched for in the catalogue.
  - The beamline measurement would get its dataset PID and could individually be referenced.
  - The previous bullet is particularly important for Tomography, as the reconstruction is usually been done later and stored as a separate dataset that would need to reference the corresponding raw data.
  - Much easier to handle the data volume in palatable pieces.
- Cons:
  - The context risks to get lost. It is not apparent that these multiple datasets belong to one experiment.
  - Inter HDF5 links are more difficult to maintain.

Discussion

- Do other facilities have the same problem? How do you handle that?
- Is there anything we could do better in ICAT?
- Maybe, we would need a way to mark several datasets as belonging to a common context and should be looked at together?