Metadata for material science at the Lightsource BESSY II

T. Birke[1], V. Laux[1], T. Mertens[1],R. Muller[1], A. Schaelicke[1], P. Schnizer[1], T. Unold[1], L. Vera Ramirez[1], J. Viefhaus[1]

[1]Helmholtz-Zentrum Berlin (HZB), Germany

11/06/2019

Motivation

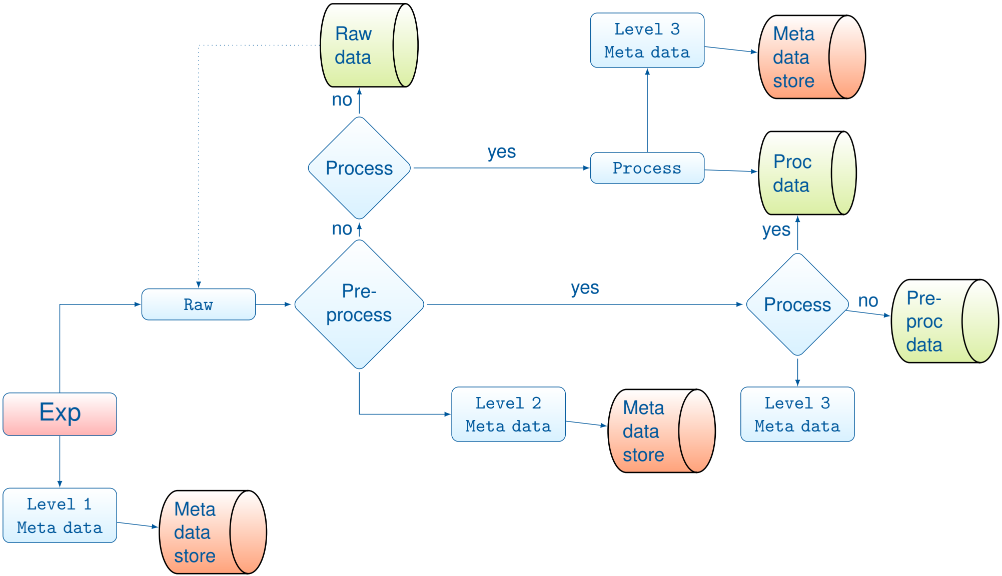
Make data conform with the FAIR concept[1]:

- Findable
- Accessible
- Interoperable
- Repurposable

[1] https://www.fairdi.eu/uploads/documents/FAIRmat_Konzeptpapier.pdf

Table of Contents

Data flow



Raw data
no
Level 3 Meta data
Meta data store
Process
yes
Process
Proc data
no
yes
Raw
Pre-process
yes
Process
no
Pre-proc data
Exp
Level 2 Meta data
Meta data store
Level 3 Meta data
Level 1 Meta data
Meta data store

Some considerations

Some (meta)data considerations

- Auto-generated?
- Manual?
  - Digital (elog entries)
  - Analog (handwritten $\rightarrow$ kicked device, then it worked)
- Persistent IDentification of data (unique, versioned,...)
- Logbook interface
- Storage: SQL/noSQL, file formats
- Scalability: can we handle the future?
- **Data Model?**
- Searchability vs amount of metadata

Some side notes (issues acc phys):

- Running software on different systems
- Version control
- File formats
- Maintainability of packages/tools
- Containerization : Singularity with a SCIentific File System (SCIF)

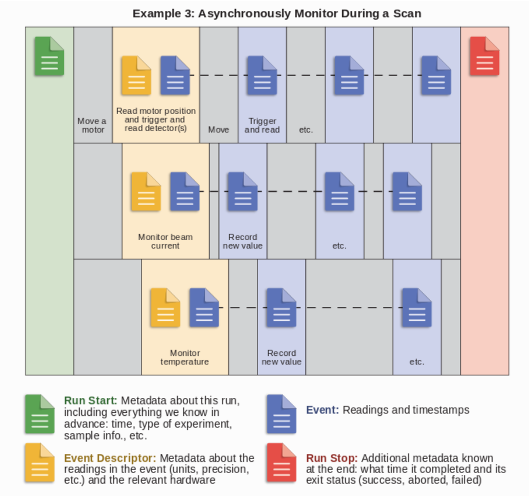
Epics, the archiver and metadata

- Control System BESSY: Epics 3.14 (3.15 under way)
- Epics variables with unique naming convention (location, device type, etc...) metadata
- all channels logged in archiver
- accelerator metadata = data
- Question: which data is metadata for some exp

- We want: Near Real Time Simulation / Analysis
- Why? Performance optimization, fast recovery, machine protection and maintenance
- Experiments: beam commissioning

Example: Bluesky and Ophyd

- Python
- Ophyd for device abstraction (epics, labview, but also extendable)
- Bluesky for experiment control and planning
- Nice data model (see right)
- Databroker available (base: sqlite and MongoDb)
- Suitcase for elasticsearch developed in house (ohters can be easily produced)
- possibility to store data in external files but keep links to data in database (adaptors can be written in straightforward way to load/save the data)
- Generates unique ID for each experiment
- Can talk to Olog
- **METADATA** and data hints
- live plotting and fitting

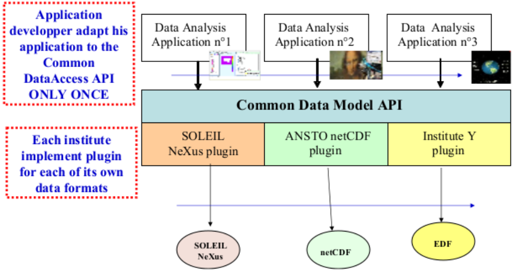


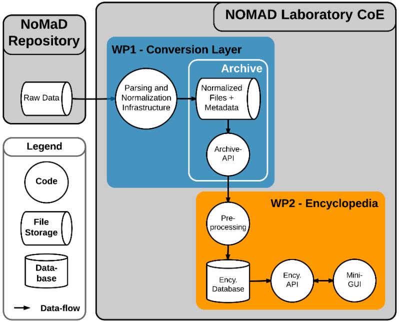
Example bluesky metadata


```
[{'_id': 'OPDFjmoBqZm8A59lj93a',
  '_index': 'start_run',
  '_score': 3.9233167,
  '_source': {'detectors': ['bpm'],
              'hints': {'dimensions': [[['motor2'], 'primary'],
                                       [['master_clock_frequency_readback'],
                                        'primary']]},
              'motors': ['motor2', 'master_clock_frequency'],
              'num_intervals': 49,
              'num_points': 50,
              'plan_args': {'cycler': "(cycler(MasterClockFrequency(prefix='', "
                                      "name='master_clock_frequency', "
                                      "parent='master_clock', settle_time=0.0, "
                                      "timeout=2.0, read_attrs=['setpoint', "
                                      "'readback', 'offset'], "
                                      'configuration_attrs=[], limits=(499626, '
                                      "499634), egu='kHz'), [499623.43033, "
                                      '499623.65255222225, 499623.87477444444, '
                                      '499624.0969966667, 499624.3192188889, '
                                      '499624.5414411111, 499624.7636633333, '
                                      '499624.98588555556, 499625.20810777775, '
                                      '499625.43033]) * '
                                      "cycler(SynAxis(prefix='', "
                                      "name='motor2', read_attrs=['readback', "
                                      "'setpoint'], "
                                      "configuration_attrs=['velocity', "
                                      "'acceleration']), [0, 1, 2, 3, 4]))",
                            'detectors': ["BPMStorageRing(prefix='', "
                                          "name='bpm', read_attrs=['stat', "
                                          "'stat.mean_x', 'stat.mean_y', "
                                          "'stat.rms_x', 'stat.rms_y', "
                                          "'waveform', 'waveform.packed_data', "
                                          "'waveform.counter', "
                                          "'waveform.ready', 'waveform.pos_x', "
                                          "'waveform.pos_y', "
                                          "'waveform.intensity_z', "
                                          "'waveform.intensity_s', "
                                          "'waveform.status', 'waveform.gain', "
                                          "'waveform.rms_x', "
                                          "'waveform.rms_y'], "
                                          "configuration_attrs=['stat', "
                                          "'waveform'])"],
                            'per_step': 'None'},
              'plan_name': 'scan_nd',
              'plan_type': 'generator',
              'scan_id': 1,
```

Beamlines

- Diverse systems, tools and requirements
- No general work-/dataflow - user/beamline specific
- New "workflows" (related to (meta)data) should not change users habits or create extra overhead
- Metadata database for static metadata (See talk Heike)
- Experiment controls: Bluesky / Tango / Sardana → Python
- Common Data Model Architecture vs FAIRmat → **Data Model?**





What is FAIRmat?

A Proposed Consortium of a German Research-Data Infrastructure (NFDI) on FAIR Data Infrastructure for Materials Science and Related Research Field, representing the materials science pillars of the association FAIR-DI (FAIR Data Infrastructure for Physics, Chemistry, Materials Science, and Astronomy e.V.).

FAIRmat

Scientific data are a significant raw material of the 21st century. To exploit its value, a proper infrastructure that makes it Findable, Accessible, Interoperable, and Re-purposable – FAIR – is a must. For the fields of computational and experimental materials science, chemistry, and astronomy, FAIRDI sets out to make this happen. This enabling of extensive data sharing and collaborations in data-driven sciences (including artificial intelligence tools) will advance basic science and engineering, reaching out to industry and society.

- Computational (NOMAD)
- Experimental
- Synthesis
- Functional
- Digital research infrastructures including cyber-security
- Artificial Intelligence Tools

source: https://fairdi.eu/fairmat

NOMAD - intro

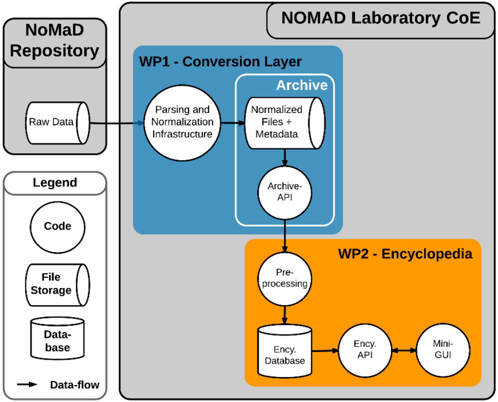
From NOMAD public report

(https://www.nomad-coe.eu/uploads/outreach/Public_Deliverables/NOMAD_D2_1_public_20June2016.pdf).

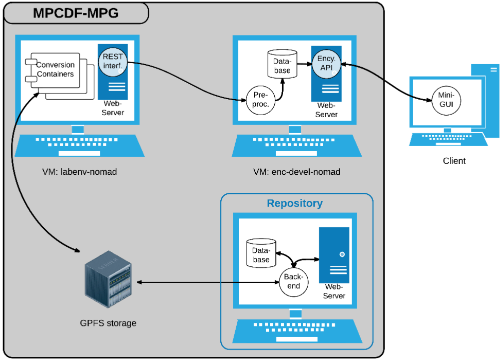
- “The Novel Materials Discovery Laboratory” (NOMAD)
- to develop a Materials Encyclopedia and Big-Data Analytics Tools for materials science and engineering
- The NoMaD Repository (https://NoMaD-Repository.eu) is an open access database
- maintained by the groups of Matthias Scheffler (FHI-MPG), Claudia Draxl (HUB), and Stefan Heinzel (MPCDF-MPG)
- Repo fulfills scientific needs and legal requirements well, e.g. by guaranteeing data storage for at least 10 years and optionally keeping the files private for up to three years
- heterogenous data and data sources $\rightarrow$ conversion layers (over 40)

NOMAD - infrastructure

Software infrastructure



System infrastructure



Something similar will be necessary to for the NFDI application with FAIRmat/ FAIR-DI.

source: https://www.nomad-coe.eu/uploads/outreach/Public_Deliverables/NOMAD_D2_1_public_20June2016.pdf

FAIRmat - HZB

- NFDI - 8 projects will be funded
- 1.6-3.9 M€ per year (personnel)
- Deadline 15 October 2019
- Community support / analysis tools / more (searchable) data sources / computer resources
- Discussions on meta(data) model - input is needed!!!
- Somebody will have to do it (Experiment control or IT group?)