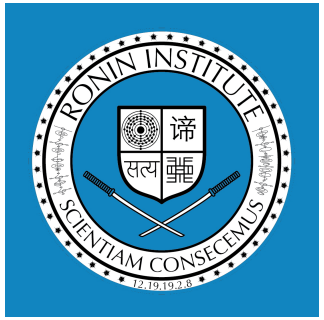
Gold Standard Format Proposal

Herbert. J. Bernstein

Ronin Institute for Independent Scholarship

Notes for HDRMX Meeting at Diamond Light Source
6-7 Nov 2019



rev 9 Dec 2019

BROOKHAVEN
NATIONAL LABORATORY

Changes in metadata

- Since we introduced the CBF format for the Dectris Pilatus detectors in 2007, there has been a recognition of the importance of controlling the metadata associated with images in order both to ensure that essential information is not lost and to minimize delays in handling the metadata
- When the Eiger detectors were introduced, the community agreed to adopt the NeXus/HDF5 format for efficiency in handling the much larger volume of data with fewer files to reduce filesystem and network burdens, but much of the metadata is carefully aligned between NeXus/HDF5 and CBF under an agreement between the NeXus International Advisory Committee (NIAC) and the IUCr Committee for the Maintenance of the CIF Standard (COMCIFS).
- With the cooperation of Dectris, the High Data Rate Macromolecular Crystallography (HDRMX) group and website were established to facilitate community discussion of the software, data and metadata.

HDRMX Discussions on Metadata

- There are signs of divergence among beamlines in Eiger formats and it is time to add new metadata, for example to identify beamlines and facilities and to record metadata that will be helpful in PDB depositions.
- The primary objective is to ensure that sufficient metadata will be provided to allow processing at a facility other than the one at which the data was produced. In particular, detailed descriptions of axis chains to be used to process the data are needed, both for sample goniometers and detector positioners.
- There were two informal HDRMX dinner meetings in summer 2019 to discuss a new “gold standard” for the NeXus HDF5 Eiger format, one at the July 2019 ACA meeting in Covington, KY and one at the August 2019 ECM32 in Vienna, AT
- We hope to come to final agreement on the changes at this HDRMX meeting at Diamond Light Source, 6-7 Nov 2019

Structure of the New Metadata

- In general the requested augmentation of metadata is divided into two groups
- First, metadata to be added via a templating mechanism in the Dectris software to be set-up before collection as static changes to the "master" files, and,
- Second, metadata to be added after collection, possibly via H5copy.

For simplicity we refer to the former as static and the latter as dynamic.

Static Metadata:

- Some tags for static (*i.e.* Dectris template) additions are already available. imgCIF defines AXIS tags needed for specification of arbitrary and very general axis chains. NeXus defines the equivalent information in the NXtransformations base class.
- Concern has been expressed about cluttering the templating mechanism with large numbers of tags used only in the most complex cases.

Static Metadata

- To avoid such clutter the input to the template can be the path to either a CBF or a NeXus file with the appropriate axis information, along with the necessary software to automatically convert between CBF and NeXus axis conventions. One way or another all diffraction geometry and all detector geometry need to be described.
- Tags have been defined to carry metadata specifying the beamline and facilty.
- Note that the detector distance, wavelength and beam center are already specified and very necessary.
- As integrating detectors or other detectors that do not count single photons come into use in this performance range, detector gain will need to be specified.
- Tags are needed for the HDF5 software version, to declare the use of non-standard local format conventions, to list the files comprising a dataset, and to give the format of each particular file.

Static Metadata Example

- As a partial example consider a beamline called XXX (ID1) at site SYNC with an omega axis, and pin_x, pin_y and pin_z translation axes stacked 5 millimetres apart, using hdf5_1.8.14 and NXmx 1.4. Then a portion of the necessary information presented as a CIF file might be:

```
data_AMX_metadata
    loop_
    _axis.id _axis.type _axis.equipment _axis.depends_on
    _axis.vector[1] _axis.vector[2] _axis.vector[3]
    _axis.offset[1] _axis.offset[2] _axis.offset[3]
    source    .    source    .    0    0    1    . . .
    gravity   .    gravity   .    0    -1   0    . . .
    pin_x     translation   goniometer   .        -1   0    0    0    0    0
    omega     rotation      goniometer   pin_x    1    0    0    -5   0    0
    pin_y     rotation      goniometer   omega    0    1    0    -10  0    0
    pin_z     rotation      goniometer   pin_y    0    0    -1   -15  0    0
```

Static Metadata Example (cont.)

```
_array_intensities.gain          1.0 #counts/photon

_diffrn_source.source            SYNCHROTRON
_diffrn_source.type              'SYNC XXX (ID1)'
_diffrn_source.pdbx_synchrotron  SYNC
_diffrn_source.pdbx_synchrotron_beamline 'XXX (ID1)'

_dataset_file_format.file_format  'hdf5_1.8.14 and NXmx 1.4'

_diffrn_radiation.beam_width  7  #micrometres
_diffrn_radiation.beam_height 5  #micrometres
_diffrn_radiation.beam_flux    400000000000 #ph/s in the beam
```

Dynamic Metadata

Many tags for dynamic (non-Dectris-template) additions are already available. For example, the monchrometer, the beam_height, beam_width, beam_flux and sample sequence can all be placed by a beamline or user in a CIF or NeXus file for merging with H5copy into an existing master metadata file. **The existing imgcif and mmcif dictionaries provide possibilities**, and more can be added. The following have been discussed:

sample provenance, sample physical characteristics, sample imagery, protein sequence, detector and sample environments, incl. temperature, sample delivery method, serial crystallography parameters (incl. pump probes), spectroscopy, sample mount, detector ROI. beamline optics, source parameters, e.g. mode, current, collection strategy, scan type, scan mode, beam profile (Gaussian, tophat), monochromator bandpass, beam divergences, beam collimation.

NXmx Gold Standard Proposal

These are the proposed changes from the NeXus 2016 NXmx 1.4. Under the Gold Standard items may be optional, recommended or required.

(entry): NXentry (required)

title: (optional) NX_CHAR

start_time: (recommended) NX_DATE_TIME **was optional**

end_time: (recommended) NX_DATE_TIME **was optional**

definition: (required) NX_CHAR

NeXus NXDL schema to which this file conforms

Obligatory value: NXmx

(data): (recommended) NXdata **was required**

data[np, i, j, k]: (recommended) NX_NUMBER **was required**

(sample): (required) NXsample **was under /entry/instrument**

name: (recommended) NX_CHAR **was optional**

depends_on: (recommended) NX_CHAR **was optional**

The axis on which the sample position depends may be stored anywhere, but is normally stored in the transformations:NXtransformations group within the NXsample group.

temperature: (optional) NX_CHAR {units=NX_TEMPERATURE}

(transformations): (recommended) Nxtransformations **was optional**

NXmx Gold Standard Proposal

(instrument): (required) Nxinstrument

name: (required) NX_CHAR

Name of instrument, i.e. beamline name

@short_name: (required) NX_CHAR

(attenuator): (optional) Nxattenuator

attenuator_transmission: (optional) NX_NUMBER {units=NX_UNITLESS}

(detector_group): (recommended) NXdetector_group **was optional**

Optional logical grouping of detector elements.

group_index[i]: (required) NX_INT

An array of unique indices for detector elements or groupings of detector elements.

group_names: (required) NX_CHAR

group_parent[group_index]: (required) NX_INT

An array of the hierarchical levels of the parents of detector elements or

A top-level element or grouping has parent level -1

NXmx Gold Standard Proposal

(detector): (required)

Normally the detector group will have the name detector. However, in the case of multiple detector elements, each element needs a uniquely named NXdetector group.

depends_on: (recommended) NX_CHAR **was required**

data[np, i, j, k]: (recommended) NX_NUMBER **was required**

description: (recommended) NX_CHAR

time_per_channel: (optional) NX_CHAR {units=NX_TIME}

distance: (recommended) NX_FLOAT {units=NX_LENGTH} **was optional**

dead_time: (optional) NX_FLOAT {units=NX_TIME}

count_time: (recommended) NX_NUMBER {units=NX_TIME} **was optional**

distance: (recommended) NX_FLOAT {units=NX_LENGTH} **was optional**

beam_center_x: (recommended) NX_FLOAT {units=NX_LENGTH} **was optional**

beam_center_y: (recommended) NX_FLOAT {units=NX_LENGTH} **was optional**

angular_calibration_applied: (optional) NX_BOOLEAN

angular_calibration[i, j, k]: (optional) NX_FLOAT

flatfield_applied: (optional) NX_BOOLEAN

flatfield[i, j, k]: (optional) NX_FLOAT

flatfield_error[i, j, k]: (optional) NX_FLOAT

NXmx Gold Standard Proposal

pixel_mask_applied: (optional) NX_BOOLEAN
pixel_mask[i, j]: (recommended) NX_INT **was optional**
countrate_correction_applied: (optional) NX_BOOLEAN
bit_depth_readout: (recommended) NX_INT **was optional**
detector_readout_time: (optional) NX_FLOAT {units=NX_TIME}
frame_time: (optional) NX_FLOAT {units=NX_TIME}
gain_setting: (optional) NX_CHAR
saturation_value: (optional) NX_INT
sensor_material: (recommended) NX_CHAR **was optional**
sensor_thickness: (optional) NX_FLOAT {units=NX_LENGTH}
threshold_energy: (optional) NX_FLOAT {units=NX_ENERGY}
type: (optional) NX_CHAR
(transformations): (recommended) Nxtransformations **was optional**
(collection): (optional) Nxcollection
(detector_module): (required) NXdetector_module
    data_origin: (required) NX_INT
    data_size: (required) NX_INT
    data_stride: (optional) NX_INT

NXmx Gold Standard Proposal

module_offset: (optional) NX_NUMBER {units=NX_LENGTH}
- @transformation_type: (required) NX_CHAR
- Obligatory value: translation
- @vector: (required) NX_CHAR
- @offset: (required) NX_CHAR
- @depends_on: (required) NX_CHAR

fast_pixel_direction: (required) NX_NUMBER {units=NX_LENGTH}
- @transformation_type: (required) NX_CHAR
  - Obligatory value: translation
- @vector: (required) NX_CHAR
- @offset: (required) NX_CHAR
- @depends_on: (required) NX_CHAR

slow_pixel_direction: (required) NX_NUMBER {units=NX_LENGTH}
- @transformation_type: (required) NX_CHAR
  - Obligatory value: translation
- @vector: (required) NX_CHAR
- @offset: (required) NX_CHAR
- @depends_on: (required) NX_CHAR

NXmx Gold Standard Proposal

(beam): (required) NXbeam **was under sample, now under instrument**

incident_wavelength: (required) NX_FLOAT {units=NX_WAVELENGTH}

In the case of a monchromatic beam this is the scalar wavelength.

In the case of a polychromatic beam this is an array of the wavelengths.
incident_wavelength_weight: (optional) NX_FLOAT

In the case of a polychromatic beam this is an array of the relative weights

of the corresponding wavelengths in incident_wavelength.

incident_wavelength_spread: (optional) NX_FLOAT {units=NX_WAVELENGTH}

The wavelength spread FWHM for the corresponding wavelength(s) in

incident_wavelength.

flux: (optional) NX_FLOAT {units=NX_FLUX}

flux incident on beam plane area in photons per second per unit area

total_flux: (required) NX_FLOAT {units=NX_FREQUENCY} **was optional**

flux incident on beam plane in photons per second

incident_beam_size[2]: (recommended) NX_FLOAT {units=NX_LENGTH}

profile: (recommended) NX_CHAR

Gaussian | Airy | top-hat | rectangular

incident_polarisation_stokes[np, 4]: (recommended) NX_CHAR

incident_wavelength_spectrum: (optional) NXdata

NXmx Gold Standard Proposal

(source): (required) NXsource

The neutron or x-ray storage ring/facility.

distance: (optional) NX_FLOAT {units=NX_LENGTH}

Effective distance from sample

name: (required) NX_CHAR

@short_name: (optional) NX_CHAR **was required**

type: (optional) NX_CHAR **was required**

type of radiation source (pick one from the enumerated list and spell exactly)

probe: (optional) NX_CHAR **was required**

type of radiation probe (pick one from the enumerated list and spell exactly)

power: (optional) NX_FLOAT {units=NX_POWER} **was required**

Source power

emittance_x: (optional) NX_FLOAT {units=NX_EMITTANCE} **was required**

Source emittance (nm-rad) in X (horizontal) direction.

emittance_y: (optional) NX_FLOAT {units=NX_EMITTANCE} **was required**

Source emittance (nm-rad) in Y (horizontal) direction.

sigma_x: (optional) NX_FLOAT {units=NX_LENGTH} **was required**

particle beam size in x

NXmx Gold Standard Proposal

sigma_y: (optional) NX_FLOAT {units=NX_LENGTH} **was required**
particle beam size in y
flux: (optional) NX_FLOAT {units=NX_FLUX} **was required**
Source intensity/area (example: s-1 cm-2)
energy: (optional) NX_FLOAT {units=NX_ENERGY} **was required**
Source energy. For storage rings, this would be the particle beam energy.
For X-ray tubes, this would be the excitation voltage.
current: (optional) NX_FLOAT {units=NX_CURRENT} **was required**
voltage: (optional) NX_FLOAT {units=NX_VOLTAGE} **was required**
frequency: (optional) NX_FLOAT {units=NX_FREQUENCY} **was required**
period: (optional) NX_FLOAT {units=NX_PERIOD} **was required**
target_material: (optional) NX_CHAR **was required**
Pulsed source target material
number_of_bunches: (optional) NX_INT **was required**
bunch_length: (optional) NX_FLOAT {units=NX_TIME **was required**
bunch_distance: (optional) NX_FLOAT {units=NX_TIME} **was required**
pulse_width: (optional) NX_FLOAT {units=NX_TIME} **was required**
mode: (optional) NX_CHAR **was required**
source operating mode

NXmx Gold Standard Proposal

top_up: (optional) NX_BOOLEAN **was required**

last_fill: (optional) NX_NUMBER {units=NX_CURRENT} **was required**

@time: (required) NX_DATE_TIME **was required**

date and time of the most recent injection.

notes: (optional) NXnote

was required

bunch_pattern: (optional) NXdata **was required**

title: (required) NX_CHAR **was required**

name of the bunch pattern

pulse_shape: (optional) NXdata **was required**

source pulse shape

geometry: (optional) NXgeometry **was required**

“Engineering” location of source

distribution: (optional) NXdata **was required**

The wavelength or energy distribution of the source

Validate Images

- Especially with new metadata being added, a fast data-driven tool for NeXus/HDF5 image validation is needed.
- The best available tool is cnxvalidate by Mark Koennecke,
  - https://github.com/nexusformat/cnxvalidate (current)
  - https://github.com/HDRMX/cnxvalidate (proposed)
- which is data driven working against
  - https://github.com/nexusformat/definitions (current)
  - https://github.com/HDRMX/definitions (proposed)

Typical call and output are

```
nxvalidate -a NXmx -l ~/definitions -e thau2_25dps_tr0p05_1_master.h5
message="Missing required global file_name attribute"
... sev=error dataPath=/ dataFile=thau2_25dps_tr0p05_1_master.h5
```

Validation Plans

- After the DLS HDRMX meeting in November 2019, the agreed changes will be integrated with the development version of cnxvalidate and submitted to Dectris and NIAC for review.

- If all goes well users should start seeing validated gold standard images in use in early 2020.

Useful Links

http://hdrmx.medsbio.org

http://github.com/HDRMX

- http://github.com/HDRMX/cnxvalidate
- http://github.com/HDRMX/definitions

http://hdrmx.medsbio.org/manual/build/html

https://zenodo.org/record/3385862 -- Small example Eiger 2X 16M data set from Diamond Light Source I04, Graeme Winter

https://zenodo.org/record/3484187 -- Small example Eiger 2X 16M data set from Diamond Light Source I04, Graeme Winter, revised for clean cnxvalidate error report, Graeme Winter, Aaron Brewster, Herbert J. Bernstein

https://zenodo.org/record/3352358 -- 68 image lysozyme dataset recorded on the Jungfrau 16M detector at SwissFEL and formatted as a NeXus file, Aaron Brewster, Meitian Wang

https://zenodo.org/record/3526738 -- 68 image lysozyme dataset recorded on the Jungfrau 16M detector at SwissFEL and formatted as a NeXus file, revised for clean cnxvalidate error report, Aaron Brewster, Meitian Wang, Herbert J. Bernstein

Acknowledgments

- Work done in part at NSLS-II, Brookhaven National Laboratory
- Supported in part by Dectris, Ltd.
- US Department of Energy Offices of Biological and Environmental Research and of Basic Energy Sciences (grant No. DE-AC02-98CH10886; grant No. E-SC0012704);
- U.S. National Institutes of Health (grant No. P41RR012408; grant No. P41GM103473; grant No. P41GM111244; grant No. R01GM117126; grant No. P30GM133893, grant No. R21GM129570)
- Our thanks to Frances C. Bernstein for helpful consultations