Gold Standard for macromolecular crystallography diffraction data

Herbert J. Bernstein

Ronin Institute for Independent Scholarship

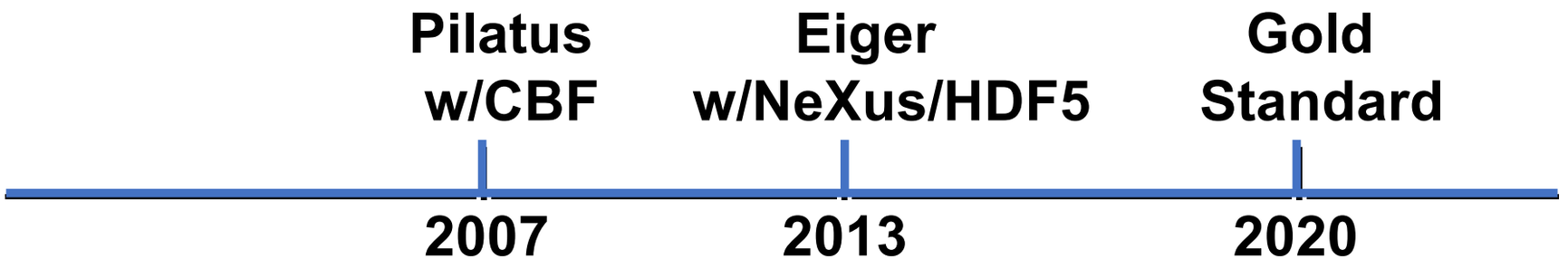
IUCr XXV, Prague, CZ, 14 -- 22 August 2021

Based on Herbert J. Bernstein, Andreas Förster, Asmit Bhowmick, Aaron S. Brewster, Sandor Brockhauser, Luca Gelisio, David R. Hall, Filip Leonarski, Valerio Mariani, Gianluca Santoni, Clemens Vonrhein, Graeme Winter
(2020) “Gold Standard for macromolecular crystallography diffraction data”, IUCrJ 7:5, ISSN: 2052-2525,
https://doi.org/10.1107/S2052252520008672

Work Supported in part by Dectris Ltd, US Department of Energy Offices of Biological and Environmental Research and of Basic Energy Sciences (grant Nos. DE-AC02-98CH10886 and E-SC0012704), National Institutes of Health (grant Nos. P41RR012408, P41GM103473, P41GM111244, R01GM117126, P30GM133893 and R21GM129570) and the Hungarian government (grant No. GINOP 2.2.1-15-2016-00012)

Introduction

HDF5 and CBF are two heavily used formats for collection of data for macromolecular crystallography (MX) at synchrotrons.



DECTRIS Ltd has dominated the market for MX detectors, first with Pilatus detectors using CBF since 2007 and recently with Eiger detectors using NeXus/HDF5 since 2013. The Gold Standard is a common set of metadata for both imgCIF/CBF and NeXus/HDF5 NXmx which helps the community to follow Findability, Accessibility, Interoperability and Reusability (FAIR) principles.

Why Do This

- **Macromolecular crystallography (MX)** is the dominant means of determining the three-dimensional structures of biological macromolecules. Over the last few decades, most MX data have been collected at synchrotron beamlines using a large number of different detectors produced by various manufacturers and taking advantage of various protocols and goniometries. These **data came in their own formats**: sometimes proprietary, sometimes open.
- Efforts to **reuse old data** by other investigators or even by the original investigators some time later were often frustrated.
- **This Gold Standard will facilitate the processing of data sets independent of the facility at which they were collected and enable data archiving according to FAIR principles, with a particular focus on interoperability and reusability.**

The Gold Standard I

- In both CBF files and NeXus/HDF5 files, the information in a Gold Standard data set is the same: **one or more diffraction-image data arrays of pixels along with sufficient metadata to allow software to determine exactly where in the laboratory coordinate system each pixel was located** and when the intensity recorded in that pixel was recorded, so that the software can locate spots, index them and integrate them.
- For example, the conversion of pixel positions relative to the detector to reciprocal-space positions requires knowledge of the **pixel size, the detector distance, the detector orientation, the wavelength and the beam center**.

NOT The Gold Standard

- In the past some of the metadata needed for this process might have been recorded in the same set of files as the image-data arrays and some of the necessary metadata might have been recorded elsewhere, for example in a laboratory notebook or in some separate electronic laboratory notebook

The Gold Standard II

- **In a Gold Standard data set, the necessary data and metadata for processing a reasonable range of use cases is recorded in the data set.** This allows the data set to be moved freely to other filesystems in other facilities and still be processed without the need to return to the original facility to recover information that had been left behind. Although the data set will normally consist of multiple files, these files should be packaged together in an appropriate container, for example a single folder in the file system at the collecting facility or under a single DOI in a data-set repository.

The Gold Standard III

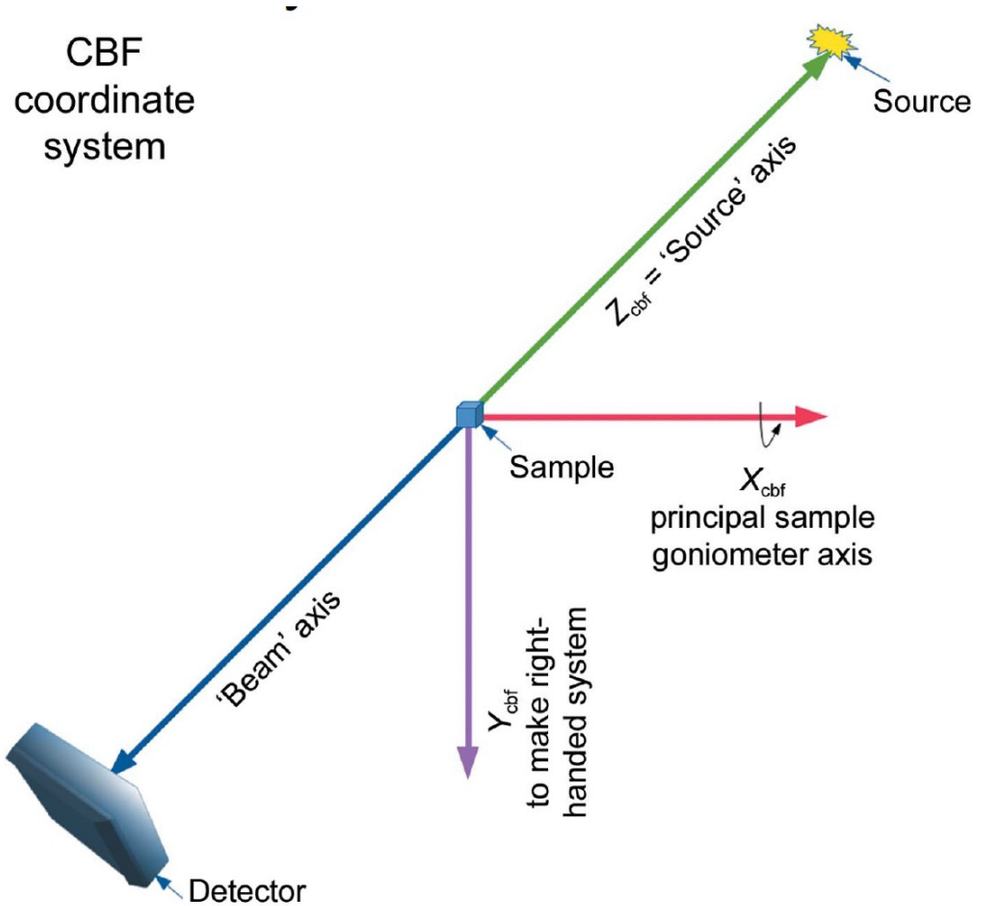
- **The specification of which metadata need be retained with the data depends on the experiment being performed and the software that will be used for processing, *i.e.* the 'use case'. The Gold Standard being discussed here is intended to be adequate for single-axis rotation experiments at synchrotrons and stills collected at XFELs and synchrotrons** and to work properly with the data-reduction programs DIALS (Waterman *et al*., 2013; Winter *et al.*, 2018), XDS (Kabsch, 2010a,b), MOSFLM (Battye *et al*., 2011), HKL-2000 (Otwinowski & Minor, 1997), the data processing pipelines xia2 (Winter, 2010) and autoPROC (Vonrhein *et al*., 2011), as well as future versions of OnDA (Mariani *et al.*, 2016).

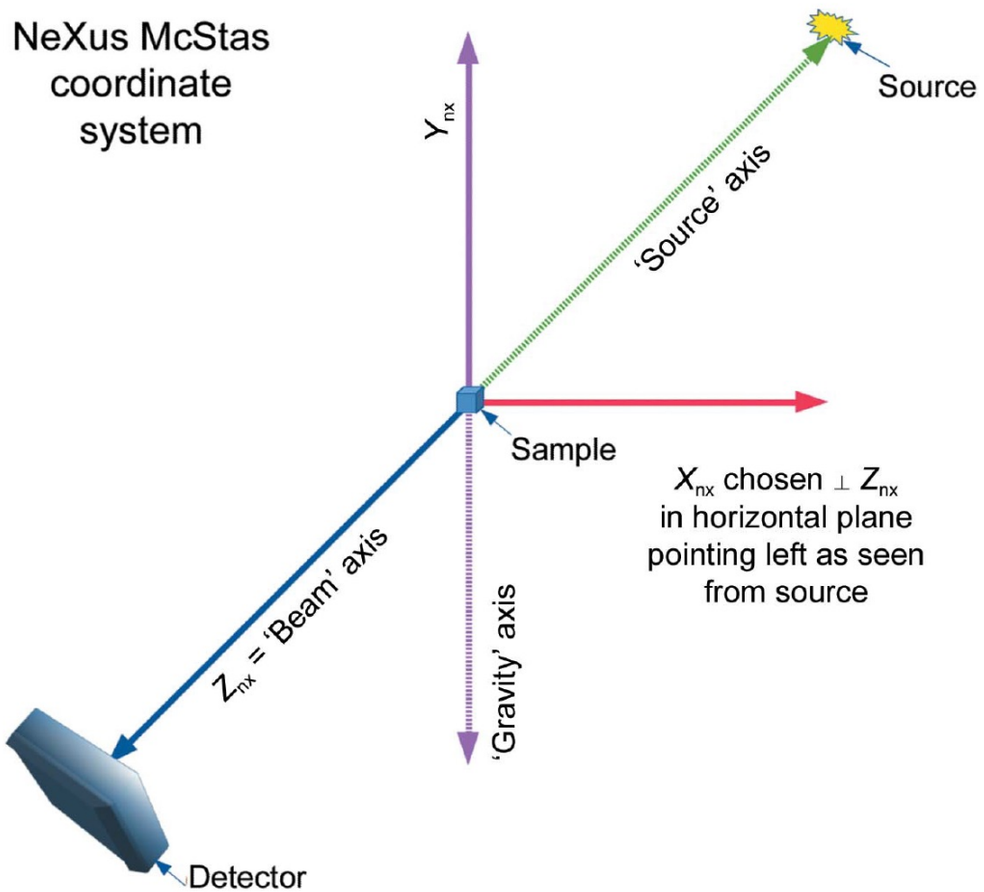
Elements of Gold Standard I

- **imgCIF/CBF vs. NeXus/HDF5**: In 1995, Andrew Hammersley proposed a 'Crystallographic Binary Format' which, after considerable discussion and revision, was adopted by the IUCr in 2005 (Bernstein, 2005; Bernstein & Hammersley, 2005; Ellis & Bernstein, 2005).

- The resulting 'imgCIF/CBF' format, metadata and supporting software was adopted by Dectris for the then-new PILATUS detector in 2007 (Powell *et al*., 2007). In subsequent years it became clear that changes would be needed to this format to support higher data rates and institutional policies (Bernstein, 2010). For the Dectris EIGER detectors, CBF was integrated with the Hierarchical Data Format (HDF5) and became the new NeXus/HDF5 NXmx format (Donath *et al.,* 2013; Könnecke *et al*., 2015; Hester, 2016; Bernstein, 2017). **Everything in the NeXus version of the Gold Standard has an equivalent in imgCIF/CBF**.

CBF and NeXus Coordinate Systems





What about Structure Factors

- This standard is focused on raw diffraction images rather than the structure factors because, in modern MX data collection, diffraction images are the primary raw data and structure factors are derived data.
- Structure factors are very important, and, even if they are derived data, they should of course be recorded, not least because since 2008 they have been mandatory for PDB depositions using the appropriate mmCIF definitions (Jiang *et al*., 1999).
- If structure factors are available, they should be added to Gold Standard files for storage, archiving and deposition. In mmCIF the REFLN category is used. In NeXus/HDF5 the NXreflections category is used.

Where and When I

- While each data set should contain all of the data and metadata necessary for processing, it also should clearly identify where and when it was collected by specifying the scientific instrument or beamline and the facility at which it was collected and the times of collection.

- In the NXmx Gold Standard, the full name of the scientific instrument or beamline is carried in the /(entry):NXentry/(instrument):NXinstrument/name field and the name of the facility is carried in the /(entry):NXentry/(source):NXsource/name field.

Where and When II

- The commonly used acronyms or abbreviations of each of the names in these name fields are carried in the associated @short_name attributes.

- The full and precise UTC ISO 8601 (Wolf & Wicksteed, 1998) time/date of the first data point collected is carried in the /(entry):NXentry/ start_time field and an estimate of the likely time of collection of the last data point is carried in the /(entry):NXentry/end_time_estimated field.

- If/when the data collection is completed, the full and precise UTC ISO8601 time/date of the last data point collected is carried in the /(entry):NXentry/end_time field, provided that it is accurately observed. The time zone of the beamline is carried in the /(entry):NXentry/(instrument):NXinstrument/time_zone field so local times may be recovered.

Experimental Geometry I

- One of the most important sets of metadata used in processing is information on where the components of the experimental setup are positioned and oriented relative to one another. We need to precisely map the events recorded in a pixel to reciprocal space, which implies a need to know or infer the sample orientation, detector position and characteristics, beam wavelength and direction at the very least.
- Essentially, we need a blueprint of the experimental setup. The set of metadata used for this purpose both in CBF and in NeXus/HDF5 describes fixed or variable positioning axes in terms of directional vectors in nested lists with optional offset vectors between pairs of axes. For an experiment with both a detector and a sample goniometer, we need to provide the nested chains of axes that determine the position and orientation of the detector and of the sample.

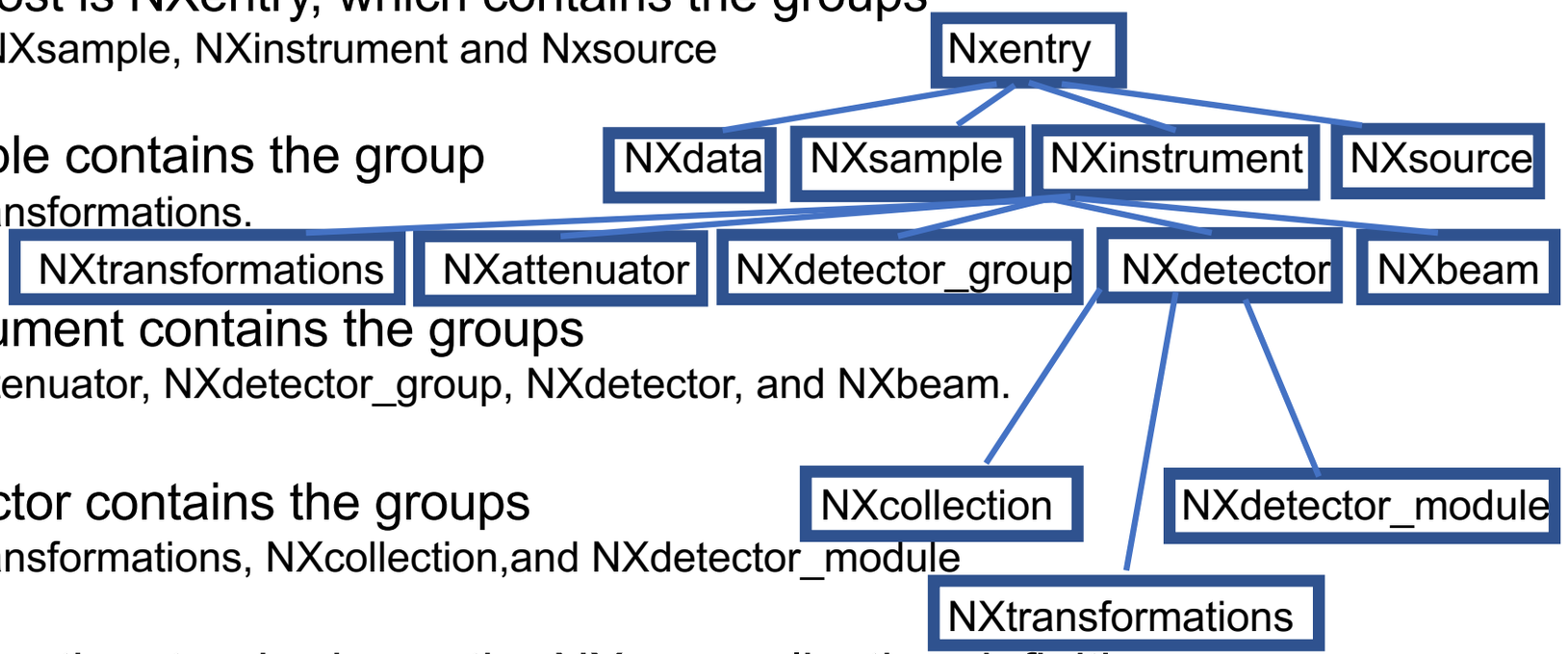
Experimental Geometry II

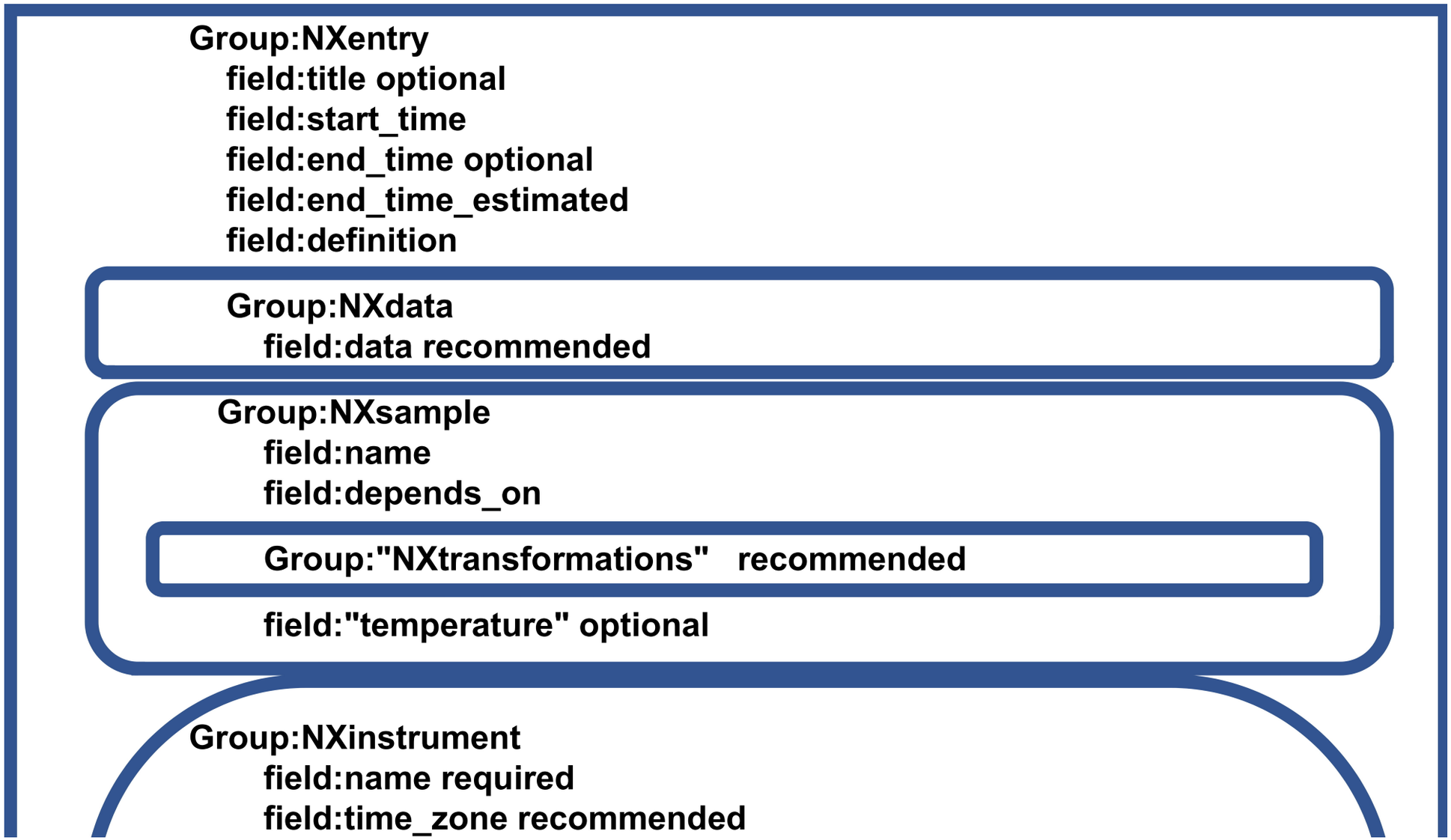
- **Axis chains:** All axis chain definitions and axis settings necessary to process the data should be clearly and explicitly described. There are cases where the values for axis settings available at the time of data collection are only approximate. In such cases, updated or refined values may be added when later calibrations and refinements make them available. Both NeXus and CBF permit the declaration of ‘variants’ to record such cases.

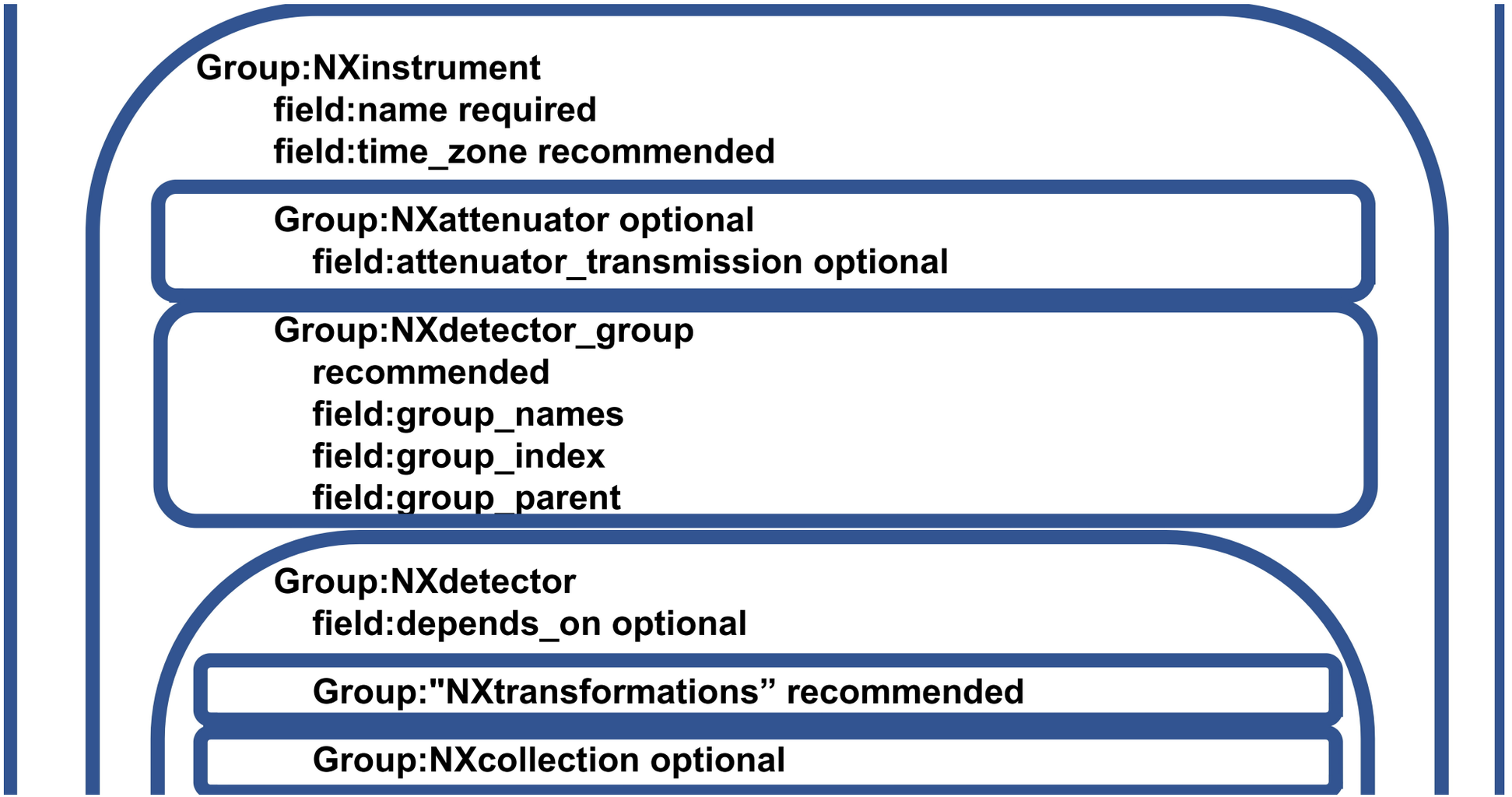
- **Axis names:** The names used for particular axes are arbitrary, provided that they are used in a consistent manner, but it is good practice to use names that enhance rather than detract from understanding. In particular if ‘Beam’ is used as an axis name it should point in the direction going from the source to the sample, and if ‘Source’ is used as an axis name it should point in the direction going from the sample to the source. It is also best never to use the same axis name in two different contexts.

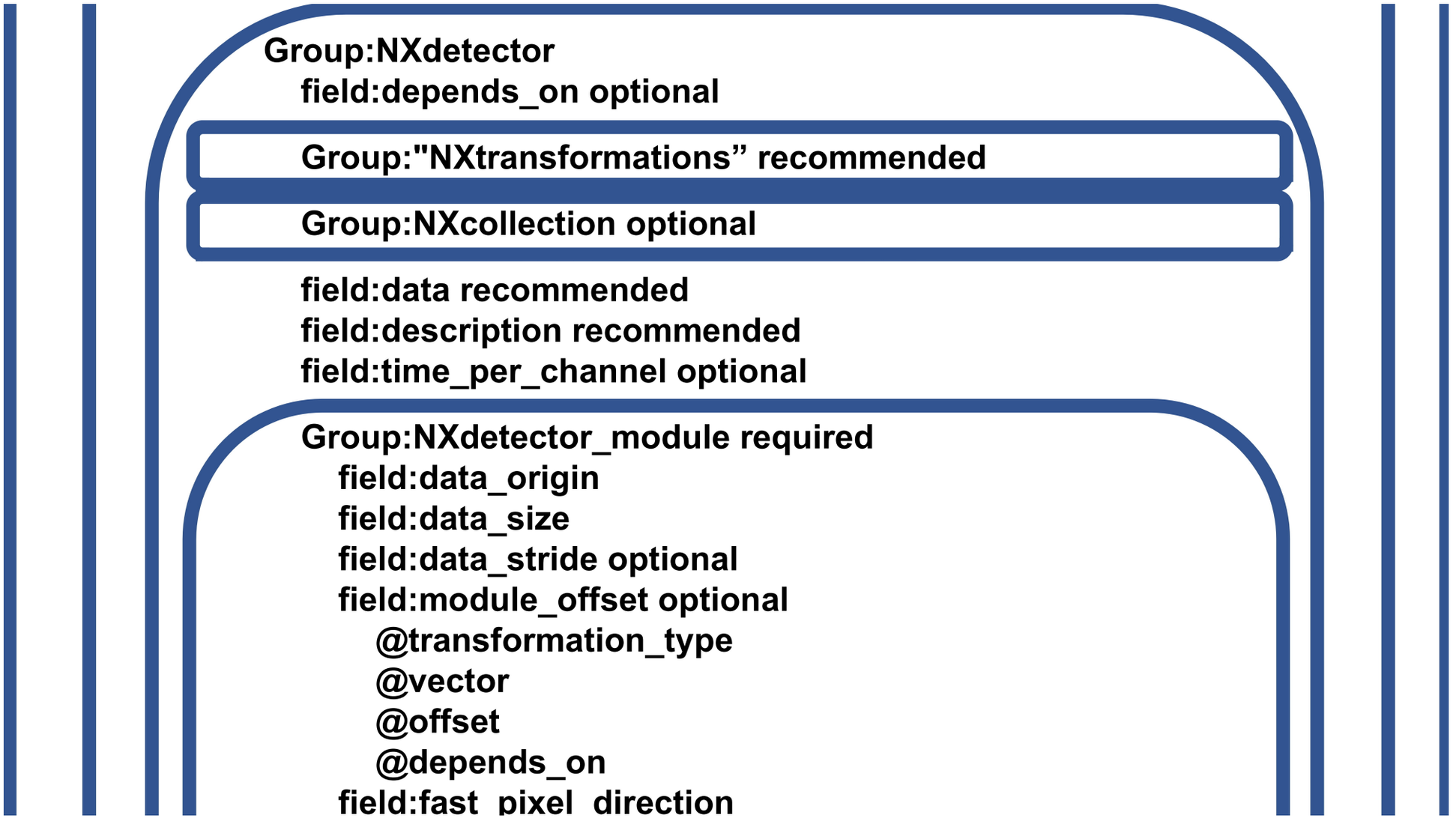
General Organization

- A NeXus/HDF5 Gold Standard file consists of a nested tree of groups.
- The outermost is NXentry, which contains the groups
  - NXdata, NXsample, NXinstrument and Nxsource
  - NXsample contains the group
    - NXtransformations.
  - NXinstrument contains the groups
    - NXattenuator, NXdetector_group, NXdetector, and NXbeam.
  - NXdetector contains the groups
    - NXtransformations, NXcollection,and NXdetector_module
- For details on the standard, see the NXmx application definition



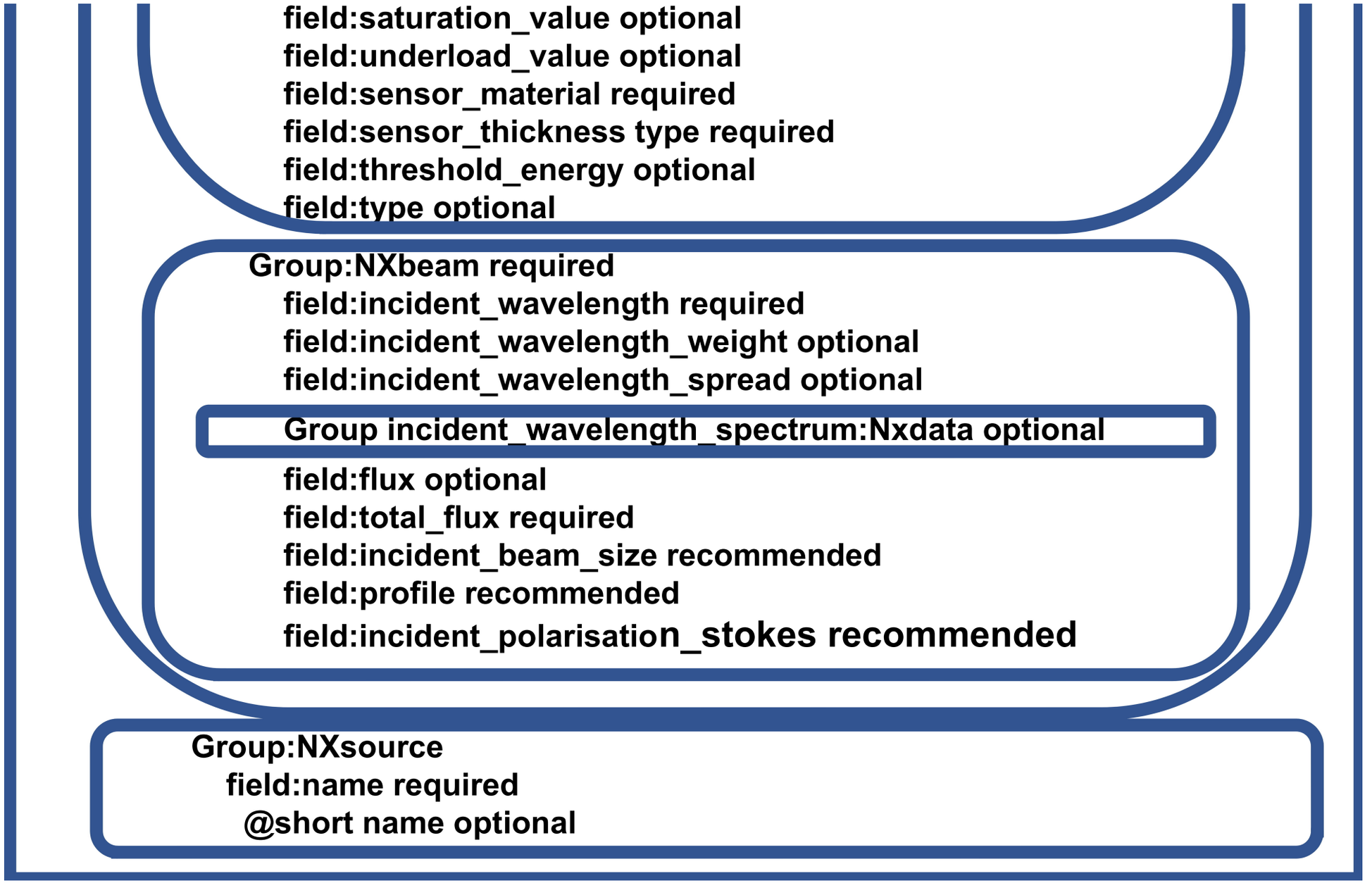

Group:NXentry
field:title optional
field:start_time
field:end_time optional
field:end_time_estimated
field:definition
Group:NXdata
field:data recommended
Group:NXsample
field:name
field:depends_on
Group:"NXtransformations" recommended
field:"temperature" optional
Group:NXinstrument
field:name required
field:time_zone recommended


Group:NXinstrument
field:name required
field:time_zone recommended
Group:NXattenuator optional
field:attenuator_transmission optional
Group:NXdetector_group
recommended
field:group_names
field:group_index
field:group_parent
Group:NXdetector
field:depends_on optional
Group:"NXtransformations" recommended
Group:NXcollection optional


Group:NXdetector
field:depends_on optional
Group:"NXtransformations" recommended
Group:NXcollection optional
field:data recommended
field:description recommended
field:time_per_channel optional
Group:NXdetector_module required
field:data_origin
field:data_size
field:data_stride optional
field:module_offset optional
@transformation_type
@vector
@offset
@depends_on
field:fast_pixel_direction

```
Group:NXdetector_module required
   field:data_origin
   field:data_size
   field:data_stride optional
   field:module_offset optional
      @transformation_type
      @vector
      @offset
      @depends_on
   field:fast_pixel_direction
      @transformation_type
      @vector
      @offset
      @depends_on
   field:slow_pixel_direction
      @transformation_type
      @vector
      @offset
      @depends_on
field:distance
field:distance_derived recommended
```

field:distance
field:distance_derived recommended
field:dead_time optional
field:count_time recommended
field:beam_center_derived optional
field:beam_center_x recommended
field:beam_center_y recommended
field:angular_calibration_applied optional
field:angular_calibration optional
field:flatfield_applied optional
field:flatfield optional
field:flatfield_error optional
field:pixel_mask_applied optional
field:pixel_mask recommended
field:countrate_correction_applied optional
field:bit_depth_readout recommended
field:detector_readout_time optional
field:frame_time optional
field:gain_setting optional


field:saturation_value optional
field:underload_value optional
field:sensor_material required
field:sensor_thickness type required
field:threshold_energy optional
field:type optional
Group:NXbeam required
field:incident_wavelength required
field:incident_wavelength_weight optional
field:incident_wavelength_spread optional
Group incident_wavelength_spectrum:Nxdata optional
field:flux optional
field:total_flux required
field:incident_beam_size recommended
field:profile recommended
field:incident_polarisation_stokes recommended
Group:NXsource
field:name required
@short name optional

Availability

Earlier versions of the NeXus NXmx application definition have been available since 2014 (https://cdn.technologynetworks.com/TN/Resources/PDF/coping-with-big-data-image-formats-integration-of-cbf-nexus-and-hdf5-a-progress-report.pdf). This presentation is based on the newest HDRMX version of the NXmx application definition as it is being proposed to the NeXus International Advisory Committee (NIAC) for adoption by NIAC. That adoption process and discussions in the community are likely to result in additions to the Gold Standard as well as changes. The latest version prior to formal adoption is available from

http://github.com/HDRMX/definitions

- The HDRMX version will be updated as needed to reflect changes during and after adoption.
- To date, the applicability of the Gold Standard has been demonstrated both for single-axis rotation data at a synchrotron

  https://doi.org/10.5281/zenodo.3484187

  and for serial crystallography data at an XFEL

  https://doi.org/10.5281/zenodo.3352357
- The imgCIF/CBF dictionary with NeXus mappings is in the CBFlib kit in

  http://github.com/CBFlib/cbflib

Acknowledgements

The Gold Standard is the result of efforts of many people over many years, including the developers of CIF, mmCIF, STAR, CBF, HDF5, NeXus and many other development efforts. The leadership of the IUCr and NeXus deserve special thanks for their thoughtfulness and community spirit in finding common ground. DECTRIS and the various funding agencies have been generous in their support of the effort.

We also wish to thank Frances C. Bernstein for infinite patience and skill in copy-editing documents and presentations over many, many years

References I

(Battye *et al*., 2011) Battye, T. G. G., Kontogiannis, L., Johnson, O., Powell, H. R. & Leslie, A. G. W. (2011). Acta Cryst. D67, 271 – 281.

(Bernstein, 2005) Bernstein, H. J. (2005). International Tables For Crystallography, Vol. G, edited by S. R. Hall & B. McMahon, pp. 199 – 205. Chester: International Union of Crystallography.

(Bernstein & Hammersley, 2005) Bernstein, H. J. & Hammersley, A. P. (2005). International Tables For Crystallography, Vol. G, edited by S. R. Hall & B. McMahon, pp. 37 – 43. Chester: International Union of Crystallography.

(Bernstein, 2010) Bernstein, H. J. (2010). HDF5 as Hyperspectral Data Analysis Format Workshop, 11 – 13 January 2010, ESRF, Grenoble, France.

(Bernstein, 2017) Bernstein, H. J. (2017). Acta Cryst. A73, a189.

References II

(Donath *et al*., 2013) Donath, T., Rissi, M. & Billich, H. (2013). Synchrotron Radiat. News, 26, 34 – 35.

(Ellis & Bernstein, 2005) Ellis, P. J. & Bernstein, H. J. (2005). International Tables For Crystallography, Vol. G, edited by S. R. Hall & B. McMahon, pp. 544 – 556. Chester: International Union of Crystallography.

(Hester, 2016) Hester, J. (2016). Data Sci. J. 15, 12.

(Jiang *et al*., 1999) Jiang, J., Abola, E. & Sussman, J. L. (1999). Acta Cryst. D55, 4.

(Kabsch, 2010a) Kabsch, W. (2010a). Acta Cryst. D66, 125 – 132.

(Kabsch, 2010b) Kabsch, W. (2010b). Acta Cryst. D66, 133 – 144.

References III

(Könnecke *et al*., 2015) Könnecke , M., Akeroyd, F. A., Bernstein, H. J., Brewster, A. S., Campbell, S. I., Clausen, B., Cottrell, S., Hoffmann, J. U., Jemian, P. R., Männicke, D., Osborn, R., Peterson, P. F., Richter, T., Suzuki, J., Watts, B., Wintersberger, E. & Wuttke, J. (2015). J. Appl. Cryst. 48, 301 – 305.

(Kroon-Batenburg & Helliwell, 2017) Kroon-Batenburg, L. M. J. & Helliwell, J. R. (2014). Acta Cryst. D70, 2502 – 2509.

(Mariani *et al*., 2016) Mariani, V., Morgan, A., Yoon, C. H., Lane, T. J., White, T. A., O'Grady, C., Kuhn, M., Aplin, S., Koglin, J., Barty, A. & Chapman, H. N. (2016). J. Appl. Cryst. 49, 1073 – 1080.

(Otwinowski & Minor, 1997) Otwinowski, Z. & Minor, W. (1997). Methods Enzymol. 276, 307 – 326.

References IV

(Powell *et al*., 2007) Powell, H., Leslie, A. & Battye, G. (2007). CCP4 Newsl. Protein Crystallogr. 46, contribution 1.

(Vonrhein *et al*., 2011) Vonrhein, C., Flensburg, C., Keller, P., Sharff, A., Smart, O., Paciorek, W., Womack, T. & Bricogne, G. (2011). Acta Cryst. D67, 293 – 302.

(Waterman *et al*., 2013) Waterman, D. G., Winter, G., Parkhurst, J. M., Fuentes-Montero, L., Hattne, J., Brewster, A., Sauter, N. K., Evans, G. & Rosenstrom, P. (2013). CCP4 Newsl. Protein Crystallogr. 49, 13–15.

(Winter, 2010) Winter, G. (2010). J. Appl. Cryst. 43, 186–190.

(Winter *et al*., 2018) Winter, G.,Waterman, D. G., Parkhurst, J. M., Brewster, A. S., Gildea, R. J., Gerstel, M., Fuentes-Montero, L., Vollmar, M., Michels-Clark, T., Young, I. D., Sauter, N. K. & Evans, G. (2018). Acta Cryst. D74, 85 – 97.

References V

(Wolf & Wicksteed, 1998) Wolf, M. & Wicksteed, C. (1998). Status for Date and Time Formats. https://www.w3.org/1998/.status/NOTE-datetime-19980827/status.