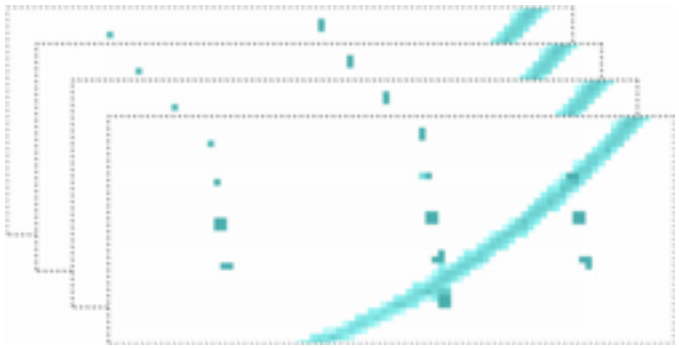


Common diffraction image metadata specification in imgCIF, HDF5 and NeXus

Herbert J. Bernstein

School of Chemistry and Materials Science

Rochester Institute of Technology, Rochester, NY, USA

Work supported in part by Dectris, Ltd, NIGMS and DOE

Work done in part at NSLS-II at BNL

Why Change Things

- New generation of fast pixel-array detectors:
  - DECTRIS Eiger X 16M
  - Cornell-SLAC Pixel Array Detector (CSPAD)
- Need to revisit and extend approaches used in the past to represent
  - the data (the diffraction images) and
  - the metadata (the information needed to reconstruct the experimental environment within which the data were collected)

What Will Change

- The data itself needs to be compressed
- The metadata needs a common ontology for multiple formats
  - CBF/imgCIF (Crystallographic Binary Format)
  - NeXus/HDF5 (format for Neutron, Xray, Muon Science, expressed in Hierarchical Data Format)
  - Compatible with standards of archives such as the PDB
  - Compatible with data processing software

Compression Issues

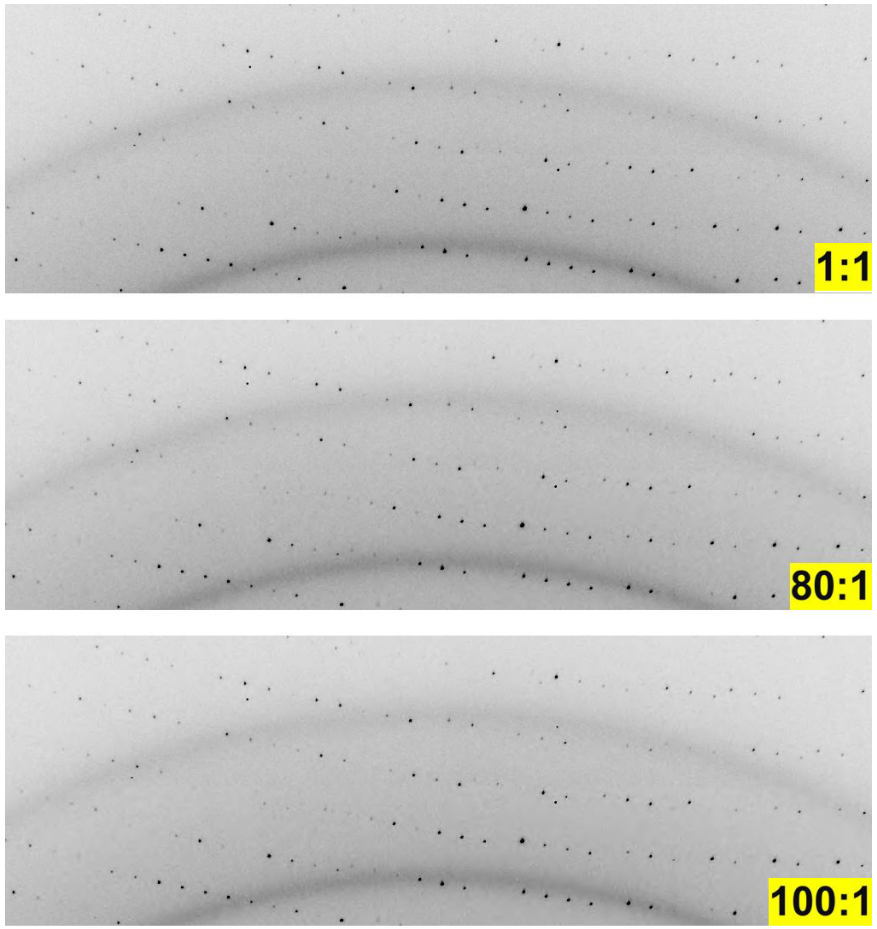
- The specific experiment determines what compression is appropriate – the compression must adapt to the science
- Some data compresses easily (e.g. medium source brightness ultra-fine-sliced MX)
- Some data is difficult to compress (e.g. high source brightness MX in air with ice rings)
- The data format must allow for multiple alternative compressions
- Lossy compression as proposed by James Holton may be needed for some data

Compression Issues (continued)

- Dectris LZ4**2 works well for very sparse data (60:1)
- Drops to 5:1 for moderately dense MX data
- Drops to 2.5:1 for high source brightness dense MX data with air scatter and ice rings
- Byte-offset and nibble-offset give 4:1 and 7:1
- Working on wavelet compression alternative to be able to move smoothly from lossless to lossy compression

Lossy Compression

This figure shows 1/27th of an ADSC Quantum 315 image with a very clean background. The entropy limit on faithful compression of this image is 8.2:1, but the spots are well preserved with compressions of 80:1 to 100:1 at the cost of some noise in the background.

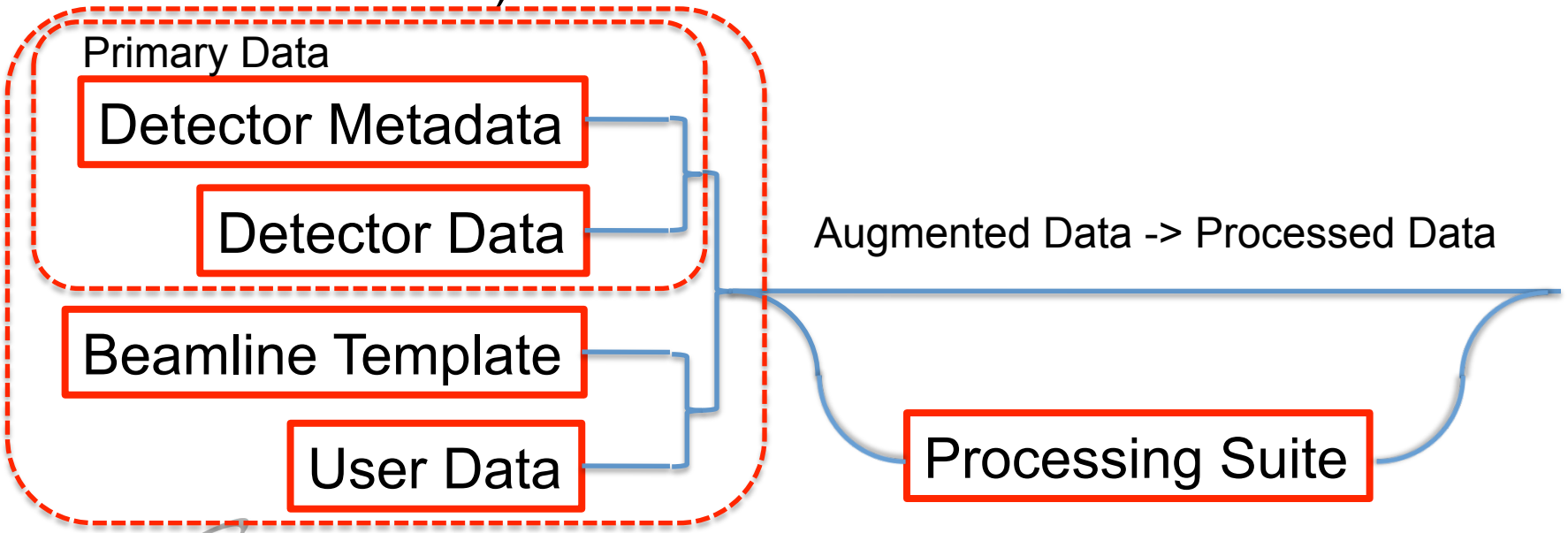


Metadata Issues

- Not all metadata is available from a single source
- Some metadata is known to the detector
- Some metadata is known to the beam line
- Some metadata is known to the user
- Some data or metadata is not known until later
- A “harvest” as proposed by Kim Henrick is needed

Merging Metadata and Data

- **Primary Data:** Merged detector metadata and data in HDF5 format, ready for preliminary spot finding
- **Augmented Data:** Merged primary data, beam line metadata (e.g. HDF5 axis configuration from CBF or HDF5 template) and user data ready for processing
- **Processed Data:** Merged primary data, beam line metadata, user data, and processing results (e.g. MX structure solution)



Metadata Issues (continued)

- HDF5 provides a container into which additional metadata and derived data can be inserted
- Metadata in different formats and with different names is often necessary
- Interoperable consistent definitions and mappings are essential to creating searchable data bases of images and to reliably combining experiments

Metadata Example -- Axes

- Axis descriptions from imgCIF
  - Describe frame-by-frame relative positions of beams, crystals and detectors
  - Mapped to new NeXus/HDF5 NXtransformations to support the new Eiger format, replacing NeXus Nxgeometry, now deprecated
  - Adopted for NeXus NXmx
  - Requires new templating scheme

Dectris Eiger Format Rotation Range

- GONIOMETER_OMEGA = [23,23.5,24]
  - @depends_on="."
  - @equipment="goniometer"
  - @transformation_type= "rotation"
  - @units= "degrees"
  - @vector=[-1, 0, 0]
- GONIOMETER_OMEGA_end = [23.5,24,24.5] ← agreed
- GONIOMETER_OMEGA_range_average=0.5 ← suggested
- GONIOMETER_OMEGA_range_total ← suggested

CBF, NeXus/HDF5 Interaction

- CBF remains CBF, NeXus remains NeXus
- Interoperablity lets either be used when needed
- New dictionaries and extensions to existing dictionaries help in documenting mappings
- Applications gain from extension to the APIs, starting with CBFlib
- HDF5 and NeXus users gain CBFlib compressions
- CBF users gain HDF5 compressions

How to Create Templates for Merging

- Templates can be NeXus/HDF5 or CBF/imgCIF
- There is already a large pool of CBF templates
- HDF5 permits rewrite in place, CBF does not
- cbf2nexus in CBFlib kit can convert template to HDF5

Portion of Eiger 1M CBF template:

```
loop_
_axis.id _axis.depends_on _axis.equipment
_axis.offset[1]   _axis.offset[2]   _axis.offset[3]   _axis.type
_axis.vector[1]  _axis.vector[2]  _axis.vector[3]
y_pixel_offset GONIOMETER_OMEGA detector -38.6 39.95 0 translation 0 -1 0
x_pixel_offset  y_pixel_offset detector 0 0 0 translation 1 0 0
DETECTOR_Z '.' detector 0 0 0 translation 0 0 -1
GONIOMETER_OMEGA DETECTOR_Z detector 0 0 0 rotation 0 0 -1
GONIOMETER_PHI GONIOMETER_KAPPA goniometer 0 0 0 rotation 1 0 0
GONIOMETER_KAPPA '.' goniometer 0 0 0 rotation 0 -1 0
```

Portion of Template in HDF5

/detector:/NXdetector
  /transformations:NXtransformations
    DATASET "GONIOMETER_OMEGA" = [0]
     @depends_on=
     "/entry/instrument/detector/transformations/DETECTOR_Z”
     @equipment="detector
     @long_name="GONIOMETER_OMEGA
     @transformation_type="rotation"
     @units="degrees
     @vector=[0, 0, 1]

The Basics of CBF/imgCIF

There are multiple types of CIF

- DDL1 CIFs (e.g. coreCIF, pdCIF)
- DDL2 CIFs (e.g. mmCIF, imgCIF)
- DDLm is here – should handle all of the above

CIF Dictionaries define the terms that can be used and their relationships

Users can add terms of their own, but your should not use an existing term with a meaning that conflicts with the meaning in a dictionary or in a way that could be confused with terms that have been officially adopted

The Basics of CBF/imgCIF (ctd)

For all CIFs:

- Information is organized into blocks of data
- Each block of data is managed in terms of tables
- Tables are called “categories” or “loops”
- The column headings are called tags” or “data names”
- Some tables have only one row of data
  then each tag can be put with its value
- Some tables have multiple rows of data
- A given tag can appear only once in a block

DDL1 CIFs treat all categories similarly

DDL2 CIFs explicitly state relationships

imgCIF is a layer on top of DDL2 mmCIF dictionary

HDF5 and NeXus

The Hierarchical Data Format Version 5 (HDF5) is a self-describing file format with a robust, well documented API capable of handling (and routinely used to handle) multi-gigabye files of data. It has a diverse user community covering a wide range of disciplines and is fully supported.

NeXus is a tree-oriented view of HDF5 (and XML and HDF4) of importance in managing neutron and x-ray data. NeXus is a convenient thin layer over HDF5 that is widely used at many physics research centers, including at synchrotrons. Together they provide a portable, extensible and efficient format for the storage and management of data.

More Information

CIF, CBF, imgCIF: See IUCR International Tables Volume G and http://www.iucr.org/resources/cif

NeXus/HDF5: http://www.nexusformat.org/

H. J. Bernstein, J. M. Sloan, G. Winter, T. S. Richter, NeXus International Advisory Committee, Committee on the Maintenance of the CIF Standard, “Coping with BIG DATA image formats: integration of CBF, NeXus and HDF5”, Computational Crystallography Newsletter (2014) 5,12 – 18, http://www.phenix-online.org/ newsletter/CCN_2014_01.pdf#page=12.

A. Förster, “EIGER HDF5 data and NeXus format”, Workshop on Metadata for raw data from X-ray diffraction and other structural techniques, 22-23 August 2015, Rovinj, Coatia, satellite of ECM 29

Acknowledgements

DECTRIS Ltd

NIH/NIGMS

DOE Office of Science

IUCR Committee on the Maintenance of the CIF Standard

NeXus International Advisory Committee

BNL Life Science Biomedical Technology Research (LSBR)

Kaden Badalian, Frances C. Bernstein, Aaron Brewster, Jean Jakoncic, Robert M. Sweet, Graeme Winter, Tobias Richter