



## Progress in adoption of imgCIF and integration with NeXus

Herbert J. Bernstein, Dowling College, Oakdale, NY USA  
Work supported in part by grants from the US Department of Energy (ER64212-1027708-0011962, ER63601-1021466-0009501), the US National Science Foundation, the US National Institutes of Health and the International Union of Crystallography

## Introduction

### What is imgCIF/CBF? (see ITVG)

1. A clearly defined set of terms to use in describing raw diffraction images and the way in which they were collected; and
2. A workable and efficient format in which to record, archive and transmit this information; and
3. Support software (e.g. CBFlib)

### What is NeXus? [Klosowski et al. 1998].

"NeXus is a data format for the exchange of neutron and synchrotron scattering data between facilities and user institutions. It has been developed by an international team of scientists and computer programmers from neutron and X-ray facilities around the world. The NeXus format uses the hierarchical data format (HDF) that is portable, binary, extensible and self-describing. ... "

22 May 2008      Bernstein: Progress in imgCIF adoption and NeXus integration      2

## imgCIF Status

**Dictionary is fairly complete, but ...**  
Need to settle what is really needed  
CIF is changing (DDLm is coming)

**Software (CBFlib) is accepted and used**  
C-library API  
Used in fit2d, mosflm, adxv, rasmol  
XDS uses miniCBF-oriented F95 code  
CN created ADSC jiffies  
NS adapted to LABELIT  
Open source, available on sourceforge  
<http://www.sourceforge.net/cbflib>  
20-40 downloads per release

Increasing detector vendor interest (Thanks to SLS and DLS)

22 May 2008      Bernstein: Progress in imgCIF adoption and NeXus integration      3

## Pending Issues?

1. Changes in CIF
  - DDLm
  - Interaction with other dictionaries
2. What is a "correct" CBF?
  - What is the necessary minimum for a CBF?
  - What should not be included?
  - Who provides the information?
3. Jiffies (mapping to and from vendor formats)
4. Integration with NeXus, HDF, XML
5. Microscopy data

22 May 2008      Bernstein: Progress in imgCIF adoption and NeXus integration      4

## Changes in CIF and interaction with other dictionaries

CIF moving to DDLm  
Adds methods to CIF  
Will allow better validation  
Support will be added to CBFlib over the next year

Need to align with PDB dictionaries  
PDBx is not mmCIF  
Changes rapidly

Some progress on SAXS  
Dictionary updated for proposed common axis definitions

22 May 2008      Bernstein: Progress in imgCIF adoption and NeXus integration      5

## What information is needed or too much

No one answer will satisfy everybody  
CIF needs to have a place to put everything  
raw and derived data  
annotation, versions

Not having information at a given time  
should not prevent processing

Merging datasets from multiple experiments requires more notational consistency

Archives and data mining demand thorough annotation

22 May 2008      Bernstein: Progress in imgCIF adoption and NeXus integration      6

## miniCBFs

**Issues**

- Not all information available at time of data collection
- Time constraints may limit complexity of what can be written
- Processing programs may need a minimal header

**Issues came to a head in setting up for SLS detector**

**Proposed solution**

- Write miniCBF at collection
- Convert to full imgCIF with convert\_minicbf program

**Templates are essential**

22 May 2008      Bernstein: Progress in imgCIF adoption and NeXus integration      7

## Where Should imgCIF, NeXus and mmCIF fit?

imgCIF dictionary extends mmCIF to manage data images

NeXus to provide a data framework within which to manage multiple sets of experimental data for multiple experiments

mmCIF provides the base within which to manage the entire crystallographic experiment

NeXus data management

imgCIF Data exp 1
imgCIF Data exp 2
imgCIF Data exp 3

22 May 2008      Bernstein: Progress in imgCIF adoption and NeXus integration      8

## Where Should imgCIF, NeXus and mmCIF fit?

NeXus is tree-oriented  
CIF is table-oriented

Easy to go from CIF to NeXus (cbf2nx available)  
Hard to go from NeXus to CIF

process similar to database normalization

Suppose we have an image for substance X

Header Describing Image

IMAGE

22 May 2008      Bernstein: Progress in imgCIF adoption and NeXus integration      9

## How to Annotate the Image?

**Simple approach:**

The header is a minimal collection of attributes of the image, sufficient to tie the image to an experiment

Any approach to annotation is equivalent to any other:

```

vendor tag=value
CIF _tag value
XML <tag>value</tag>
...
    
```

Header Describing Image

IMAGE

22 May 2008      Bernstein: Progress in imgCIF adoption and NeXus integration      10

## How to Annotate the Image? (cont)

**More complex approach**

- Multiple images for substance X
- Multiple images for multiple substances X, Y, Z ...
- Need ways to organize and relate multiple images and headers
- Not clear if an experiment/substance is an attribute of an image or if each image is an attribute of an experiment/substance

Header  
IMAGE

Header  
IMAGE

Header  
IMAGE

Header  
IMAGE

Header  
IMAGE

Header  
IMAGE

22 May 2008      Bernstein: Progress in imgCIF adoption and NeXus integration      11

## Jiffies

**Small programs to convert to/from imgCIF**  
**Recommended by Hawaii imgCIF workshop in 2006**

A way to deal with SLS miniCBFs  
C. Nielsen's ADSC jiffies released  
Some MAR to CBF available

More needed. Collaborations welcome.

22 May 2008      Bernstein: Progress in imgCIF adoption and NeXus integration      12

## Software status

CBFlib (<http://www.sourceforge.net/projects/cbflib>) provides

Now up to CBFlib 0.9.1

API (C function library, under GPL or LGPL, your choice)  
some fortran support, J. Wright's Python bindings

Manual and sample files

Utilities (under GPL only)

convert\_image (works for Mar)  
convert\_miniCBF (e.g. for SLS images)  
new ADSC jiffies  
cif2cbf  
vcif2

mosflm (<http://www.mrc-lmb.cam.ac.uk/harry/mosflm/>) supports imgCIF

adxv (<http://www.scripps.edu/~arvai/adxv.html>) supports imgCIF

22 May 2008

Bernstein: Progress in imgCIF adoption and NeXus integration

13

## The Basics of imgCIF

### There are multiple types of CIF

DDL1 CIFs (e.g. coreCIF, pdCIF)  
DDL2 CIFs (e.g. mmCIF, imgCIF)  
DDLm is coming

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

Users can add terms of their own

Do 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.

22 May 2008

Bernstein: Progress in imgCIF adoption and NeXus integration

14

### For all CIFs:

Information is organized into blocks of data  
Each block of data is managed essentially 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  
e.g. parent-child relationships

**imgCIF is a DDL2 dictionary that extends the macromolecular CIF (mmCIF) dictionary.**

22 May 2008

Bernstein: Progress in imgCIF adoption and NeXus integration

15

## imgCIF Categories

### ARRAY\_DATA

presents the actual numeric data  
(e.g. the numeric values of the pixels in an image)

### ARRAY\_INTENSITIES

tells you what you need to do to recover intensities from ARRAY\_DATA values

### ARRAY\_STRUCTURE

how the bits and bytes are organized

### ARRAY\_STRUCTURE\_LIST

how the array dimensions are organized

### ARRAY\_STRUCTURE\_LIST\_AXIS

how axis settings relate to array indices

### AXIS

the physical parameters of each axis

22 May 2008

Bernstein: Progress in imgCIF adoption and NeXus integration

16

## imgCIF Categories (cont.)

### DIFFRN

mmCIF category describing diffraction data

#### DIFFRN\_DATA\_FRAME

details about each frame of data

#### DIFFRN\_DETECTOR

information about each detector

#### DIFFRN\_DETECTOR\_AXIS

information about each detector axis

#### DIFFRN\_DETECTOR\_ELEMENT

layout of detector elements

#### DIFFRN\_MEASUREMENT

goniometer information

#### DIFFRN\_MEASUREMENT\_AXIS

information about each goniometer axis

22 May 2008

Bernstein: Progress in imgCIF adoption and NeXus integration

17

## imgCIF Categories (cont.)

### DIFFRN\_RADIATION

incident radiation (crossfire, polarization, etc.)

### DIFFRN\_REFLN

reflection-by-reflection parameters for each frame

### DIFFRN\_SCAN

relationship of axis settings to scans

### DIFFRN\_SCAN\_FRAME

relationship of particular frames to scans

### DIFFRN\_SCAN\_FRAME\_AXIS

relationship of axis settings to particular frames

22 May 2008

Bernstein: Progress in imgCIF adoption and NeXus integration

18

## imgCIF Categories (cont.)

### Categories under development

#### MAP

density maps and masks

#### MAP\_SEGMENT

bricks, slices and other segments of maps

Similar categories for compressed binary arrays are being considered.

22 May 2008

Bernstein: Progress in imgCIF adoption and NeXus integration

19

## CIF Syntax

A collection of data blocks

Each data block contains data names (tags) and their values

White space delimits tokens

Tags start with a leading underscore ("\_") to distinguish them from values

Values that might be confused with data names or keywords or that contain whitespace are quoted

#### Quoting

single quote (single line only)

double quote (single line only)

semicolon in column 1 (multiple lines OK)

terminal quote mark must be followed by whitespace

22 May 2008

Bernstein: Progress in imgCIF adoption and NeXus integration

20

### Characters with special meaning

Underscore

Quote marks

Period (".") or question mark ("?") (null value)

Hash mark ("#") (comment)

#### Reserved words

"global\_", "data\_", "loop\_", "stop\_", and "save\_"

In addition to the underscore, and the three quote marks, three other characters have special meaning: the period ("."), the question mark ("?") and the hash mark ("#"). The period is used when no value is specified. The question mark is used when a value is desired but not available. The hash mark indicates that the remaining characters on that line are part of a comment.

There are a small number of reserved words:

"global\_", "data\_", "loop\_", "stop\_", and "save\_"

The last two reserved words are not used by CIF but are reserved to prevent conflict with the language from which CIF is derived (STAR). "global\_" and "data\_" mark the start of a data block.

22 May 2008

Bernstein: Progress in imgCIF adoption and NeXus integration

21

"data\_" should be followed immediately with the name of the block, without intervening whitespace.

If "loop\_" appears, it is followed by a sequence of tags without intervening data values. Those tags are considered as the column headings of a table. These are followed by rows of data values corresponding to those column headings.

Outside of a table, tags and data values appear in simple alternation. Within a data block a given tag may appear only once.

The meaning of a CIF document is not altered by changing the order of presentation of data blocks nor is it altered by changing the order of presentation of tags within a block.

There are two styles of CIF in use for crystallography: DDL1 and DDL2.

22 May 2008

Bernstein: Progress in imgCIF adoption and NeXus integration

22

## DDL1 CIF (e.g. coreCIF, pdCIF)

Partial example of a small molecule coordinate list [Longridge 98]

```
loop_
_atom_site_label
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_U_iso_or_equiv
_atom_site_adp_type
_atom_site_calc_flag
_atom_site_refinement_flags
_atom_site_occupancy
_atom_site_disorder_assembly
_atom_site_disorder_group
_atom_site_type_symbol
Fe1 1 0 1 .0084(2) Uani d S 1 . . Fe
Na1 .50907(11) .13980(8) 1.09450(9) .0185(3) Uani d . 1 . . Na
Na2 .89904(10) .37128(8) 1.21657(9) .0171(3) Uani d . 1 . . Na
C1 .7997(2) -.01740(18) 1.0419(2) .0110(4) Uani d . 1 . . C
N1 .6788(2) -.02885(18) 1.0696(2) .0166(4) Uani d . 1 . . N
C2 .9306(3) -.01004(16) .8075(3) .0130(4) Uani d . 1 . . C
```

22 May 2008

Bernstein: Progress in imgCIF adoption and NeXus integration

23

## DDL2 CIF (e.g. mmCIF, imgCIF)

Partial example of a macromolecular CIF (1CRN) as converted to mmCIF by the program pdb2cif [Bernstein et al. 98]

```
loop_
_atom_site.label_seq_id
_atom_site.group_PDB
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.auth_seq_id
_atom_site.label_alt_id
_atom_site.cartn_x
_atom_site.cartn_y
_atom_site.cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.footnote_id
_atom_site.label_entity_id
_atom_site.id
1
ATOM N N THR * 1 . 17.047 14.099 3.625 1.00 13.79 . 1 1
1
ATOM C CA THR * 1 . 16.967 12.784 4.338 1.00 10.80 . 1 2
```

22 May 2008

Bernstein: Progress in imgCIF adoption and NeXus integration

24

### ImgCIF Binary Data

```

_array_structure.id ARRAY1
_array_structure.encoding_type "signed 32-bit integer"
_array_structure.compression_type packed
_array_structure.byte_order little_endian
_array_data.array_id ARRAY1
_array_data.binary_id 1
_array_data.data
;
--CIF-BINARY-FORMAT-SECTION--
Content-Type: application/octet-stream;
conversions="x-CBF_PACKED"
Content-Transfer-Encoding: BINARY
X-Binary-Size: 3745758
X-Binary-ID: 1
X-Binary-Element-Type: "signed 32-bit integer"
Content-MD5: 1zsJWPfoI2GYI2V+QsXrw==
'ΠP«q «q FA• fj/E• aR~u<~>k2`b j5ß ...
    
```

22 May 2008      Bernstein: Progress in imgCIF adoption and NeXus integration      25

### How to Make Changes to the imgCIF Dictionary

1. Get the best current version of the dictionary from the IUCr
2. Check that what you propose is not already there, or if there is at least an appropriate category
3. To avoid conflicts with others doing the same thing, get a prefix from Brian McMahon ([bm@iucr.org](mailto:bm@iucr.org))
4. If you are going to be sending files to other people, discuss your new definition with them and, please, on the [imgcif-l](mailto:imgcif-l) list
5. If this will remain just a local change, use it in good health
6. If you think this should be added to the main dictionary for community use, please say so on the [imgcif-l](mailto:imgcif-l) list, and, if appropriate, on other lists
7. If there is sentiment to add it to the main imgCIF dictionary, we will post a revised dictionary for comments, and then, if the dictionary working group agrees, forward the dictionary to COMCIFS for adoption

22 May 2008      Bernstein: Progress in imgCIF adoption and NeXus integration      26

### How to Use and Make or Propose Changes to CBFlib

**Use:**

1. Download the package (source or binary)
2. If source, build for your machine
3. If you need help building, contact [yaya@dowling.edu](mailto:yaya@dowling.edu)
4. If you are using the utilities, install them in your favorite location for binaries and use them
5. If you are building an application against the API, install the library in your favorite location and use it

**Changes:**

1. Changes in your own programs that just use the API: Just do it (LGPL)
2. Changes to the API or Program: Do it, but follow the GPL/LGPL rules on changes (making source available, carrying the license forward)

**Credit**

We would appreciate a credit and knowing about changes. Please cite [Bernstein, Ellis 2005] (see below)

22 May 2008      Bernstein: Progress in imgCIF adoption and NeXus integration      27

### Where to Find imgCIF Information

**IUCr Crystallographic Information Framework:**  
 International Tables, Volume G  
<http://www.iucr.org/iucr-top/cif/index.html>  
 official copies of dictionaries and stable releases of software

**Image CIF/Crystallographic Binary File (imgCIF/CBF)**  
<http://www.bernstein-plus-sons.com/software/CBF>  
 development versions of dictionary and software  
<http://www.iucr.org/iucr-top/cif/cbf/imgcif-l>  
<http://scripts.iucr.org/mailman/listinfo/imgcif-l>  
 imgCIF discussion list (please join)

**Management of Experimental Data in Structural Biology (MEDS BIO)**  
<http://www.medsbio.org>  
 A broader perspective (imgCIF, NeXus, ...) concentrating on interfaces  
<http://www.medsbio.org/meetings>  
 information on this workshop and future ones of interest  
<http://scripts.iucr.org/pipermail/medsbio-l/>  
<http://scripts.iucr.org/mailman/listinfo/medsbio-l>  
 MEDS BIO discussion list (please join)

**Protein Data Bank**  
<http://www.pdb.org>  
 Information on dictionaries and file format, BioSync, etc.

22 May 2008      Bernstein: Progress in imgCIF adoption and NeXus integration      28

### References

[Allen et al. 1973] Allen, F. H., Kennard, O., Motherwell, W. D. S., Town, W. G. & Watson, D. G. (1973), "Cambridge Crystallographic Data Centre. ii. Structural Data File", *J. Chem. Doc.* 13, 119 – 123.

[Berman et al. 2000] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000), 'The Protein Data Bank', *Nucleic Acids Research* 28, 235 – 242.

[Bernstein et al. 1977] Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, Jr., E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977), 'The protein data bank: a computer based archival file for macromolecular structures', *J. Mol. Biol.* 112, 535 – 542.

[Bernstein 2005] Bernstein, H. J. (2005) "The Classification of Image Data", chapter 3.7 in *International Tables For Crystallography, Volume G: Definition and Exchange of Crystallographic Data*, S. R. Hall and B. McMahon, eds. International Union of Crystallography, Springer, Dordrecht, NL, pp. 199 – 205.

22 May 2008      Bernstein: Progress in imgCIF adoption and NeXus integration      29

### References (Cont.)

[Bernstein, Bernstein 2005] Bernstein, H. J., Bernstein, F. C. (2005) "Databanks of Macromolecular Structure", chapter 4 in *Database Annotation in Molecular Biology*, A. M. Lesk, ed., Wiley, Chichester, UK, pp. 63 – 79.

[Bernstein, Ellis 2005] Bernstein, H. J., Ellis, P. J. (2005), "CBFlib: An ANSI C Library for Manipulating Image Data", chapter 5.6 in *International Tables For Crystallography, Volume G: Definition and Exchange of Crystallographic Data*, S. R. Hall and B. McMahon, eds., International Union of Crystallography, Springer, Dordrecht, NL, pp. 544 – 556.

[Bernstein, Hammersley 2005] Bernstein, H. J., Hammersley, A. P. (2005) "Specification of the Crystallographic Binary File (CBF/imgCIF)", chapter 2.3 in *International Tables For Crystallography, Volume G: Definition and Exchange of Crystallographic Data*, S. R. Hall and B. McMahon, eds., International Union of Crystallography, Springer, Dordrecht, NL, pp. 37 – 43.

22 May 2008      Bernstein: Progress in imgCIF adoption and NeXus integration      30

### References (Cont.)

- [Bray, Paoli, Sperberg-McQueen 98] Bray, T., Paoli, J., Sperberg, C. M., eds. "Extensible Markup Language (XML)", W3C Recommendation 10-Feb-98, REC-xml-19980210, <http://www.w3.org/TR/1998/REC-xml-19980210>
- [CCP4 1994] COLLABORATIVE COMPUTATIONAL PROJECT, NUMBER 4, 1994. "The CCP4 Suite: Programs for Protein Crystallography". *Acta Cryst. D50*, 760-763.
- [Gewirth 2003] Gewirth, D. (2003). "THE HKL MANUAL, A Description of the Programs Denzo, XDisplayF, Scalepack An Oscillation Data Processing Suite for Macromolecular Crystallography", 6th ed. (written with the cooperation of the program authors Zbyszek Otwinowski and Wlodek Minor, revised and updated by Wladyslaw Majewski". [http://www.hkl-xray.com/hkl\\_web1/hkl/manual\\_online.pdf](http://www.hkl-xray.com/hkl_web1/hkl/manual_online.pdf)
- [Hall, Allen, Brown 1991] Hall, S. R., Allen, F. H. & Brown, I. D. (1991), 'The crystallographic information file (CIF): a new standard archive file for crystallography', *Acta Cryst. A47*, 655 – 685.

### References (Cont.)

- [Hall, McMahon 2005] Hall, S. R. & McMahon, B. (2005). *Volume G: Definition and Exchange of Crystallographic Data, International Tables For Crystallography*, Springer: Dordrecht, chapter 1.1. Genesis of the Crystallographic Information File
- [Hammersley, Bernstein, Westbrook 2005] Hammersley, A. P., Bernstein, H. J., Westbrook, J. D. (2005). "Image Dictionary (imgCIF)", chapter 4.6 in "International Tables For Crystallography, Volume G: Definition and Exchange of Crystallographic Data," S. R. Hall and B. McMahon, eds., International Union of Crystallography, Springer, Dordrecht, NL, pp. 444 – 458.
- [Klosowski et al. 1998] Klosowski, P.; Koennecke, M.; Tischler, J. Z.; Osborn, R. (1998). "NeXus: A common format for the exchange of neutron and synchrotron data", *Physica B: Physics of Condensed Matter*, 241, 1-4, pp. 151-153.
- [Otwinski, Minor 1997] Otwinowski, Z., Minor, W. (1997). "Processing of X-ray Diffraction Data Collected in Oscillation Mode", *Methods in Enzymology*, 276: *Macromolecular Crystallography*, part A, pp. 307 – 326.

### References (Cont.)

- [Powell 2001] Powell, H. (2001), "Recent improvements to Mosflm - version 6.11", *CCP4 Newsletter on Protein Crystallography*, 39, 18. See [http://www.ccp4.ac.uk/newsletters/newsletters39/18\\_mosflm.html](http://www.ccp4.ac.uk/newsletters/newsletters39/18_mosflm.html).
- [Westbrook et al. 2003] Westbrook, J., Feng, Z., Chen, L., Yang, H., Berman, H. M. (2003) "The Protein Data Bank and structural genomics," *Nucleic Acids Research*, Vol. 31, No. 1 489-4

### ADDITIONAL READING

- [BIOXHIT 2004] "Bioxhit: biocrystallography (X) on a highly integrated technology platform for European structural genomics," *EU Genomics News*, No. 3, November 2004. See <http://icarus.embl-hamburg.de/bioxhit/index.html>
- [Fitzgerald et al. 2005] Fitzgerald, P. M. D., Westbrook, J. D., Bourne, P. E., McMahon, B., Watenpugh, K. D. (2005). "Macromolecular dictionary (mmCIF)", chapter 4.5 in "International Tables For Crystallography, Volume G: Definition and exchange of crystallographic data," Vol. G, S. R. Hall and B. McMahon, eds., International Union of Crystallography, Heidelberg: Springer, pp. 444 – 458.
- [Szebenyi, Arvai, Ealick, Laluppa, Nielsen, 1997] Szebenyi, D. M. E., Arvai, A., Ealick, S., Laluppa, J. M., Nielsen, C. (1997) "A System for Integrated Collection and Analysis of Crystallographic Diffraction Data", *J. Synchrotron Rad.* 4, 128-135. For advx see <http://www.scripps.edu/~arvai/advx.html>

### People Involved

- Frances C. Bernstein  
imgCIF Workshops:  
Dowling: Herbert J. Bernstein  
BNL: Robert M. Sweet  
ARCiB Lab:  
Dowling College: Herbert J. Bernstein, Isaac Awuah  
Asimah, Darina Boycheva, Georgi Darakev, Nikolay  
Darakev, John Jemilawon, Nan Jia, Georgi Todorov  
SVEVSL Project:  
Dowling College: ARCiB Lab group  
RIT: Paul A. Craig, Jared Carter, Brett Hanson, Scott  
Mottarella, Charlie Westin  
**And many more over the years.**