Discussion Materials for
WK.02 The Management of Synchrotron Image Data:
The imgCIF File System and Beyond
at the 2006 Meeting of the American Crystallographic Association
July 22 to July 27, 2006 in Honolulu, Hawaii

Herbert J. Bernstein, yaya@dowling.edu
Robert M. Sweet, sweet@bnl.gov

Sponsored by DOE under grant ER64212-1027708-0011962, NSF under grant DBI-0610407 and Area Detector Systems Corporation (ADSC)
and run thanks to the help and cooperation of
the ACA, the ACA Continuing Education Committee and the ACA Data, Standards, and Computing Committee

Oahu Room, 8:30 am – 5:00 pm, Saturday, 22 July 2006

Brief refresher on the structure and flexibility of imgCIF and available supporting software and mechanisms for making changes to both

Herbert J. Bernstein

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://arcib.dowling.edu/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 (MEDSBIO)

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

MEDSBIO discussion list (please join)

Protein Data Bank

http://www.pdb.org

Information on dictionaries and file format, BioSync, etc.

Software status

CBFlib (http://arcib.dowling.edu/CBF) provides

- **API (C function library, under GPL or LGPL, your choice)**
- **Manual and sample files**
- **Utilities (under GPL only)**
  - **convert_image (works for Mar or ADSC)**
  - **cif2cbf**
  - **vcif2**

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

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

Frequently Asked Questions

What is imgCIF/CBF?

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)

Do I have to use it?

No, of course not. Do what works best for the science you are doing.

Can I change it?

Yes, please do. We would appreciate:

- New ideas
- New items for the dictionary
- New support software
- Bug fixes and improvement for the existing open source code
- Good ways to translate to and from other presentations

But please don't use existing terms in ways that conflict with their meanings

- Define a new term with a new name instead

How can I change it?

Send email to imcif-l@iucr.org, write code, get your own dictionary prefix

The BIG Frequently Asked Question

Can I make proprietary software using imgCIF and CBFlib?

Yes, the API in CBFlib is available under the LGPL.

If you change CBFlib itself, you must publish the changed source code under the LGPL, but even if you change CBFlib, you do not have to make your program into an open source program.

The Basics of imgCIF

There are multiple types of CIF

- **DDL1 CIFs (e.g. coreCIF, pdCIF)**
- **DDL2 CIFs (e.g. mmCIF, imgCIF)**
- **DDL3 is coming**

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

Users can add terms of their own, but you 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.

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.

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 form 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

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

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

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**

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 a 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.**

"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.

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
```

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
```

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: 1zsJjWPfol2GYl2V+QSXrw==

□□P«q□qE□A•□j Æ•ã□R˜u<˘›k2˙b□5ß …
```

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)**

4. **If you are going to be sending files to other people, discuss your new definition with them and, please, on the 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 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 appropriate, forward the dictionary to COMCIFS for adoption**

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**
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)

READING

[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 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," Vol. G, S. R. Hall and B. McMahon, eds. International Union of Crystallography, Heidelberg: Springer, pp. 199 – 205.

[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," Vol. G, S. R. Hall and B. McMahon, eds., International Union of Crystallography, Heidelberg: Springer, pp. 544 – 556."

[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, 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," Vol. G, S.

R. Hall and B. McMahon, eds., International Union of Crystallography, Heidelberg: Springer, pp. 37 – 43.

[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

[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 Wladek Minor, revised and updated by Wladyslaw Majewski", http://www.hkl-xray.com/hkl_web1/hkl/manual_online.pdf

[Fitzgerald et al. 2005] Fitzgerald, P. M. D., Westbrook, J. D., Bourne, P. E., Mcmahon, B., Watenpaugh, 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.

[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," Vol. G, S. R. Hall and B. McMahon, eds., International Union of Crystallography, Heidelberg: Springer, 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.

[Longridge 98] Longridge, J. J., "Tetrasodium Hexacyanoferrate(II) Decahydrate", Acta Cryst. C54, 1998, CIF-Access paper, IUCR9800028.cif.

[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/newsletter39/18_mosflm.html.

[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 adxv see http://www.scripps.edu/~arvai/adxv.html