Issues for CIF wonks

What is CIF?

(1) Crystallographic Information FRAMEWORK

- domain-specific ontology (machine-readable collection of data identifiers, attributes and relationships)
- tight coupling to crystallographic topics: small-unit-cell structures (inorganic, organic, organometallic); biological macromolecules (proteins, nucleic acids); single-crystal structure determination experiments; powder diffraction; modulated and composite structures; electron density; symmetry; image data
- looser coupling to other structural science topics: biological NMR structures; cryoelectron microscopy; protein production; ligand structures; molecular modelling

(2) Crystallographic Information FILE:

- specific file-based implementation of the crystallographic information framework
- ASCII, flat data structure, simple syntax, well suited to archive and exchange purposes, relatively free format, but manageable with even Fortran I/O capabilities
- some idiosyncracies (*e.g.* curious character and text-field delimiters, awkward line ending problems, constrained line widths)
- easy to transform to XML, relational database schemas

(3) Crystallographic BINARY File:

- isomorphous to crystallographic information file
- contains binary data
- metadata still expressed with ASCII character set and typical CIF syntax for ease of ASCII/binary (CIF/CBF) interconversion

Why CIF?

- because it is there.
- tools
    - Fortran libraries (CIFtbx)
    - C libraries (CBFlib; CIFOBJ/CIFParse etc.)
    - Python toolkit (PyCifRW; mmLib)
    - Perl modules (STAR::Parser 2tc.)
    - support in CCP4, mmLib etc.
- interoperability acros crystallographic/structural informatics domains
- ease of conversion to XML, SQL, PDB
- large existing data store (>37,500 structures in PDB, >366,000 structures in CCDC, >22,600 structures available from IUCr journals)
- well developed integration of 'data' and 'metadata'

Which DDL?

DDL (dictionary definition language) is a machine-readable formalism for specifying and validation attributes of data identified by specific tags (data names).

DDL1 (Hall & Cook, 1995) is modelled loosely on a relational database model, supports very few data types, and allows tags to be looped together "never", "always" or "sometimes".

DDL2 (Westbrook et al., 2006) is isomorphous with a relational database model, provides extended data typing through use of regular expressions, and has rigorous validation of key and foreign key relationships.

A next-generation DDL is under development that will allow for dynamic specification of methods, complex data types and better structured data hierarchies.

imgCIF is written in DDL2 for compatibilty with mmCIF applications.

Standards certification

CIF is a private standard managed internally by the IUCr Committee for the Maintenance of the CIF Standard (COMCIFS), reporting directly to the Executive Committee.

Each dictionary has a dedicated Working Group (prior to initial acceptance) or Management Group (when the dictionary has been accepted by COMCIFS). These groups have autonomy to work within their relevant communities to develop or maintain the dictionaries.

Drafts are submitted to COMCIFS for technical validation and ultimately for approval.

The COMCIFS Chair (I. David Brown) acts as coordinator of different dictionary activities.

Formal mechanisms (namespace register) exist to allow informal or private dictionary development. Private dictionaries can serve many useful purposes:

- during initial development of a new topic area
- to test new areas or ideas
- for use in non-crystallographic communities that do not report to IUCr
- to implement reasonable extensions during the sometimes lengthy COMCIFS review and approval process

Data items developed under private data names can be integrated with "official" dictionaries through an aliasing mechanism. Aliasing also allows interoperability of DDL1- and DDL2-based applications. *E.g.* the small-unit-cell core dictionary item _cell_length_a is aliased to the mmCIF equivalent _cell.length_a.

Interoperability initiatives

- mmCIF has many extra-crystallographic extensions implemented in CIF:
  - 3D EM Extension Dictionary
  - 3D EM Exchange Dictionary
  - BIOSYNC Extension Dictionary
  - MDB Modeling Extension Dictionary
  - CCP4 Harvest Extension Dictionary
- mmCIF also interoperates with
  - NMRSTAR Dictionary (CIF-like format with hierarchical data model)
- mmCIF ontology expressed as
  - XML (PDBML)
  - object libraries/APIs
  - CORBA
- coreCIF
  - Chemical Markup Language (CML)
  - MIF (molecular information file, also in STAR format)
  - developing metadata dictionary for institutional repositories (eBank)
- imgCIF
  - NeXuS, HDF applications
  - possible relevance to image data stores (ISIS, Rutherford Appleton Lab)

Documentation

http://www.iucr.org/iucr-top/cif/

International Tables for Crystallography (2005). *Volume G: Definition and exchange of crystallographic data*, edited by S. R. Hall & B. McMahon. Dordrecht: Springer. (http://it.iucr.org/g/)

Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *The Crystallographic Information File (CIF): a new standard archive file for crystallography. Acta Cryst.* A**47**, 655-685. [historical]

Hall, S. R. & Cook, A. P. F. (1995). *STAR dictionary definition language: initial specification. J. Chem. Inf. Comput. Sci.* **35**, 819-825.

Westbrook, J. D., Berman, H. M. & Hall, S. R. (2005). Specification of a relational dictionary definition language (DDL2). Chapter 2.6 of *International Tables for Crystallography Volume G: Definition and exchange of crystallographic data*, edited by S. R. Hall & B. McMahon. Dordrecht: Springer.