The vital role of Crystallographic Information Files in chemical and biological crystallography to underpin the databases' validation reports

Brian McMahon

International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, UK

Email: bm@iucr.org

The clue is in the name…

“There is a traditional hierarchy of components of understanding: **data**, **information**, **knowledge**, **wisdom** (the DIKW model).

Crystallographic **information** is the component that bridges the gap between the raw experimental **data** and the global **knowledge** bases represented by the Protein Data Bank, Cambridge Structural Database, International Centre for Diffraction Data, Inorganic Crystal Structure Database, Crystallography Open Database etc.”

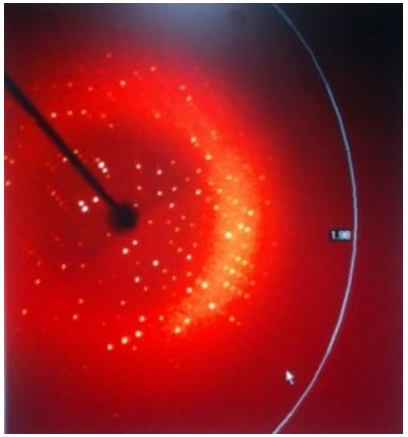
Summary: CIF and the *checkCIF* paradigm

- CIF for small molecules led to the *checkCIF* validation service
- *checkCIF* has a threefold approach:
  - Completeness
  - Correctness
  - Context
- CIF embraces all validation requirements (publication/database)
  - Handles metadata on same footing as data
  - Suitable for data query
  - Extensible

Types of crystallographic data described by CIF

‘Raw’ data

Numerical data collected directly from an experimental apparatus



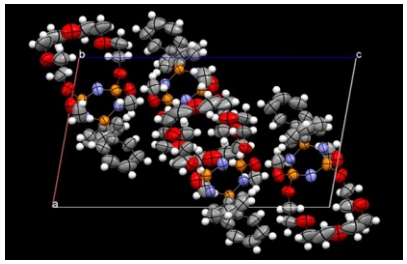
‘Processed’ data

Reduced, calibrated, processed numerical observations

```
# h,k,l, Fo-squared, Fo-squared, sigma(Fo-squared) and status flag
#
data_6
_shelx_title ' 01SRC413 in P2(1)/n'
_shelx_refln_list_code          4
_shelx_F_calc_maximum      183.83
_expt1_crystal_F_000      1144.00
_refIns_d_resolution_high  0.7705

loop_
 _symmetry_equiv_pos_as_xyz
 'x, y, z'
 '-x+1/2, y+1/2, -z+1/2'
 '-x, -y, -z'
 'x-1/2, -y-1/2, z-1/2'

_cell_length_a    11.8293
_cell_length_b    10.3312
_cell_length_c    21.6318
_cell_angle_alpha  90.000
_cell_angle_beta  100.203
_cell_angle_gamma  90.000

_shelx_F_squared_multiplier     1.000

loop_
 _refln_index_h
 _refln_index_k
 _refln_index_l
 _refln_F_squared_calc
 _refln_F_squared_meas
 _refln_F_squared_sigma
 _refln_observed_status
   2   0   0     772.37     856.47     28.20 o
   4   0   0    1445.15    1446.80     39.55 o
   6   0   0    1130.79    1097.08     30.62 o
   8   0   0    1347.13    1490.27     55.41 o
  10   0   0    3273.01    3545.64    154.91 o
  12   0   0      48.20      40.50      4.56 o
  14   0   0      79.87      63.02      7.91 o
   2   1   0    2093.70    1975.83     47.36 o
   3   1   0   33795.10   34884.29   1287.71 o
   4   1   0    2298.16    2035.72     38.24 o
   5   1   0       9.73      36.06      5.59 o
   6   1   0     449.80     506.89     11.92 o
   7   1   0       1.81       7.91      5.59 o
   8   1   0      43.36      28.81      6.79 o
   9   1   0      64.18      48.51      6.02 o
  10   1   0    1412.22    1628.54     45.96 o
  11   1   0     242.68     279.96      9.70 o
  12   1   0      14.96      10.52      3.84 o
  13   1   0      16.87      15.76      4.56 o
  14   1   0      16.46       7.91      7.91 o
  15   1   0       0.00       3.95      5.59 o
   0   2   0    2443.71    2679.14     61.27 o
   1   2   0   23397.80   23770.90    546.30 o
   2   2   0   20572.37   19502.51    520.01 o
   3   2   0    8854.88    8282.53    169.57 o
```

‘Derived’ data

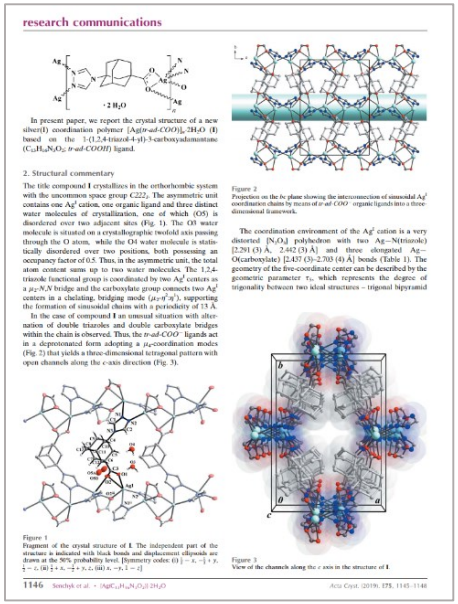
Numerical description of the parameters of a calculated structure model



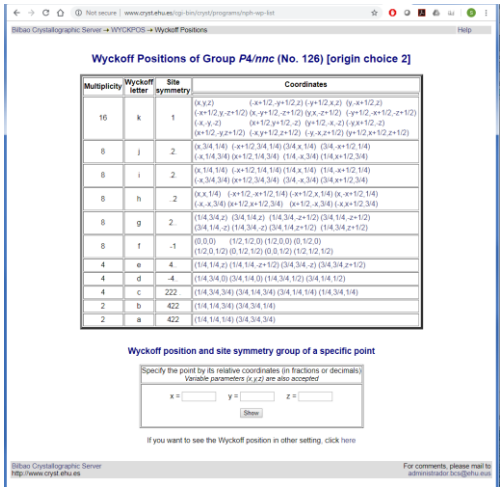
‘Interpretative’ data

Variable parameters in the experimental set-up or numerical modelling and interpretation

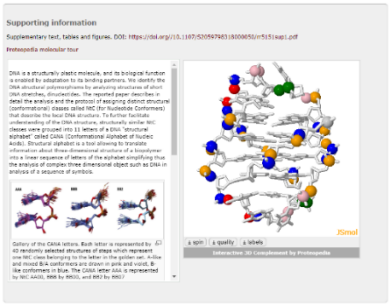
Commentary



‘Reference’ data



Annotation



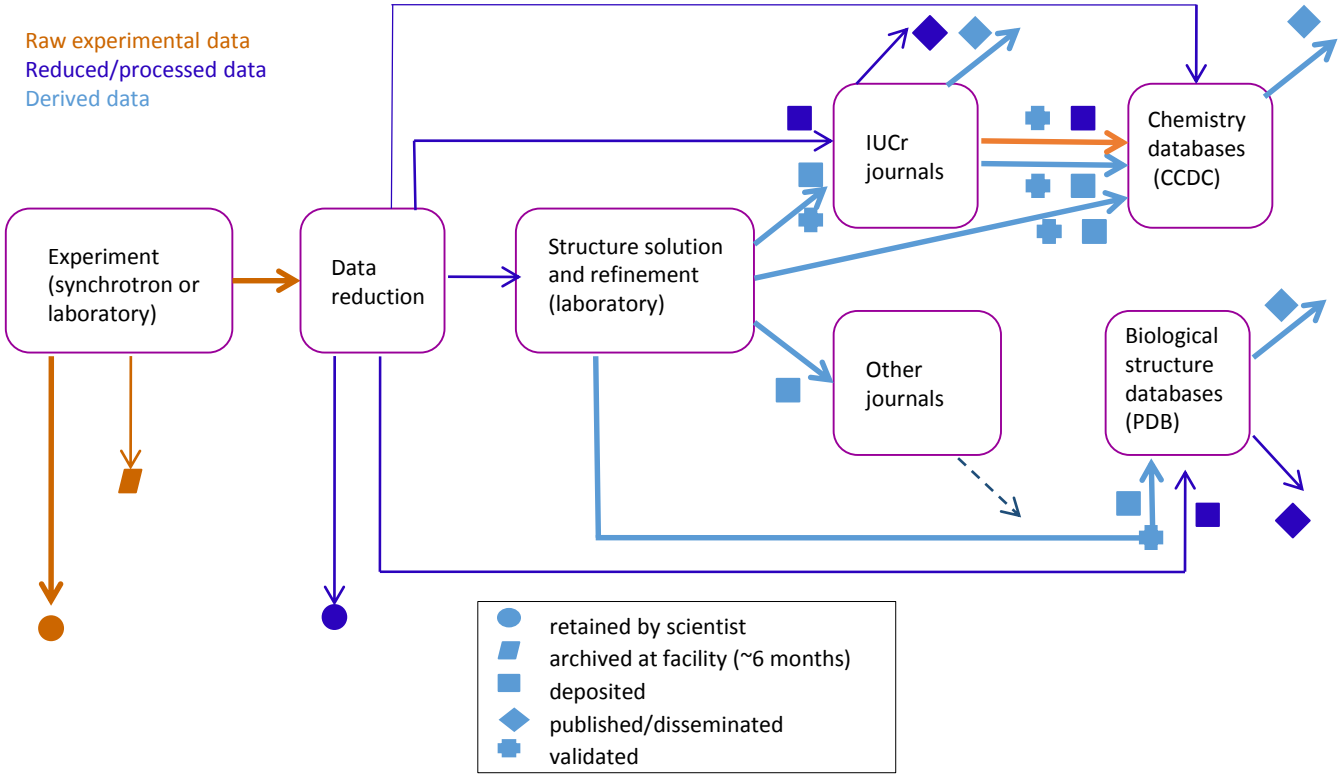
CIF dictionaries (COMCIFS)

- Crystallographic Core (coreCIF) – 1991
- Crystallographic Restraints – 2011
- Crystallographic Powder Diffraction (pdCIF) – 1997
- Modulated and Composite Structures (msCIF) – 2002
- Multipole Electron Density (rhoCIF) – 2003
- Crystallographic Twinning – 2014
- Magnetic Structures (magCIF) – 2016
- Lattice topology (topoCIF) – 2018
- Crystallographic Symmetry (symCIF) – 2001
- Diffraction Images (imgCIF) – 2000
- High pressure – under development
- Crystallographic Macromolecular Structure (mmCIF) – 1997

CIF dictionaries (wwPDB)

- Crystallographic Macromolecular Structure (mmCIF) – 1997
- PDB Exchange Dictionary (PDBx/mmCIF) – 1997 and ongoing
- Integrative/Hybrid (I/H) methods – 2017
- 3DEM Extension Dictionary – 2004
- NMRSTAR Dictionary – 2013
- Biological Small Angle Scattering– 1998
- Model Archive Extension Dictionary – 2018
- BIOSYNC Extension Dictionary – 2000
- NMR Exchange Format Dictionary – 2016

A coherent information flow



Early drivers for CIF

- Requirement to input raw data from many diffractometers
- Exchange data between software packages in the solution/refinement pipeline
- Provide a mechanism for electronic publication of articles
- Help referees to check the structural model
- Improve the accuracy of data capture by databases

Standardisation of raw data input

- Championed by Howard Flack for point detectors

J. Appl. Cryst. (1992). **25**, 455–459

DIFRAC, single-crystal diffractometer output-conversion software. By H. D. Flack, *Laboratoire de Cristallographie, University of Geneva, 24 quai Ernest-Ansermet, CH-1211 Genève 4, Switzerland* and E. Blanc and D. Schwarzenbach, *Institut de Cristallographie, University of Lausanne, BSP Dorigny, CH-1015 Lausanne, Switzerland*

(*Received* 25 *October* 1991; *accepted* 9 *January* 1992)

Abstract

Software is described that converts single-crystal diffractometer output files as produced by maufacturers' software into a standardized instrument-independent form consisting of a clear, complete, well documented record of the sample and the diffraction measurements performed upon it. Information not already available in the manufacturers' diffractometer files is obtained in an interactive question-and-answer session. The software is written in a modular way. Available modules can deal with Enraf–Nonius CAD-4, Philips PW1100 and Siemens $P2_1$ single-crystal diffractometers and produce output in CIF or SCFS format.

Introduction

Users of several models of four-circle single-crystal diffractometers may well have been struck by the diversity of form and content of the data files generated by diffractometer-manufacturers' software. The form of the file can create difficulties in its transfer to other computing equipment and further renders the corresponding computer programs specific to a certain type or types of diffractometer.

*"The difficulties in the content of the files manifest themselves principally by **the paucity of the available information** necessitating additional input to the data-treatment software (e.g. type of radiation, wavelength of radiation, scan width, …). The problem has become aggravated in recent times by the rapid development in electronic data exchange. ... The advent of machine-readable submission for publication and data-base and supplementary-material deposition further highlights **the problem of missing data**."*

Standardisation of raw data input

- Championed by Howard Flack for point detectors
- Relatively unenthusiastic uptake by vendors
- CIFs produced by commercial vendors often faulty
- imgCIF/CBF a second-wave attempt for diffraction images
- Perhaps slightly more effort from vendors
- ‘mini-CBFs’ showed significant diversity
- Reluctance to record essential experimental metadata

Data exchange between software packages

- STAR File modelled loosely on *Xtal* internal data structures
- Particular early adopters were collaborative packages (*NRCVAX*, *CCP*4) or general-purpose analysis programs (*PLATON*)
- Adoption by 'market leader' *SHELXL* crucial for success
- Reluctant uptake of mmCIF because of existing PDB format
- In general, has stood up well to the test of time
  - Generic presentations like XML not popular (but watch JSON!)
  - Limitations of PDB format eventually led to mmCIF as standard
  - Adoption of HDF5/NeXus in image capture driven by throughput needs
  - Relational nature of DDL dictionaries has become more important

Electronic publication

- Adoption by IUCr Journals (1991) very important
- Mandatory data format for *Acta C* by 1996
- *Notes for Authors* stipulate and validate mandatory metadata by listing required CIF data items
- Most journals reporting crystal structures now require CIFs as supporting information
- For online publication, availability of CIF makes the publication interactive

Technical peer review

- IUCr journals always required reviewers to check chemical structures
- CIF allowed ease of input to popular checking programs (*PARST*, *MISSYM*, *UNIMOL*, *CREDUC*)
- This led to semi-automated validation
  - IUCr checking reports
  - checkCIF dedicated tests
  - Consolidation with *PLATON*
- Elimination of ‘Marshing’
- Subsequent ability to validate structure factors and re-do refinement

Improved quality of structures in databases

- Improved quality of small-unit-cell structures from IUCr journals
- Convergence of checking criteria with CCDC
- Adoption of *checkCIF* for CCDC direct depositions
- **wwPDB database structure isomorphic to mmCIF**
- PDB requirement for structure factors in CIF format
- Validation Reports developed alongside extension CIF dictionaries for novel techniques

mmCIF ⟷ RDBMS

- CIF originally developed as tagged file format with one-dimensional loops
- York mmCIF workshop (1993) identified similarities with relational database structure
- DDL2 developed 1994/5 as a fully relational description of CIF data items
- The mmCIF/DDL2 schema is the basis for the relational database schema used by the wwPDB
- Latest iteration of DDLm retains this relational nature

mmCIF ⟷ RDBMS

_computational BIOLOGY

COMMENTARY

Overhauling the PDB

Amanda C Schierz, Larisa N Soldatova & Ross D King

The Brookhaven Protein Data Bank was once a pioneering database, but its organization of structural data is now outdated and in need of an upgrade.

Although structural biology was once a leader in both the development of standards for the preservation and sharing of scientific data and for database development, this lead has been lost. The main data standard used by PDB, mmCIF, does not meet state-of-the-art standards in biology for ontologies; this has serious repercussions for the analysis and sharing of data, and has led to problems in database design. The main database, the reengineered Brookhaven Protein Data Bank (PDB), although intended to be a relational database, does not conform to relational database design standards and principles; this diminishes the quality of information stored in PDB and has serious implications for data-storage capacity. If structural biology is to realize the full potential of its wealth of stored data and regain its lead in data standards and databases, then both mmCIF and PDB need to be extensively reengineered. In this article, we detail some of the more serious faults we have found in the mmCIF and PDB. We then sketch a way to

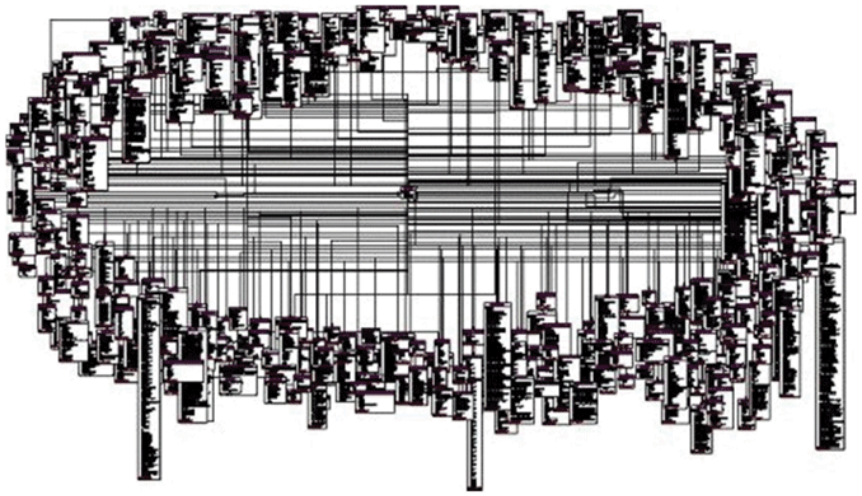
Since the early 1970s, structural biology has been at the forefront of the development of standards for the preservation and sharing of scientific data. The PDB was set up in 1971 (more than 10 years before the first sequence databases, such as EMBL-Bank and GenBank). Structural biology journals were among the first to require submission of data to international databases as a condition of publication, and the mmCIF dictionary[3], published in 1997, was one of the first standards internationally agreed upon for reporting experimental results.

The PDB was recently reengineered to become a 'relational' database based on the mmCIF schema[4]. As wwPDB, it now includes the PDB at the Research Collaboratory for Structural Bioinformatics (RCSB, Brookhaven), the Macromolecular Structure Database (MSD) at the European Bioinformatics Institute (Hinxton Hall, UK), the Protein Data Bank (Osaka University, Japan) and more recently the Biological Magnetic Resonance Data Bank. Our discussion here focuses mainly on the RCSB PDB, as we consider it to have the most problems.

Our analysis indicates that the reengineering process was not a success and has led to problems of data repetition, redundancy, inconsistency and integrity. These problems are important: they can result directly in incorrect answers to queries or more indirectly lead to erroneous results through the incomplete updating of the database due to the formation of uncontrolled redundancies. They also seriously inhibit the future possibility of designing intelligent analysis tools to analyze the data in PDB and aid the process of knowledge discovery.

Why mmCIF is not a true ontology

Several biological ontologies are now in use, including FMA (Foundational Model of Anatomy), which details the key concepts and relations in anatomy[5], and FuGO (The



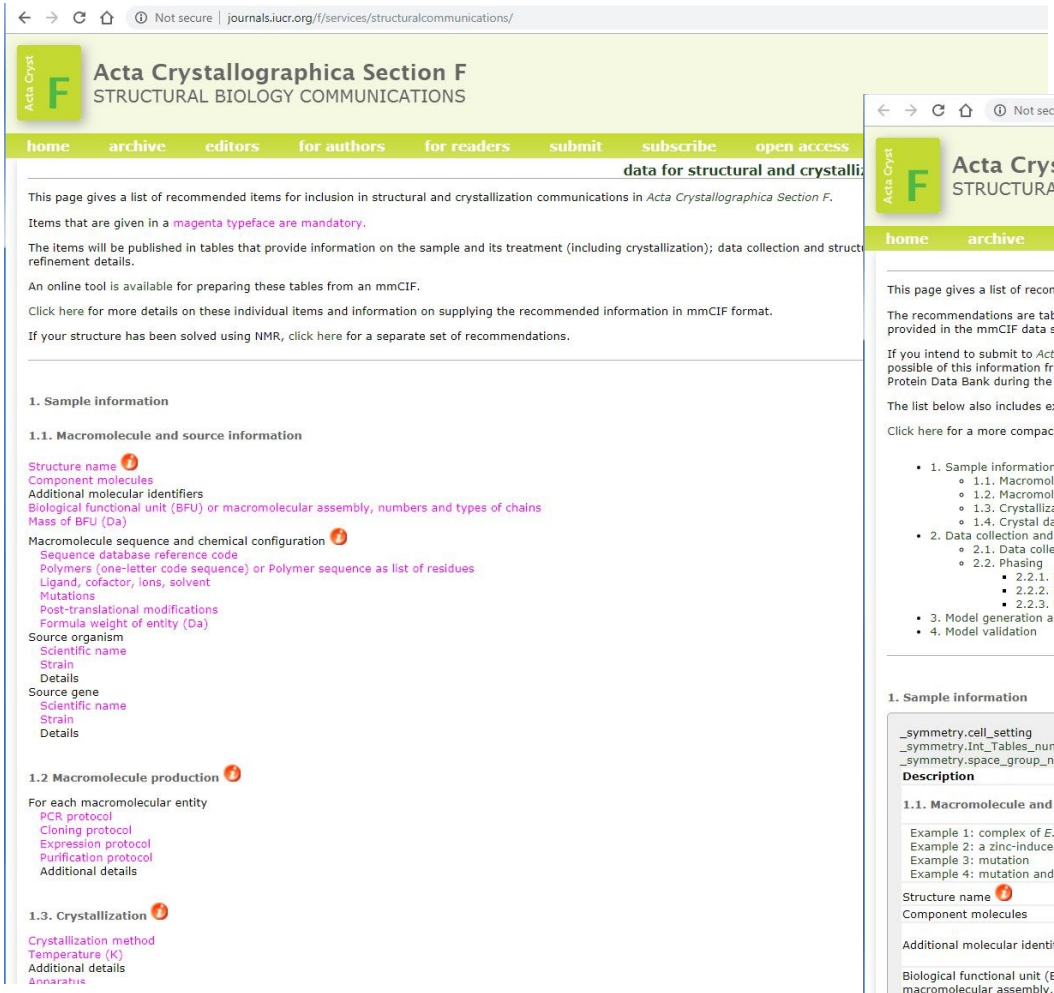
Global schema map of the entire PDB relational database; *from* Schierz, A. C., Soldatova, L. N. & King, R. D. (2007). *Nature Biotechnology*, **25**, 437-442

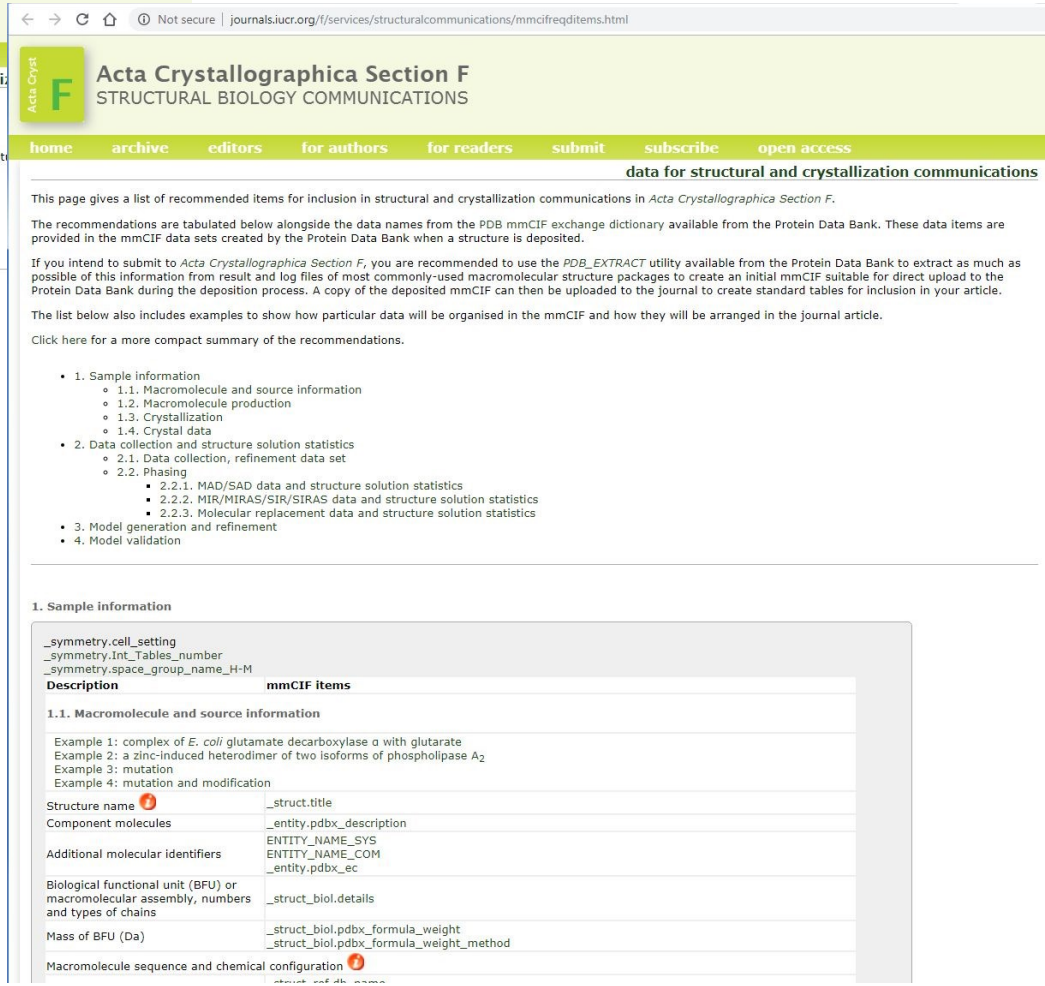
The Brookhaven Protein Data Bank was once a pioneering database, but its organization of structural data is now outdated and in need of an upgrade.

mmCIF ⟷ RDBMS

- *Critique*: too many tables in the database empty or nearly empty
  → poorly defined schema
- *Reality*: each table represents a category of information needed to describe the structure completely
- *Problem*: depositors unwilling to provide that information (*i.e.* populate the tables)
- Most of the poorly populated tables capture experimental *metadata*

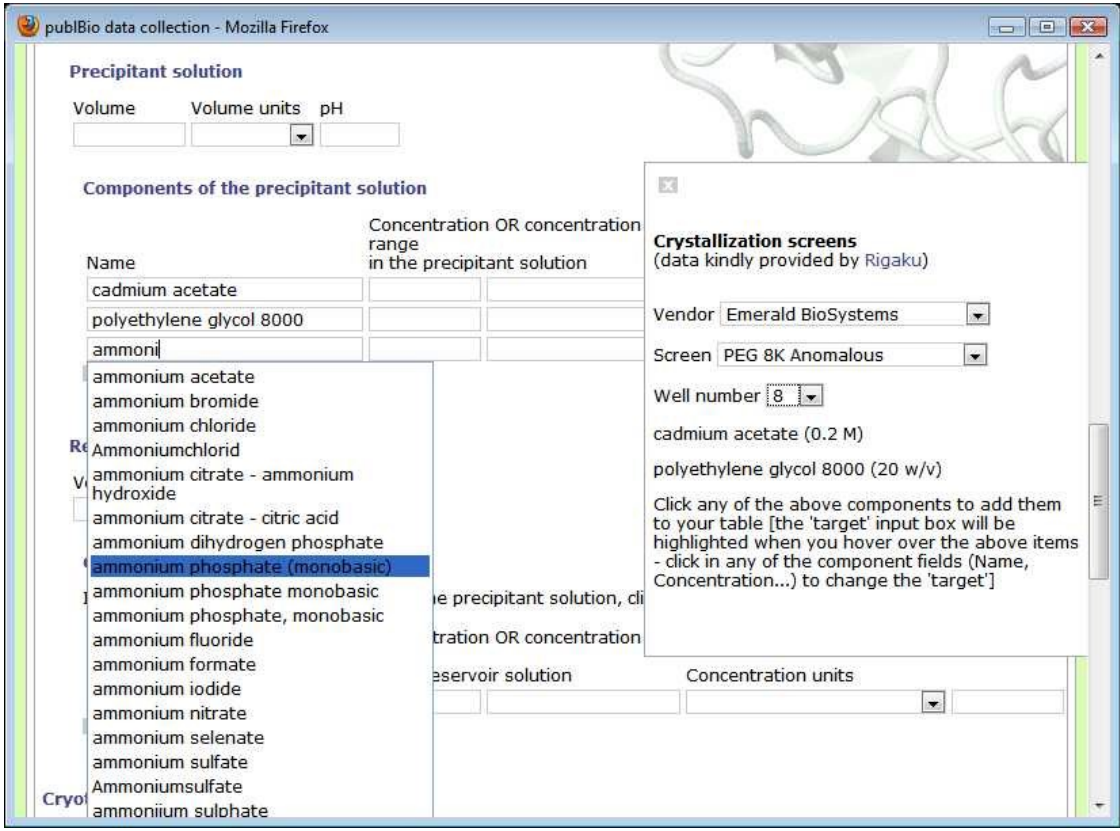
Metadata capture in macromolecular reports





Authoring tools to assist metadata capture (*publBio* publisher)

- For MX a particular shortcoming is details of sample preparation and crystallization
- *publBio* offers a helpful interface
- Still a drought of crystallization papers
- Perhaps a role for *IUCrData*?



The *checkCIF* paradigm

Three aspects to *checkCIF* validation:

1. Completeness (the ‘metadata’ problem)

The metadata problem

- *checkCIF* for small structures achieves this reasonably well
  - For IUCr journals: published requirements in Notes for Authors
  - For other journals: in principle, customised ‘request lists’; in practice, *de facto* acceptance of IUCr recommendations with *ad hoc* pick and choose
  - Possibility of better uptake through automated LIMS systems
- For protein structures
  - Reluctance in specifying mandatory data items
  - Inertia in design of standard experimental protocols in some facilities
  - Automated LIMS and other systems would help
- For raw data
  - CommDat sponsoring ‘*checkCIF* for raw data’ initiative

The *checkCIF* paradigm

Three aspects to *checkCIF* validation:

1. Completeness (the ‘metadata’ problem)
2. Internal self-consistency (relational methods)

Internal self-consistency

- IUCr *checkCIF* tests perform wide range of consistency checks
- Many of these also in *PLATON*
- Relational nature of DDL2 ensures category integrity within mmCIF
- DDLm offers a generic approach to specifying integrity relationships in dictionaries
- *dREL* a specific implementation of DDLm methods

The *checkCIF* paradigm

Three aspects to *checkCIF* validation:

1. Completeness (the ‘metadata’ problem)
2. Internal self-consistency (relational methods)
3. Comparison with related structures (the ‘knowledge’ problem)

The knowledge problem

- *PLATON* includes huge amount of crystallographic and chemical knowledge
- *Mogul* offers opportunity to compare new structures with existing curated ones in CSD
- PDB Validation Reports build on existing knowledge bases
- These tools are tuned to molecular geometry and other chemical properties
- Prospect of interrogating any ensemble of CIFs for any other defined properties through a non-specific API

Pain points for CIF in the 21st century

- Reluctance/difficulty in supplying experimental metadata
- Reticence in revising core dictionaries
- Lax approach to detailed standards in some software packages
- Lack of resources in developing CIF2/DDLm applications
- Lack of resources for validation software maintenance and development

- None are critical – many show some (but slow) progress
- Important to keep momentum and train new generation

“There is a traditional hierarchy of components of understanding: **data**, **information**, **knowledge**, **wisdom** (the DIKW model).

Crystallographic **information** is the component that bridges the gap between the raw experimental **data** and the global **knowledge** bases represented by the Protein Data Bank, Cambridge Structural Database, International Centre for Diffraction Data, Inorganic Crystal Structure Database, Crystallography Open Database etc.

This school will teach students to respect their raw data, extract the most reliable information they can, and disseminate that information in a complete and verifiable manner. In this way they will contribute to the sum total of scientific knowledge with rigour and integrity.

Crystallographic wisdom is outside the scope of this course.”

Crystallographic information in the FAIR era