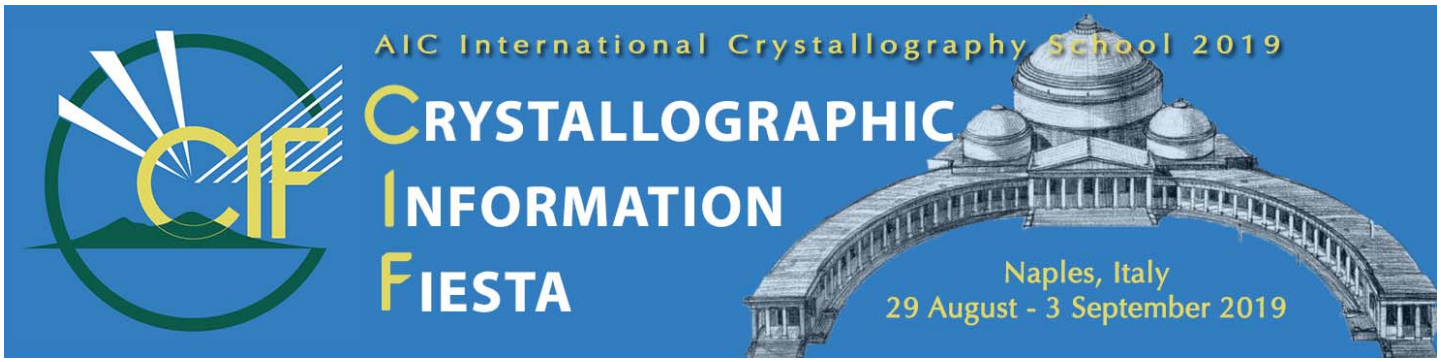


The devil is in the detail: sharing data

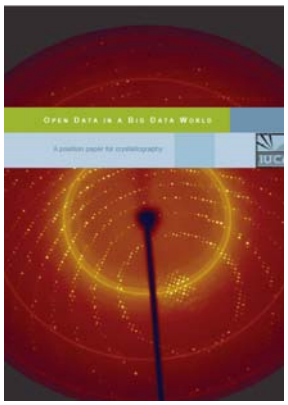
Brian McMahon

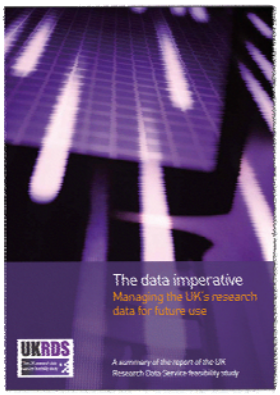
International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, UK

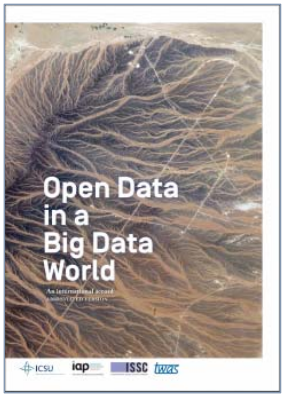
bm@iucr.org

A historical account of the development of a standard exchange mechanism for crystallographic software. What began as an exercise in defining a common data exchange format has become a fully-featured Crystallographic Information Framework with wide-ranging uses in publishing, data validation and archiving.

Big data leading to big policies!







A phrase much in vogue in recent years has been “Data Deluge” – the recognition that there are vast quantities of experimental and statistical data being generated by scientific research worldwide, and that the quantity of such data threatens to overwhelm our ability to process, store and re-use it effectively. A response to this prospect has been the proliferation of science policies addressing the current and future needs of the scientific community, and the potential societal effects of this new data-driven scientific paradigm. There is now huge investment in building up data management infrastructure and the expertise to use it properly. For crystallography, which is a science where the detailed analysis of data has always been crucial, there is in a sense nothing new about this. Excellent data retention policies were the hallmark of the pioneers of crystal structure determination by X-ray diffraction; the crystallography journals of the IUCr since 1948 have required data deposition and supplied experimental and derived data in machine-readable form as soon as electronic publishing became established; high-quality curated structural databases have been part of the crystallographer’s research tool set since the 1930s. That is not to say that no challenges remain – the quantity of crystallographic data is still growing rapidly, particularly if one includes time-resolved XFEL experiments. This lecture focuses on the development of the standards for information management that the IUCr has developed over the last quarter century, and their future directions.

Fit for purpose?

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.

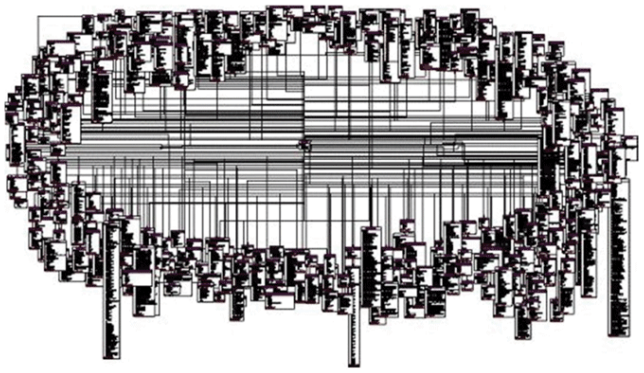
The PDB was recently reengineered to become a 'relational' database based on the mmCIF schema[4]. As wwPDB, it now includes the PDB at the Research Collaboratory for Structural Bioinformatics (RCSB, Brookhaven), the Macromolecular Structure Database (MSD) at the European Bioinformatics Institute (Hinxton 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

There is a potential problem with having been around for a long time – there is the danger of being considered obsolete in the face of new trends. In 2007 a surprising paper was published in Nature Computational Biology. Written by database specialists, their analysis concluded that the description of macromolecular structures stored in the Protein Data Bank was very inefficient. Many tables in the relational database were almost empty, a sign (they claimed) that the underlying data model was inefficient and poorly thought out. What they failed to recognise was that the "data model" had been built up over many years of careful thought and study, and with a 20-year experience of the needs and limitations of an existing database (the original PDB implementation at Brookhaven National Laboratories). The new PDB relational schema was designed to capture the many scientific relationships between experiments, software models and macromolecular structures. The underpopulation of database tables reflected the community's relative unwillingness to collect and store data or metadata that would actually enhance the overall value of the database as a scientific research tool. That is in part a failure of the current research funding and literature cultures, but it is also in part a reflection of the lack of understanding of the scientists themselves of the value of all the detailed data they are in a position to collect. It is partly to enhance that understanding that this school has been conceived. But crystallography is in an excellent position compared with many other sciences because of the thought and vision that has been put into its information and data characterisation over many years. Let us begin with a historical review of how we come to be in that happy position.

The initial challenge

Jim Stewart

Howard Flack

- Import diffraction data from many different instruments
- Relatively simple data description (intensity counts, goniometer axes, time)
- Necessary to capture certain metadata (instrument geometry, X-ray wavelength, reference reflections)
- Useful to capture additional metadata (incident-beam characteristics, chemical formula, environmental conditions, cell parameters,…)
- *Xtal* program system
- *XRAY* computational model

Our story really begins during the 1980s, when the many different crystallography research groups began to exploit burgeoning computer power to write software for a variety of purposes in data collection, evaluation, reduction, processing and analysis. At the same time, instrument vendors were outputting their measurements in ways that could be directly fed into these new programs. With the growing multiplicity of software and instrumentation, it soon became apparent that some sort of standardization was important. Jim Stewart was a pioneer of coordinating software into multi-contributor packages (*XRAY*67), and its descendant, *Xtal*, led by Stewart and Syd Hall, was attracting many international collaborators. One of these was Howard Flack, who was particularly concerned with capturing data from instruments produced by many different vendors.

The rise of the databanks

- Powder Diffraction File (1938)
- Cambridge Structural Database (~1965)
  - ASER, BCCAB file formats
- Protein Data Bank (~1971)
  - PDB file format
- Nucleic Acid Database (1991)
- Inorganic Crystal Structure Database (~1978)
- NBS/NIST Crystal Data (~1965)
- CRYSTMET (~1960, 1974)

Olga Kennard

Walter Hamilton

Alan Mighell Tom Koetzle

Helen Berman

Also significant was the rise of the structural databases, which were interested in capturing as much information as possible about as many aspects as they could harvest of the crystallographic experiment and the derived crystal structures. Prior to the age of electronic publishing, data had to be transcribed by hand from journal papers, a laborious and error-prone procedure. However, to help mitigate the propagation of errors, the databases did develop validation programs to cross-check the data that were entered for internal self-consistency and, increasingly, for chemical reasonableness based on statistical analysis against the other structures they were collecting. These were in some ways precursors of the checkCIF and PDB validation tools that were to be developed in later years. The individuals on this slide were involved in the establishment of many of the most important structural databases.

A Standard Crystallographic File Structure

- 1978: IUCr Warsaw: Working Party appointed by Data and Computing Commissions
- 1981: IUCr Ottawa adopted SCFS-81
- 1985: Revised version SCFS-84
- 1987: Final version SCFS-87
- FORTRAN i/o paradigm
- Copy of SCFS-87 on *ITG* CD-ROM

I. David Brown

A very significant advance in the standardisation of crystallographic information was the development of a Standard Crystallographic File Structure (SCFS) in 1981. It was developed by a group headed by David Brown, and had important support from Howard Flack. David has a very precise mind, and was enthusiastic in the task of collecting together all the different concepts that needed to be codified in a data exchange standard. He was also aware of the far-seeing terms of reference adopted by the working group, among which were: “1. The file structure must be extendable to include all types of crystallographic data. 2. It must be compatible with current and future methods of data transmission.” In the end it was very little used, partly because its rigid formatting rules, appropriate for FORTRAN input/output, were less suitable for the newer computer languages that were beginning to emerge. Nevertheless, the organisation of the material considered necessary for an effective data management pipeline was an essential foundation for the next major advance.

Electronic publication requirements

- IUCr Editorial office, Chester, 1990
- *Acta Cryst. Section C: Crystal Structure Communications*
- Basic computer facilities
- Need to automate text-processing
- Helpful to have richer annotation of body part of file

Sidney Abrahams

Acta Cryst. (1991). C47, 1721–1723

Lewis-Base Adducts of Group 11 Metal(I) Compounds. 60. Binuclear Adducts of Copper(I) Halides with 2-Hindered Pyridine Bases

BY PETER C. HEALY

School of Science, Griffith University, Nathan, Australia 4111

AND JOHN D. KILDEA, BRIAN W. SKELTON, A. FIONA WATERS AND ALLAN H. WHITE

Department of Physical and Inorganic Chemistry, University of Western Australia, Nedlands, Australia 6009

(*Received* 22 *June* 1990; *accepted* 8 *November* 1990)

Abstract. (I): Di-μ-bromo-bis[bis(2-bromopyridine)-copper(I)], $[Cu_2Br_2(C_5H_4BrN)_4]$, $M_r = 918.9$, triclinic, $P\bar{1}$, $a = 10.224(4)$, $b = 8.935(2)$, $c = 7.892(2)$ Å, $\alpha = 68.84(2)$, $\beta = 71.96(3)$, $\gamma = 83.23(3)°$, $V = 639(1)$ Å^3, $Z = 2$, $D_x = 2.39$ g cm^{-3}, Mo $K\alpha$ radiation, $\lambda = 0.71069$ Å, $\mu = 109$ cm^{-1}, $F(000) = 864$, $T = 293$ K, final $R = 0.052$. (II): Di-μ-chloro-bis[bis(2-benzylpyridine)copper(I)], $[Cu_2Cl_2(C_{12}H_{11}N)_4]$, $M_r = 874.9$, triclinic, $P\bar{1}$, $a = 16.441(5)$, $b = 9.183(5)$, $c = 7.661(2)$ Å, $\alpha = 76.87(4)$, $\beta = 81.65(3)$, $\gamma = 73.17(4)°$, $V = 1074(1)$ Å^3, $Z = 2$, $D_x = 1.35$ g cm^{-3}, Mo $K\alpha$ radiation, $\lambda = 0.71069$ Å, $\mu = 8.1$ cm^{-1}, $F(000) = 452$, $T = 293$ K, final $R = 0.047$. wR(observed reflections) = 0.059, S(all reflections) = 2.29, S(observed reflections) = 2.77, weights based on measured σ's; $(\Delta/\sigma)_{max} = 0.088$, $(\Delta/\sigma)_{mean} = 0.004$, $\Delta\rho_{max} = 0.966$, $\Delta\rho_{min} = -0.946$ e Å^{-3}, no correction for secondary extinction.

(II): Plate, colourless, crystal size 0.5 × 0.2 × 0.03 mm, scintillation counter. Diffraction measurement method: $2\theta/\theta$, diffraction temperature 293 K. Absorption correction type: Gaussian. $A^*_{min} = 1.03$, $A^*_{max} = 1.17$. $\theta_{min} = 1.30$, $\theta_{max} = 24.99°$; h 0→19, k −10→10, l −8→9. Eight standard reflections, measured every 100 reflections, intensity variation 0%. Criterion for observed reflections: $I > 3\sigma(I)$. Full-matrix least-squares refinement, 2189 reflections 'observed' out of 3778 indepen-

By the start of the 1990s, a clear requirement for an effective data exchange mechanism was that it should be able to facilitate the publication of crystal structure reports, which were full of numerical data, and the error-free transfer of the data associated with such publications to the database. During his term of office as Editor of *Acta Crystallographica*, Sidney Abrahams had developed a structured publication format for *Acta Cryst. C*, which established the norm for publication of structural information and (through Notes for Authors) established what were the most important experimental metadata needed to reproduce and validate the scientific arguments in the paper. These publication requirements, together with the needs of the crystallographic databases, were to define clearly the content of the next data exchange standard that the IUCr wanted to develop.

Working Party on Crystallographic Information

- Appointed by IUCr in response to proposal for electronic submission of manuscripts at IUCr XIV Congress, Perth (1987)
- Convened ECM11, Vienna (1988)
- Proposed adoption of a STAR-file-based standard

Ted Maslen

Frank Allen

Charlie Bugg

And so, at the direction of the incoming Editor-in-Chief of IUCr Journals, Charlie Bugg, a working party on Crystallographic Information was formed under the leadership of Ted Maslen. Its initial purpose was to specify a new information exchange standard that could be used for the electronic submission of manuscripts. The group included Frank Allen, at that time leading scientist at the Cambridge Crystallographic Data Centre, David Brown, who had created the SCFS standard, and Syd Hall, a colleague of Ted's from the University of Western Australia, who was exploring a novel approach to information storage using a general, free-form extensible format known as STAR (for Self-Defining Text Archive and Retrieval) File.

Self-Defining Text Archive and Retrieval (STAR) format

- A Universal Archive File:
  - Used to store *all types* of data
  - *Not* (*necessarily*) a database file
  - *Machine independent*
  - *Simple* to read and access
  - *Flexible* to future change

Syd Hall

Nick Spadaccini

The STAR file: a new format for electronic data transfer and archiving

Sydney R. Hall *J. Chem. Inf. Comput. Sci.* (1991), **31**, 326 – 333

[doi: 10.1021/ci00002a020]

The STAR File was developed with a similar, but even more wide-ranging set of forward-looking requirements as SCFS. It had the potential to be adopted in any discipline, and STAR-based applications were developed in quantum chemistry, botanical taxonomy, NMR structure representation. Initially it was based on the binary data packing strategy used in Xtal, with a header which indexed the location of specific data items later in the body of the file. But in early experiments with the IUCr publishing office, it became apparent that the header was unnecessary, and a simpler approach was to declare each item as it appeared in the file with an initial tag, and then present the value of that tag. Where an item took on multiple values, it could be looped together with other items closely related to it. In fact the tag-value approach was not unlike the approach of XML – but this was five years before XML appeared!

Simple tag, value structure

```
#......................................................................
data_manuscript
_manuscript_summary
; This is some dummy text to show how a multiple data-block STAR file works!
;

data_crystal_structure
 _chemical_formula       'C13 H12 O5'
 _chemical_name
; 3-(2,5-dihydro-4-hydroxy-5-oxo-3-phenyl-2-furyl)propionic acid
;
 _publication_title
; Structure of WF-3681, 3-(2,5-Dihydro-4-hydroxy-5-oxo-3-phenyl-2-furyl)propionic Acid.
;

_cell_a                  18.757(8)
_cell_b                   7.282(2)
_cell_c                  17.511(8)
_cell_alpha              90
_cell_beta               91.20(3)
_cell_gamma              90
_cell_volume             2391(3)

_symmetry_space_group    '-C 2yc'
loop_ _symmetry_pos_in_xyz
 'x,y,z'              '-x,-y,-z'           '-x,y,1/2-z'
 'x,-y,1/2+z'         '1/2+x,1/2+y,z'      '1/2-x,1/2-y,-z'
 '1/2-x,1/2+y,1/2-z' '1/2+x,1/2-y,1/2+z'
```

October 22, 1989

The example on this slide demonstrates how from the beginning the format was intended to accommodate the full text of a scientific paper. The original idea was that there would be a single tagged field (**_manuscript_summary**) into which the author would place the entire text of the manuscript, including all the crystallographic data, tables etc. that consisted of material that was also placed elsewhere in the file by the structure solution and refinement software. It soon became clear that since errors can always arise when copying data from one location to another, a better approach would be to synthesise the reporting of data within the paper by using the other data items present in the file. This was greatly helped by the existing standardization of presentation of data that the Editor had imposed within the journals for many years.

We will not discuss the format in detail in this lecture, as it will be revisited later in the School, but you will see that it is very simple. Each data name is recognized because it has a leading underscore. The value follows, separated by white space (the amount of white space does not matter). If the value itself *contains* white space, it should be wrapped in quote marks or surrounded by semicolons in the first column. Looped values are laid out as if in a table, where the data names are listed together after the **loop_** keyword and the corresponding values follow in strict rotation.

Early adopters

- *Xtal*
- *NRCVAX*
- *PARST*
- *DIFFRAC*
- *SHELXL*-93
- *PLATON/PLUTON*
- *CRYSTALS*


Eric Gabe


Mario Nardelli


Howard Flack


George Sheldrick


Ton Spek


David Watkin

The standard was published in a classic article in the flagship IUCr journal *Acta Crystallographica Section A**, and was accompanied by editorials in *Acta Cryst. A*, *B* and *C* inviting electronic submission of articles reporting small-unit-cell crystal structures in the new format. The first paper submitted in CIF format appeared alongside the editorial in *Acta C*. That issue also contained a number of other papers with supplementary CIF data sets, as the editorial staff began to validate the numerical data in submitted papers using crystallographic programs that could read the new format. By 1996, *Acta Cryst. C* was able to make CIF its mandatory submission format. Such a relatively rapid adoption of the new standard by a diverse community was possible because of the enthusiasm with which software developers and instrument vendors recognised the benefits of full interoperability. A number of the most significant early adopters are recognised on this slide.

* Hall, S. R., Allen, F. H. & Brown, I. D. (1991). The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Cryst.* (1991). A**47**, 655-685 [DOI: 10.1107/S010876739101067X]

Dictionary definition language

```
data_atom_site_attached_hydrogens
    _name                       '_atom_site_attached_hydrogens'
    _category                     atom_site
    _type                         numb
    _list                         yes
    _list_reference             '_atom_site_label'
    _enumeration_range            0:8
    _enumeration_default          0
    loop_ _example
          _example_detail         2    'water oxygen'
                                  1    'hydroxyl oxygen'
                                  4    'ammonium nitrogen'
    _definition
;             The number of hydrogen atoms attached to the atom at this site
              excluding any hydrogen atoms for which coordinates (measured or
              calculated) are given.
;
```

Although the adopted file format was convenient for computational purposes, by far the most important element of the CIF standard was the comprehensive and detailed definitions that it included. Another early innovation was the recognition that the attributes of the definitions were themselves data, and so were amenable to documentation using the same format as the data files themselves. This raised the exciting possibility that software capable of reading CIF data could in fact read CIF data definitions also, and begin to validate data against the constraints of their machine-readable definitions. The data items that were created to describe the attributes of CIF data items were part of what became known as a dictionary definition language (DDL) – a term you will hear a lot during this School.

New CIF dictionaries

- 1997 Powder diffraction (pdCIF) – B. H. Toby
- 2002 Modulated and composite structures (msCIF) – G. Madariaga
- 2003 Precision electron density (rhoCIF) – P. R. Mallinson
- 2011 Restraints – I. D. Brown and I. Guzei
- 2014 Twinning – V. Young, I. D. Brown and J. R. Hester
- 2017 Magnetic – B. Campbell, J. M. Perez-Mato, V. Petříček, J. Rodriguez-Caval and W. Sikora
- 2018 Topology – D. Proserpio and V. Blatov
- Further extensions to Core


Brian Toby


Gotzon Madariaga


Paul Mallinson


Ilia Guzei


Vic Young


Branton Campbell


Václav Petříček


Davide Proserpio


Vladislav Blatov

Since CIF had been designed to be extensible and general, it soon became apparent that its use in other areas of structure determination was inevitable. A number of new sets of definitions (data dictionaries) were commissioned and produced, to cover different areas of structural science or to add new material to the core dictionary. These will be discussed in some detail later in the School, so here we simply recognize some of the people who were involved in creating these new dictionaries. As with all the portraits shown in this presentation, the purpose is partly to acclaim these individuals, but also to demonstrate how wide ranging within the crystallographic community had become the recognition of the significance of detailed definition and consequent data management. Many of the extension dictionaries, though small in extent, took a long time to be completed – a surprising result to many of the authors, who had not appreciated in fact how difficult it is to define new concepts with the sort of precision needed in computational applications.

The mmCIF workshops

- Macromolecular CIF dictionary working group
- York workshop (April 1993)
- Tarrytown workshop (October 1993)
- Brussels workshop (October 1994)
- Development of ‘relational’ DDL2


Paula Fitzgerald


Eleanor Dodson


Phil Bourne


Shoshana Wodak

Another large area of crystallographic interest that seemed very suitable for applying the CIF approach to was structural biology, and in particular the X-ray crystal structure determinations of protein and nucleic acid molecules. A working group was set up under the leadership of Paula Fitzgerald, and set about the task by involving eminent macromolecular crystallographers and the staff of the Protein Data Bank. However, it became apparent that this task was a major undertaking. Not only are biological macromolecules very large and complex, but their structure needs to be described in many ways, their intra- and intermolecular interactions are diverse and complex, and their actual determination from experiment through phasing and refinement can be very difficult and complicated. Crucial to the development of a formalism rich enough to capture all this information of vital scientific importance was a series of three important interdisciplinary workshops in the UK, USA and Belgium, that brought together crystallographers and information scientists. These resulted in an extension to the dictionary definition language that would allow description of the multiplicity of relationships amongst the many, many data items that were needed for a full description of a macromolecular structure and the experiment(s) from which it was derived.

Relational dictionary definition language

```
save__atom_site.attached_hydrogens
    _item_description.description
;               The number of hydrogen atoms attached to the atom at this site
                excluding any hydrogen atoms for which coordinates (measured or
                calculated) are given.
;
    _item.name                  '_atom_site.attached_hydrogens'
    _item.category_id             atom_site
    _item.mandatory_code          no
    _item_aliases.alias_name    '_atom_site_attached_hydrogens'
    _item_aliases.dictionary      cif_core.dic
    _item_aliases.version         2.0.1

loop_
    _item_range.maximum
    _item_range.minimum           8   8
                                  8   0
                                  0   0
    _item_type.code               int
    loop_
    _item_examples.case
    _item_examples.detail         2  'water oxygen'
                                  1  'hydroxyl oxygen'
                                  4  'ammonium nitrogen'
     save_
```

And here is an example of a data definition using the new dictionary definition language DDL2 that emerged from those workshops. Superficially, it does not look very much different from the DDL1 dictionary definitions, and that is a testament to the quality of the original CIF design. It does include a number of technical differences which make it more relational in nature, and in fact this approach produces a data model which was used by the Protein Data Bank to re-engineer its database on a fully-featured relational database platform. This is the same schema that was criticized by the *Computational Biology* paper alluded to previously, but the important thing to remember is that the scientific requirements were determined first, and a suitable computational representation and storage solution built to accommodate those requirements.

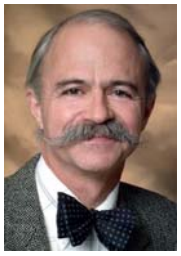
DDL2 applications

- Adoption of mmCIF by the Protein Data Bank
  - PDB Exchange Dictionary (PDBx/mmCIF) Version 5.0 supporting the data files in the current PDB archive
  - Integrative/Hybrid (I/H) methods extension dictionary
  - 3DEM Extension Dictionary
  - NMRSTAR Dictionary
  - Small Angle Scattering Dictionary
  - Model Archive Extension Dictionary
  - BIOSYNC Extension Dictionary
  - NMR Exchange Format Dictionary
- Symmetry CIF dictionary (symCIF)
- imgCIF Workshop, Brookhaven (October 1997)
- Crystallographic Binary File (CBF)

Helen Berman

John Westbrook

Andy Hammersley



Bob Sweet

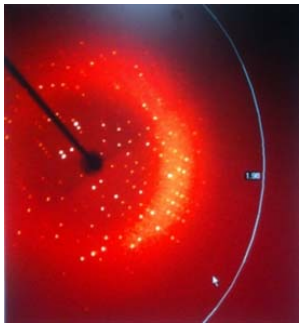
Herbert Bernstein

With the formal publication of mmCIF in 1997 and its use as the basis for the PDB database platform came a plethora of extension dictionaries in structural biology, which, like the small-unit-cell extensions, provided for new areas of research and novel techniques. This family of dictionaries, under the umbrella name of ‘PDBx’, is maintained and developed by the Worldwide Protein Data Bank, and John Berrisford will say more about these in his presentation. The dictionary definition language DDL2 developed to enable mmCIF/PDBx was also adopted for other applications. A symmetry CIF dictionary developed by David Brown (2001) extended the crystallographic symmetry definitions from the core dictionary into a larger set suitable for describing all the symmetry operations described in International Tables for Crystallography Volumes A (properties of space groups), A1 (relationships between space groups) and E (properties of layer and frieze groups), as well as allowing the description of higher-dimensional symmetry in CIFs reporting quasicrystal, magnetic and modulated structures. [This dictionary will be absorbed back into the latest versions of the core dictionary.] Another significant development was the creation of the imgCIF dictionary and associated CBF format allowing diffraction images – the raw data from most crystallographic structure determination experiments – to be brought within the growing family of CIF descriptions. By the late 1990s this family was being referred to as the Crystallographic Information Framework, in recognition of its applicability across different physical file formats and applications.

Types of crystallographic data described by CIF

‘Raw’ data

Numerical data collected directly from an experimental apparatus



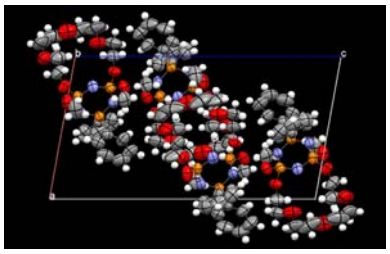
‘Processed’ data

Reduced, calibrated, processed numerical observations

```
# h,k,l, Fc-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
_exptl_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

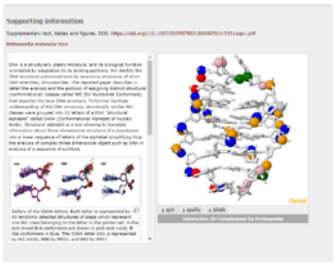
Numerical description of the parameters of a calculated structure model



‘Interpretative’ data

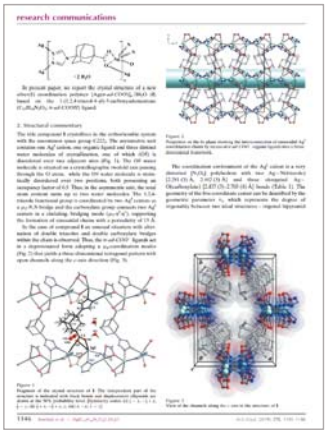
Variable parameters in the experimental set-up or numerical modelling and interpretation

Annotation



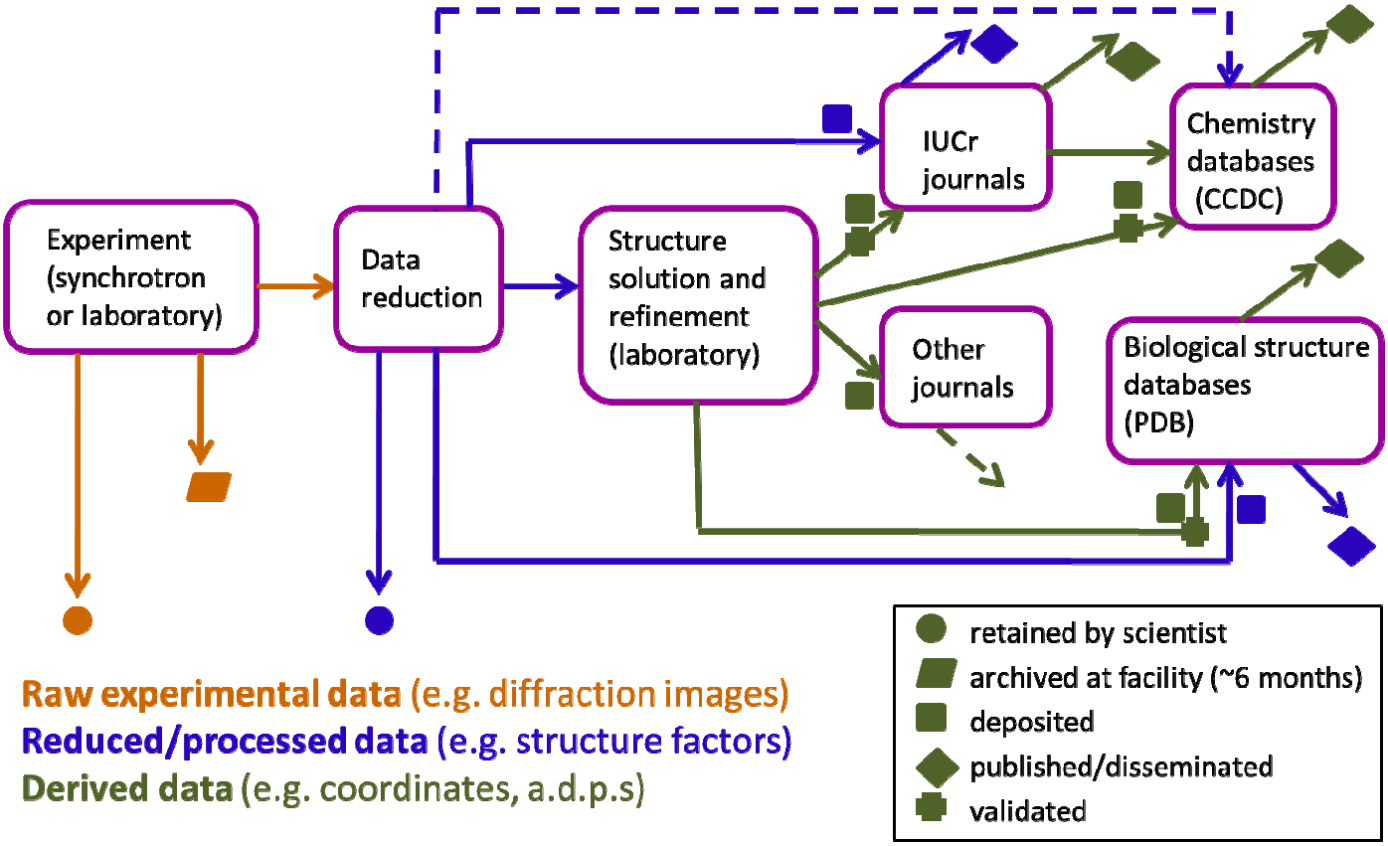
‘Reference’ data

Commentary



Through the generality of the underlying file format and the lack of differentiation between ‘data’ and ‘metadata’ classifications CIF has grown over the last quarter-century into a framework which spans the whole range of ‘data’ as it may be understood in a classical X-ray diffraction experiment and its subsequent interpretation and sharing of results. This slide illustrates some of the types of data handled by CIF. Raw data might be diffraction images or intensities measured on a point diffractometer. Processed data might appear as a set of scaled structure factors. Derived data could be the six-dimensional structural model (atomic positional coordinates and anisotropic displacement parameters). Interpretative data could include the refinement restraints and constraints in the structure solution process. Annotation includes descriptive relationships between subsets of the included structural and experimental data (automatically or manually-generate; the example shown is a Protopedia molecular tour served as supporting information to a journal article). Commentary refers in this case to a publication in the traditional scientific literature. Examples of reference data are the symmetry relationships stored in the Bilbao Crystallographic Server (shown here), the PDB ligand database, compilations of bond valence parameters *etc*.

Data flow in crystallography



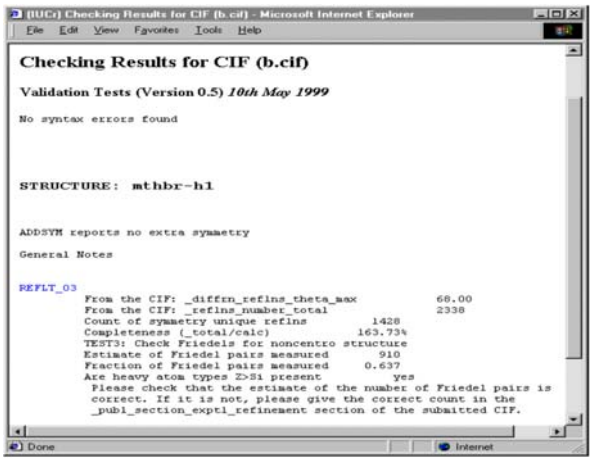
Raw experimental data (e.g. diffraction images)
Reduced/processed data (e.g. structure factors)
Derived data (e.g. coordinates, a.d.p.s)

A corollary of this is that the CIF acronym can be put to another use – that of describing a Cohesive Information Flow from experimental data, through analysis, interpretation and validation, to the synthesis of a narrative, its publication and the deposition of all supporting data in associated repositories. In this way it spans the Data – Information – Knowledge components that we have identified as significant elements of the message we propagate at this School.

Data validation: (*Acta Cryst.*) 1999

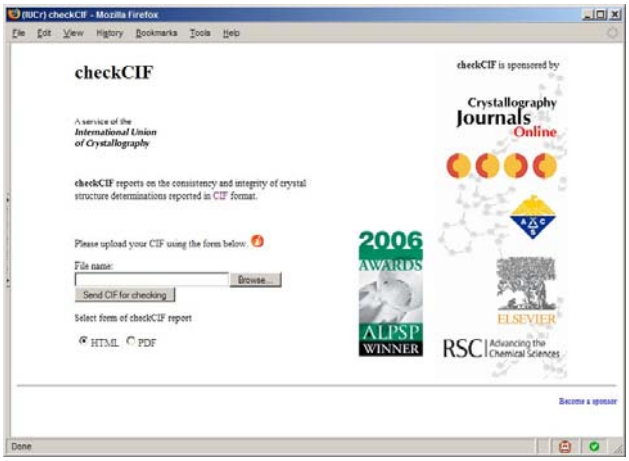
- In-house checks for correct syntax, consistency, completeness, style
- Crystallographic / geometry / crystal chemistry checks: *Xtal/PREPUB* (du Boulay & Hall), *PLATON* (Spek)
- New checking procedures : *e.g. BUNYIP* (Hester & Hall)
- Web-published checking algorithms

James Hester



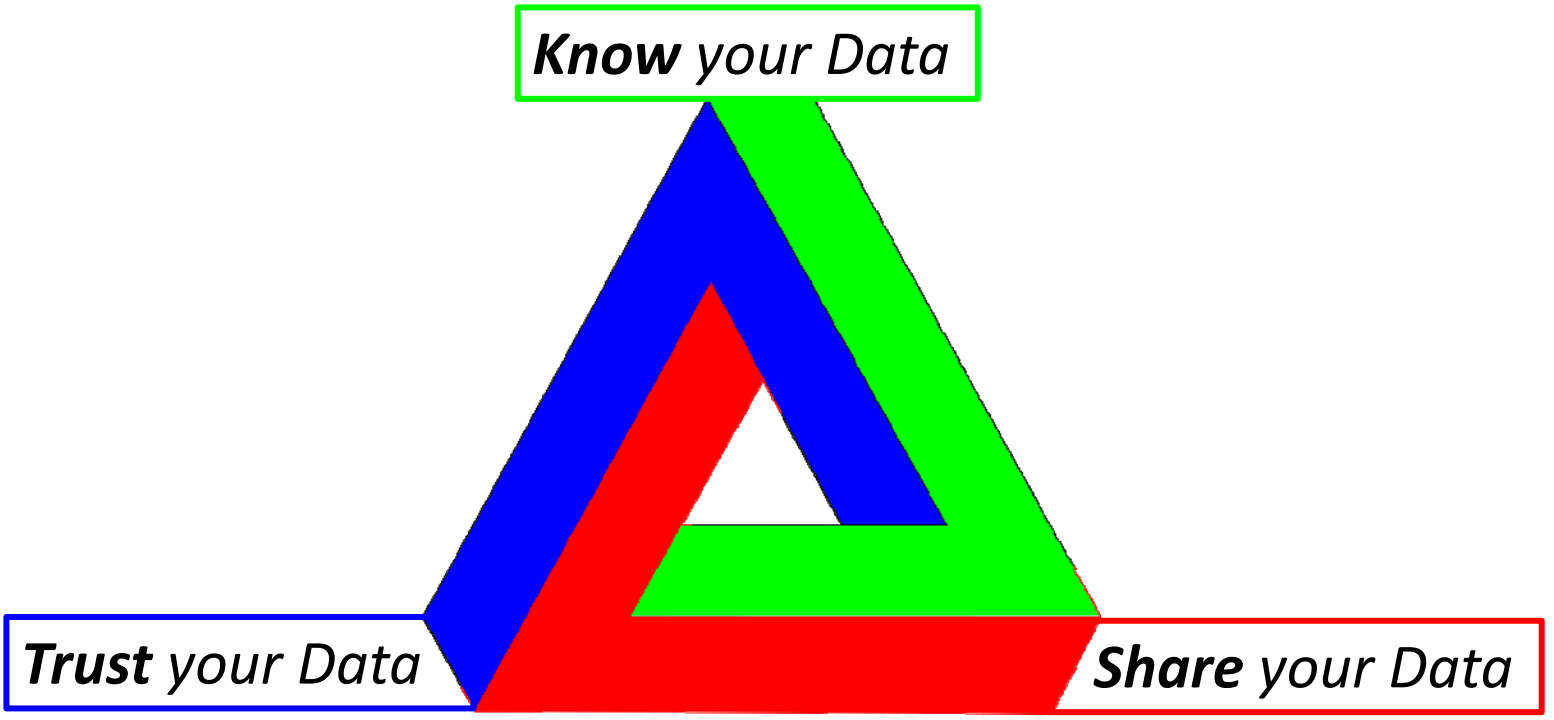
Data validation: 2007

- Public *checkCIF* service
- Sponsored by publishers / databases
- Community standard
- Recognised by ALPSP award for Innovation in Publishing
- Validation module for commercial submission systems



A very significant corollary of the development of machine-readable data definitions and attribute descriptors (an approach that is often described in modern information science as describing a “domain ontology”) is that automated systems could be devised for validating a data set, both for internal consistency and for reasonable values amongst an ensemble of similar structures. The development of these validation procedures is an early form of “artificial intelligence”, and you will hear much about the implications of this later in the School, both in the development of *checkCIF* for small-molecule structures and a wide range of macromolecular checks. I put this slide here partly to acknowledge James Hester as a pioneer in the more manual side of validation that became possible with access to machine-readable structural data. James is now Chair of COMCIFS, the IUCr CIF management committee, which is actively developing more automated and powerful validation and evaluation methods using DDLm and dREL.

Crystallographic information in the FAIR era



Some final thoughts arising from the introductory lectures to this School. We have seen that the FAIR principles are widely understood and respected – research data should be Findable, Accessible, Interoperable and Reusable. Crystallography is in a special position in addressing these requirements. We have a long tradition of precise careful measurement, statistical analysis and interpretation of experimental data, now backed up by the formal definitions in CIF dictionaries. We have highly refined computational tools for weighting and interpreting correlations and model-building constraints and restraints, now backed up by the intricate cross-checking and similarity analysis of *checkCIF* and other validation software. And we have a very open, well structured and well curated treasure-house of peer-reviewed journals and structural databases, also built on the knowledge-based approach that the Crystallographic Information grew from, and is continuing to perpetuate. Here I characterise these attributes as three vertices of an Escher triangle where knowledge of the data (in the sense of “the devil is in the detail”) inspires that trust that is needed for you and the scientific community at large to take the fullest advantage of it.