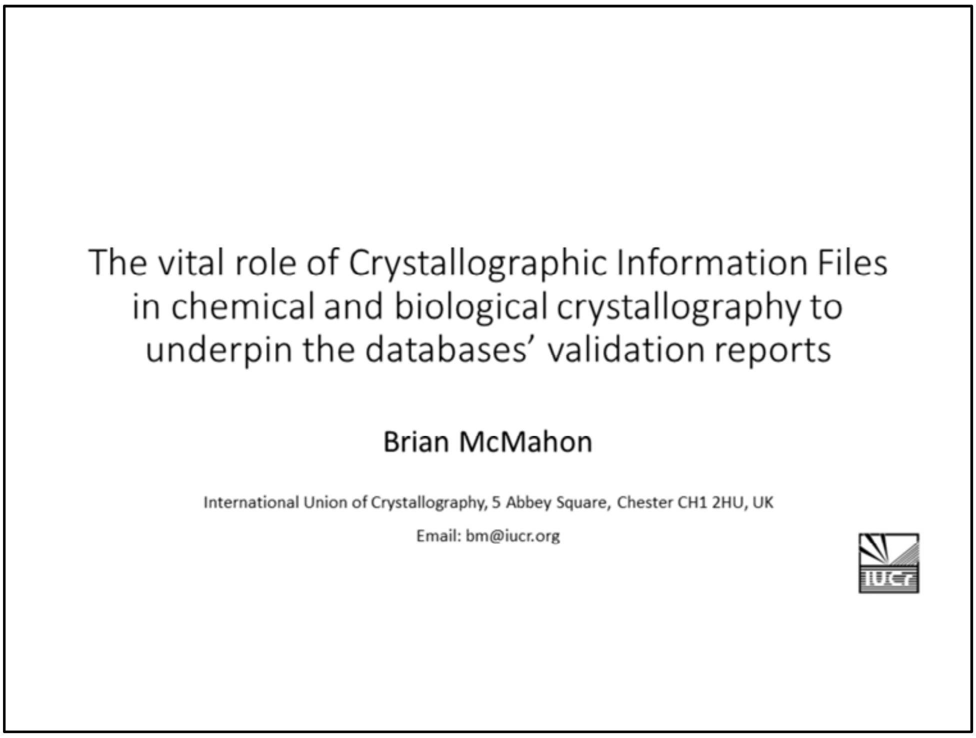


The vital role of Crystallographic Information Files in chemical and biological crystallography to underpin the databases' validation reports

Brian McMahon

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

Email: bm@iucr.org

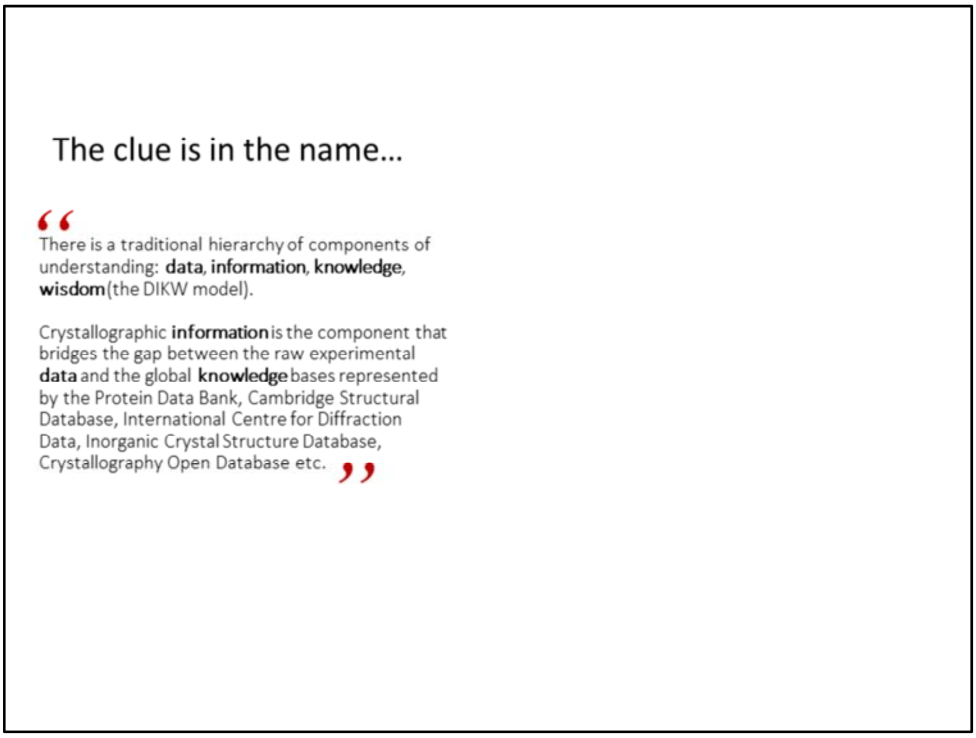
The Crystallographic Information File (CIF) was introduced in 1991 to facilitate data exchange between computer programs [1], but it soon became apparent that its most important feature was the *precise* definition of the data items that were manipulated by the programs. With standardisation of terminology across essentially all crystallographic software came the prospect of validating any structure, and indeed, to some extent, the experimental data upon which it was modelled. Structural databases already carried out extensive validation in their curating of stored data sets [2], but the development of protocols such as *checkCIF* [3] permitted extensive validation and evaluation of quality by the journals and indeed by the submitting author. The recent incorporation of software methods in the CIF dictionaries *via* the DDLm protocol [4] opens the door to even greater automation in both quality control and information retrieval from any solved crystal structure.

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.

See, for example, Karen, V. L. & Mighell, A. (1996). Special Issue: NIST Workshop on Crystallographic Databases. *J. Res. Natl Inst. Sci. Technol.* **101**, 205-381; Allen, F. H. & Glusker, J. P. (2002). Crystallographic Databases. Joint special issue. *Acta Cryst.* B**58**, 317-422; *Acta Cryst.* D**58**, 879-920.

Spek, A. L. (2009). Structure validation in chemical crystallography. *Acta Cryst.* D**65**, 148-155.

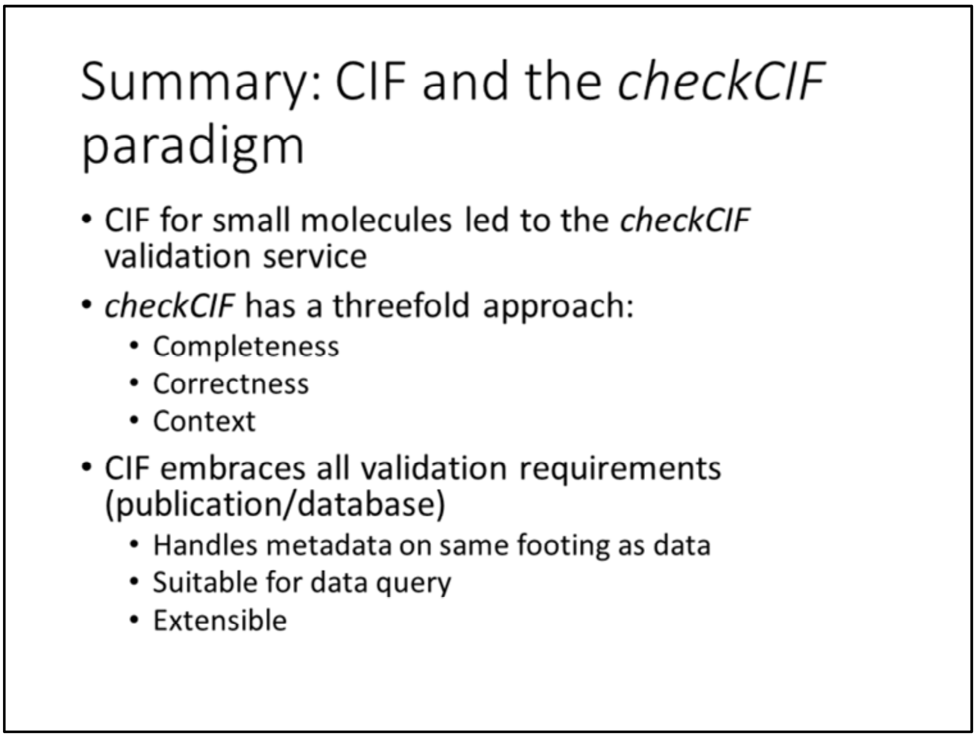
Spadaccini, N. & Hall, S. R. (2012). DDLm: a new dictionary definition language. *J. Chem. Inf. Model.* **52**, 1907-1916.



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.”

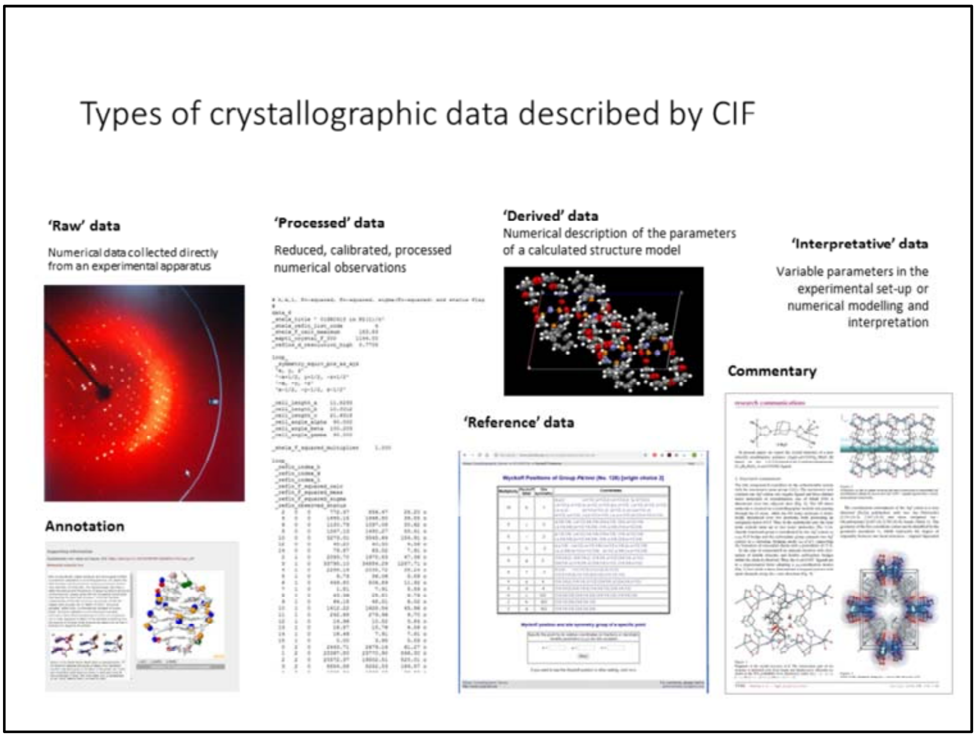
We begin by emphasising the care that went into the choice of name for the IUCr’s standardisation initiative – the Crystallographic Information File was always intended to be more than a simple file format for packaging numerical data. The source of the quotation will be revealed at the end of the presentation.



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

The presentation comes within a session of the Workshop entitled ‘The *checkCIF* paradigm’. I will review the three prongs of the checkCIF approach – here characterised succinctly as ‘completeness’, ‘correctness’ and ‘context’. Exactly what is meant by these terms will be discussed as we review each in turn. *checkCIF* of course rose naturally from the adoption of CIF as a data exchange standard, but there are three aspects of CIF that make it particularly appropriate that this should be so. The original file format makes no distinction between ‘metadata’ and ‘data’, so that everything can be collected together in the one file and kept together during processing. In practice this isn’t always the case – IUCr journals request that structure factor data be uploaded separately from the journal article, for example – but the conceptual design makes it easier to produce an integrated information management system. It’s also of significance that the definition of data items is also maintained (as CIF dictionaries) within the same formalism. This allows ‘bootstrapping’ of a knowledge system, from attribute identifiers to attributes of defined terms, to applications of those definitions. The discrete and well defined nature of each data item also means that a CIF can be used directly to construct a query against other CIFs. For example, the original electronic implementation of the *World Directory of Crystallographers* was as a collection of STAR Files. Only later was it loaded into a relational database. The extensibility of CIF means that it is ready to be used for novel techniques and applications just as soon as new definitions are minted.



Types of crystallographic data described by CIF

'Raw' data

Numerical data collected directly from an experimental apparatus

'Processed' data

Reduced, calibrated, processed numerical observations

'Derived' data

Numerical description of the parameters of a calculated structure model

'Interpretative' data

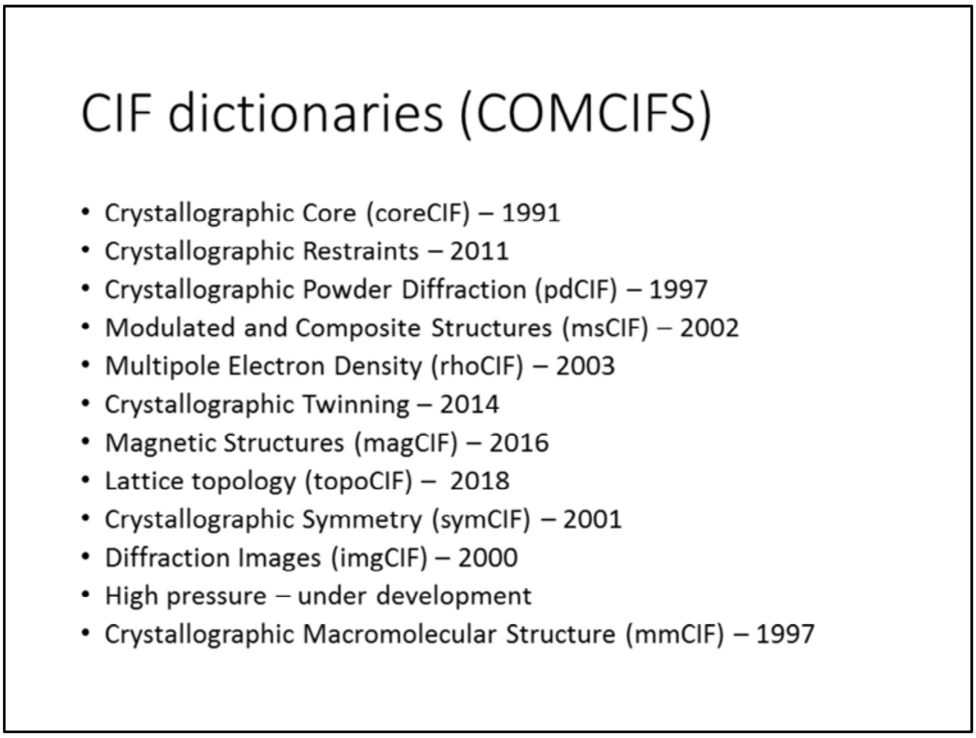
Variable parameters in the experimental set-up or numerical modelling and interpretation

Commentary

'Reference' data

Annotation

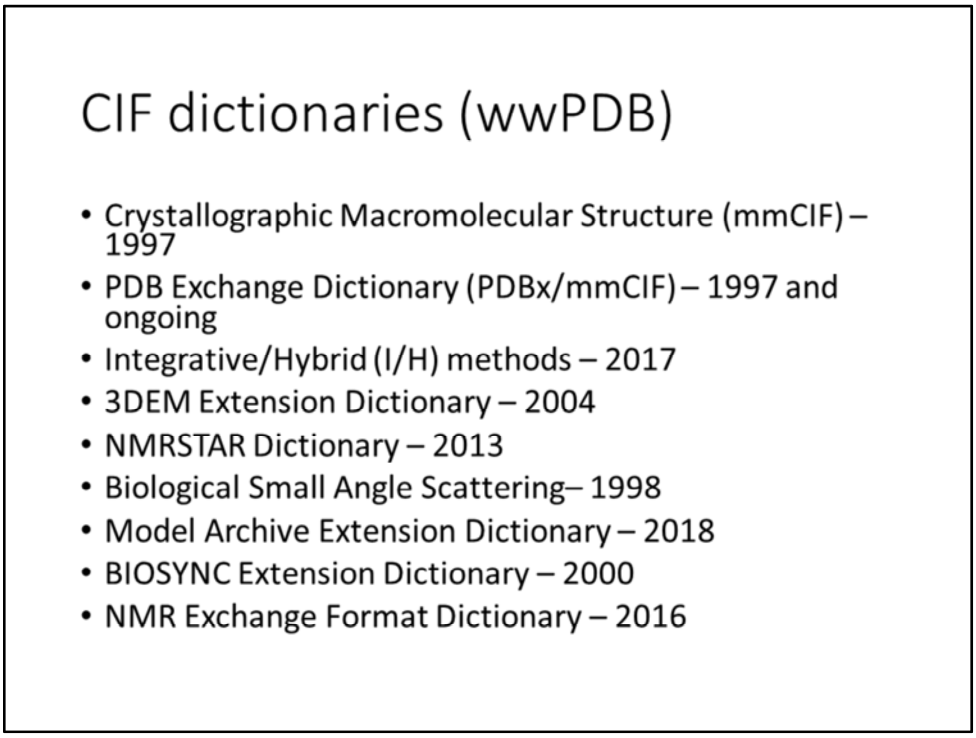
This slide emphasises the lack of distinction between 'metadata' and 'data', but at the same time it also illustrates what a variety there may be in types of data – a fact that is not necessarily well understood by policy makers who prescribe or mandate different approaches to research data management.



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

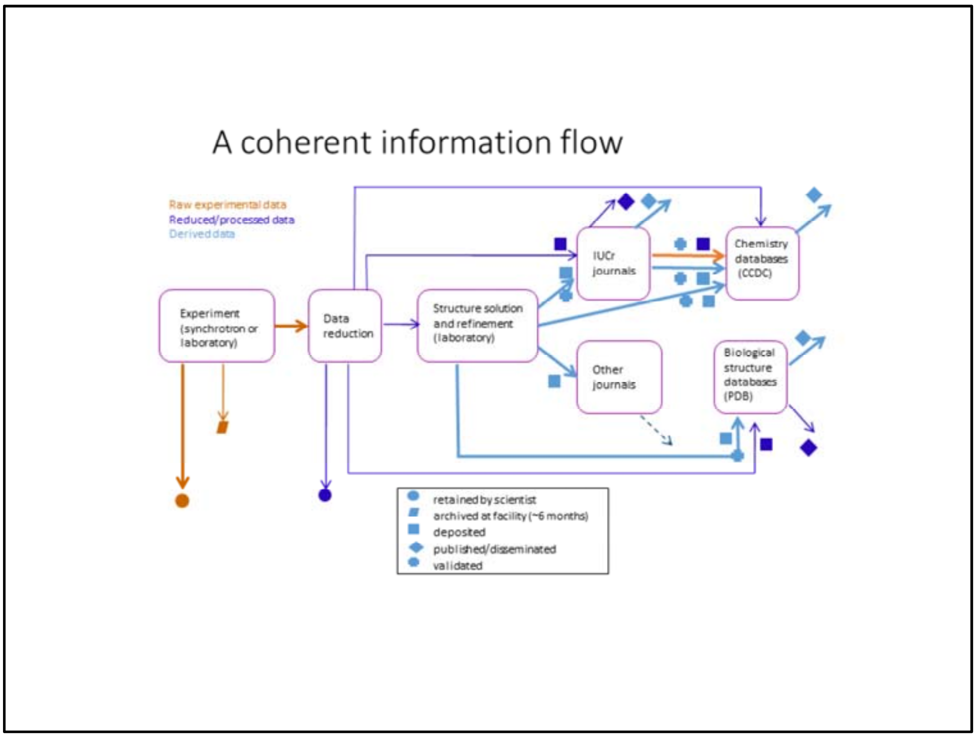
This tabulation of the CIF dictionaries under COMCIFS management illustrates the breadth of subject areas encompassed by the crystallography-based ontologies, and also the fact that they continue to be developed over time.



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

The Worldwide Protein Data Bank has been managing CIF dictionaries relevant to biological databases and research since 2006. Again, there is a wide spread of subject areas and a continuing development effort.



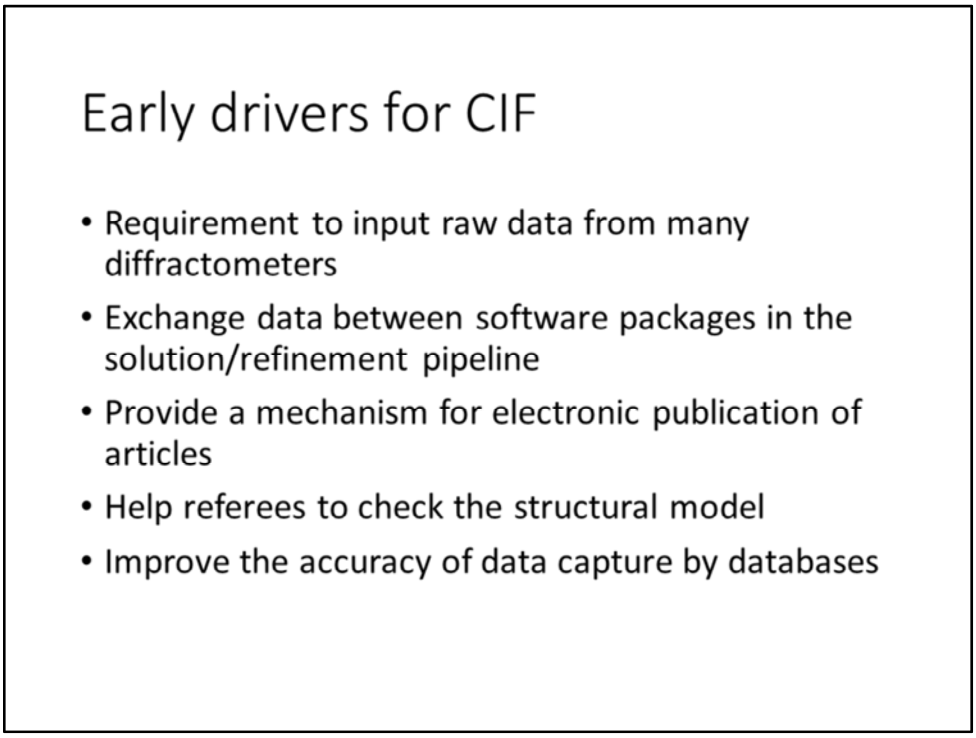
A coherent information flow

Raw experimental data
Reduced/processed data
Derived data

Experiment (synchrotron or laboratory)
Data reduction
Structure solution and refinement (laboratory)
IUCr journals
Other journals
Chemistry databases (CCDC)
Biological structure databases (PDB)

● retained by scientist
▰ archived at facility (~6 months)
■ deposited
◆ published/disseminated
● validated

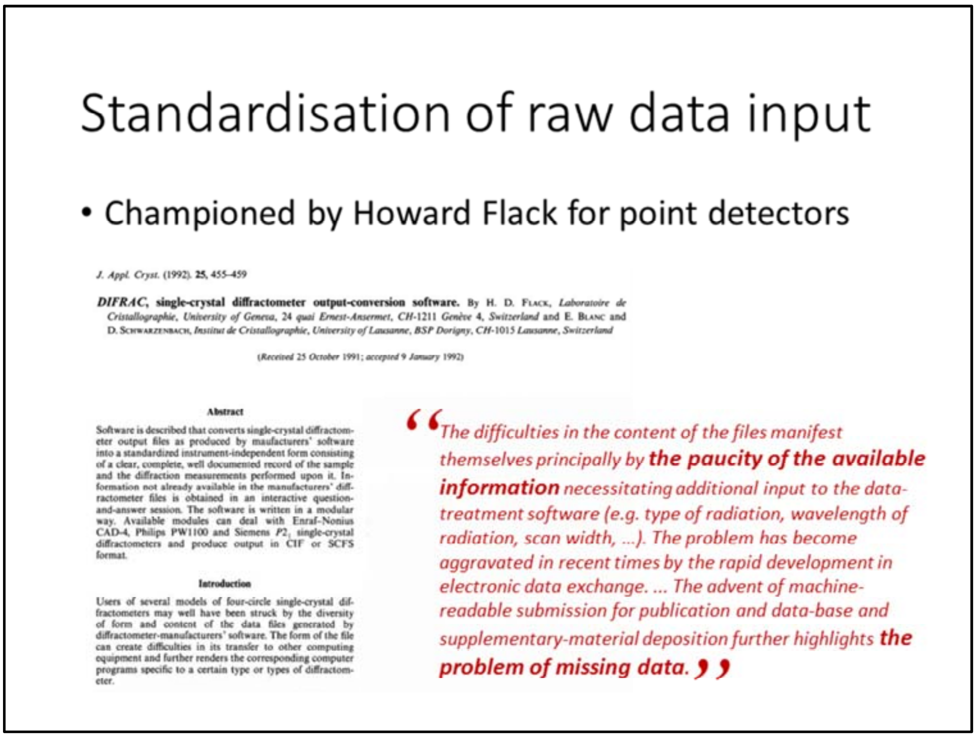
The acronym CIF – originally ‘Crystallographic Information File’, and subsequently ‘Crystallographic Information Framework’, as it became apparent that the actual file format was less important than the collection of precisely defined terms that the dictionaries provided – has also been subverted to this diagram, which illustrates the ‘coherent information flow’ all the way from the diffraction apparatus to the journals and databases. Again, in practice CIF files are not always involved at every stage of the pipeline – high data-rate image capture in synchrotrons increasingly uses HDF5/NeXus files. Nevertheless, the conceptual framework unites all the various processes, and informs approaches to frictionless information transfer (the NeXus profile for macromolecular crystallography is built very much on the imgCIF dictionary for diffraction images).



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

In this part of the talk I carry out a historical review of the design principles guiding CIF from early days, in part to highlight the foresight of Syd Hall, Frank Allen and David Brown in their initial design, but also to recognise that the community has always been able to conjure up a holistic view of its data management requirements.



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

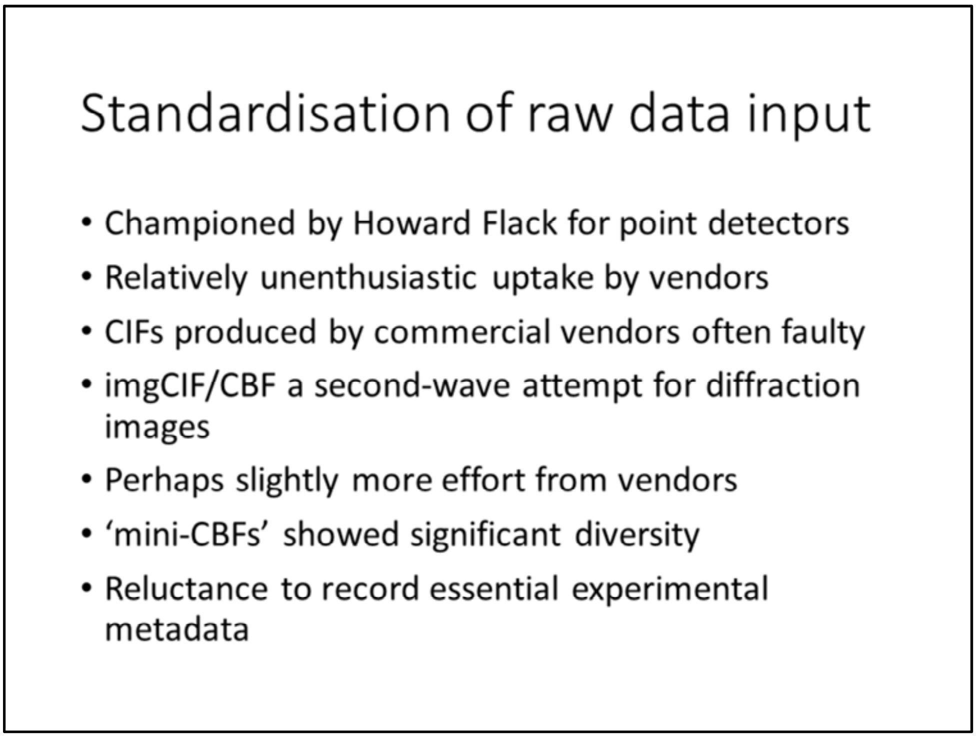
Software is described that converts single-crystal diffractometer output files as produced by maufacturers' software into a standardized instrument-independent form consisting of a clear, complete, well documented record of the sample and the diffraction measurements performed upon it. Information not 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.*** ”

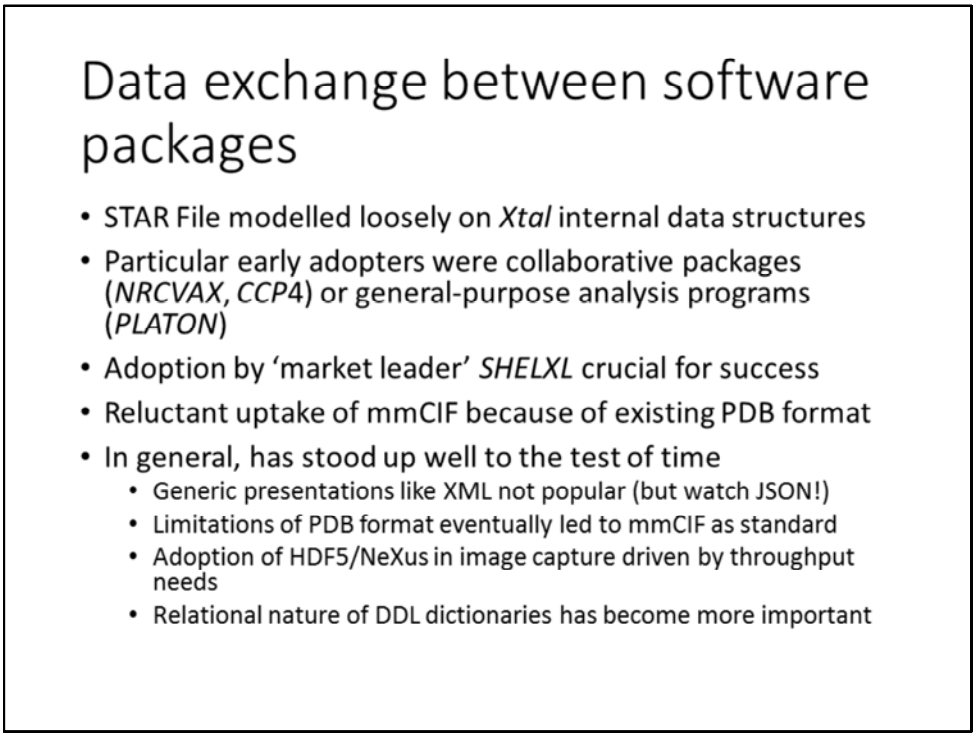
Howard used to insist to me that CIF came about because he challenged Syd Hall to produce a format that could be used by all the various diffractometer vendors. This was a topic Howard was much exercised by, and he was responsible for the *DIFRAC* module of *Xtal* that captured the wide and growing range of raw (point diffractometer) input data. The complaint in his 1992 paper that the raw data files usually lacked essential metadata is hardly less relevant today.



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

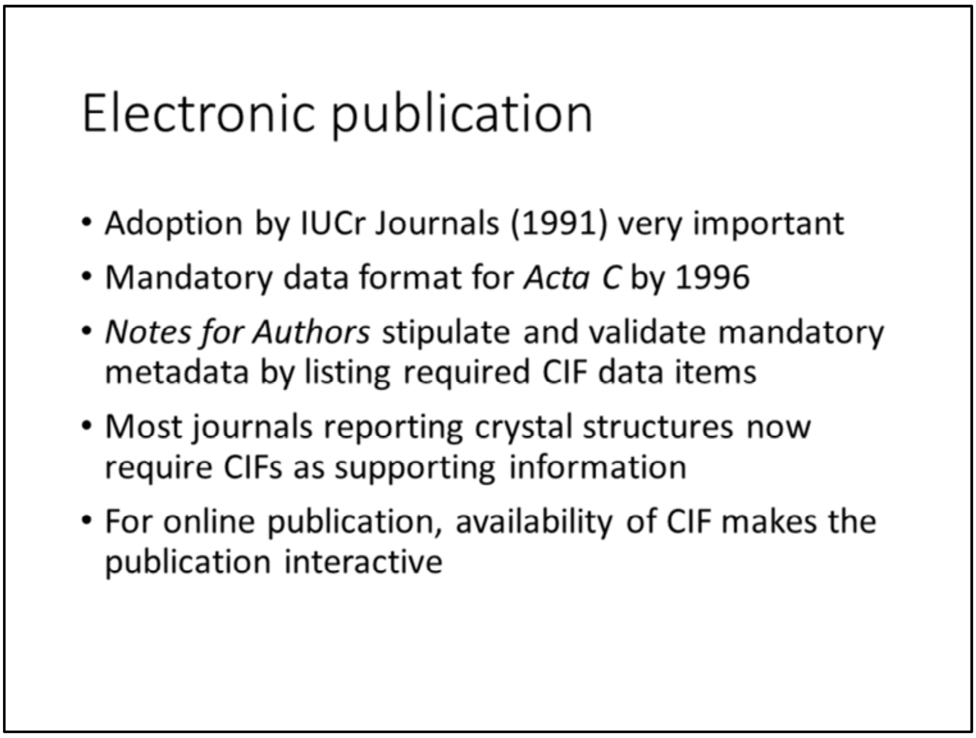
Howard used to insist to me that CIF came about because he challenged Syd Hall to produce a format that could be used by all the various diffractometer vendors. This was a topic Howard was much exercised by, and he was responsible for the *DIFRAC* module of *Xtal* that captured the wide and growing range of raw (point diffractometer) input data. The complaint in his 1992 paper that the raw data files usually lacked essential metadata is hardly less relevant today. When area detectors became more widespread, the imgCIF initiative addressed the metadata requirements for diffraction images. There was considerable effort to try to encourage the vendors to populate each frame with relevant metadata, but only partial success, and different vendors responded to different degrees and to different extent. While some standardisation is better than none, there is still scope for persuading the vendors to make more metadata available more easily. The IUCr’s Committee on Data (CommDat) is sponsoring a ‘checkCIF for raw data’ exercise that ha sthis as its goal.



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

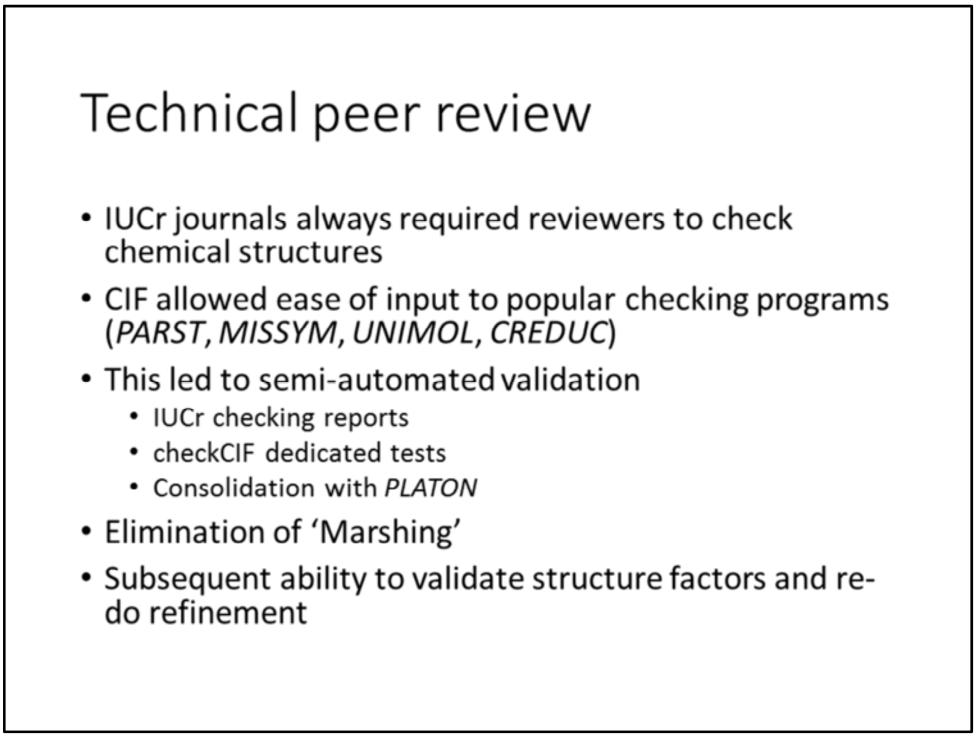
The Crystallographic Information File (*i.e.* the file format associated with CIF) has been reasonably successful, because in many ways it is fit for purpose. The exact purposes for which it was defined are described in the original paper [Hall, S. R., Allen, F. H. & Brown, I. D. (1991). The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Cryst*. A**47**, 655-685]. The continuing evolution of computer technology, coupled with new computational requirements, recently led to a revised specification [Bernstein, H. J., Bollinger, J. C., Brown, I. D., Gražulis, S., Hester, J. R., McMahon, B., Spadaccini, N., Westbrook, J. D. & Westrip, S. P. (2016). Specification of the Crystallographic Information File format, version 2.0. *J. Appl. Cryst*. **49**, 277-284], although this is not yet in widespread use. There is still some friction between the small-unit-cell and macromolecular CIF file formats, though this is arguably a feature of the underlying conceptual model at least as much as the formatting differences [Brink, A. & Helliwell, J. R. (2019). Why is interoperability between the two fields of chemical crystallography and protein crystallography so difficult? *IUCrJ* **6**, DOI:10.1107/S2052252519010972]. Nevertheless, it still delivers a lightweight, easily-editable medium for information exchange, and is easily convertible with other formats.



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

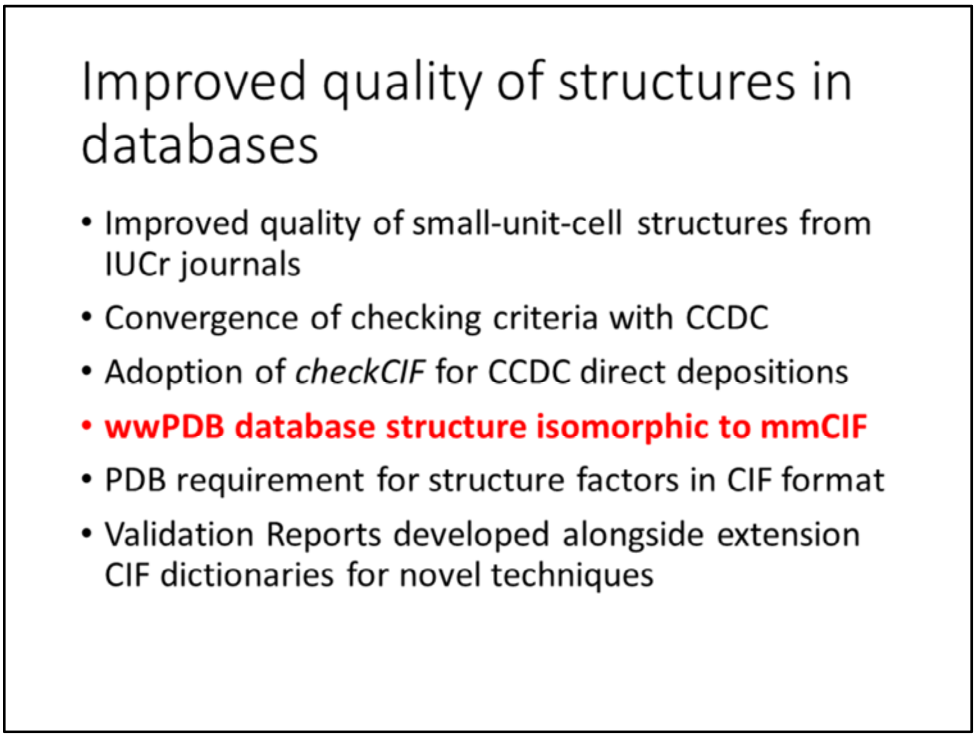
The adoption of CIF as a publication mechanism by IUCr journals was undoubtedly a very strong factor in its success. Authors were initially resistant to the new approach, just as people are generally resistant to innovation, but the pressure to publish overcame that resistance. Some young researchers, not born when CIF was published, can hardly imagine crystallography without CIF. However, it is noteworthy (and to me somewhat disappointing) that other publishers still do not see the fundamental difference between publishing an article derived from a data-rich well-structured file such as a CIF and publishing the conventional article with a semi-detached set of files of supporting information.



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

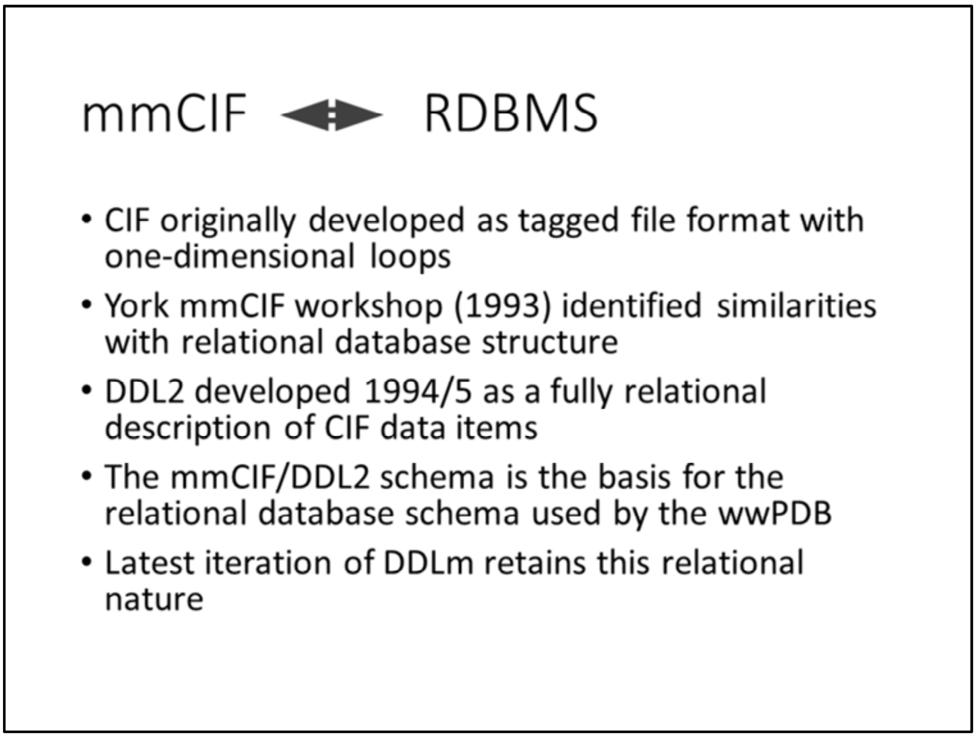
It was the burden of undertaking the technical review of the published structure that drove the introduction of the *checkCIF* service. While it has undoubtedly relieved most of the tedium of the reviewer, it has introduced the danger that ‘alerts’ are seen as determinative and not indicative. Reviewers can relax their guard if a *checkCIF* report does not have A alerts; conversely, authors can massage results to ‘get around’ the admonishments of *checkCIF*, as they are sometimes perceived.



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

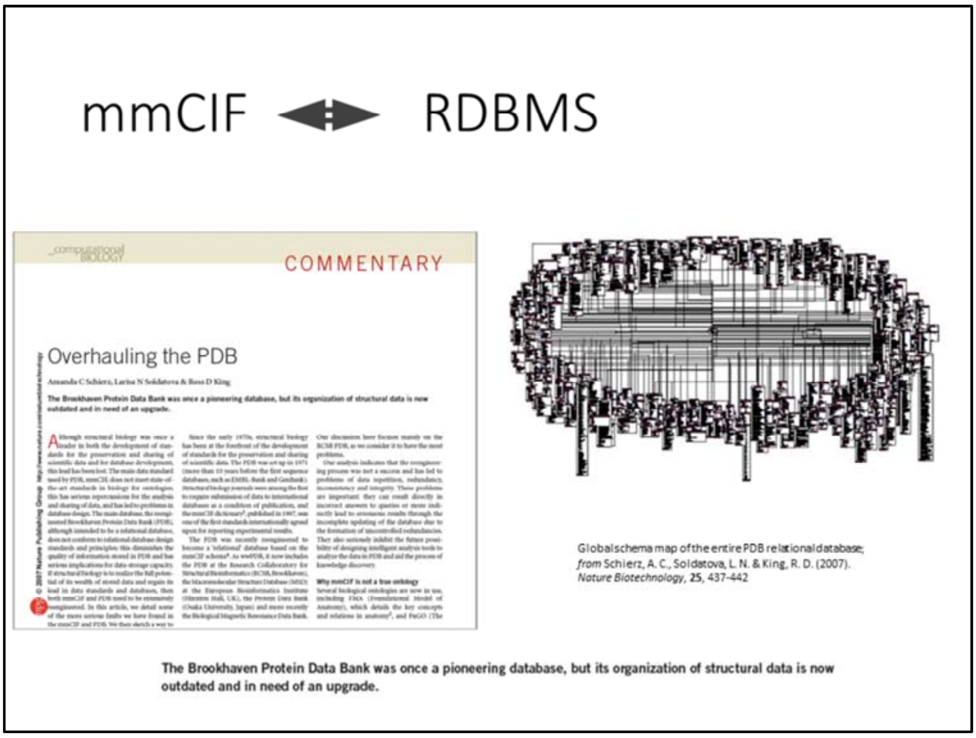
The improvement in quality of structures finding their way into small-unit-cell databases after *checkCIF* screening is apparent. However, a particular factor that I also want to highlight in the case of biological macromolecules is the importance of the mmCIF design as a relational data structure.



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

Following the brief history outlined in the slide, CIF has become very strongly a relational data model. This is satisfying, partly because relational databases are founded on strong basic mathematical principles [Date, C. J. (1995). *An Introduction to Database Systems*, 6th ed. Reading, MA: Addison-Wesley] and partly because the relational model can be transformed into different object representations by applying appropriate denormalisations [Hester, J. R. (2016). A robust, format-agnostic scientific data transfer framework. *Data Science Journal*, **15**:12, 1–17]. An immediate consequence of this is that a well-formed mmCIF file can be imported directly into the Protein Data Bank with full internal referential integrity. As well as keeping the database engineers happy, this ensures that the database entries are all consistent descriptions of macromolecules and their associate properties (such consistency does not of itself guarantee correctness, of course).



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[1], 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[2]. 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

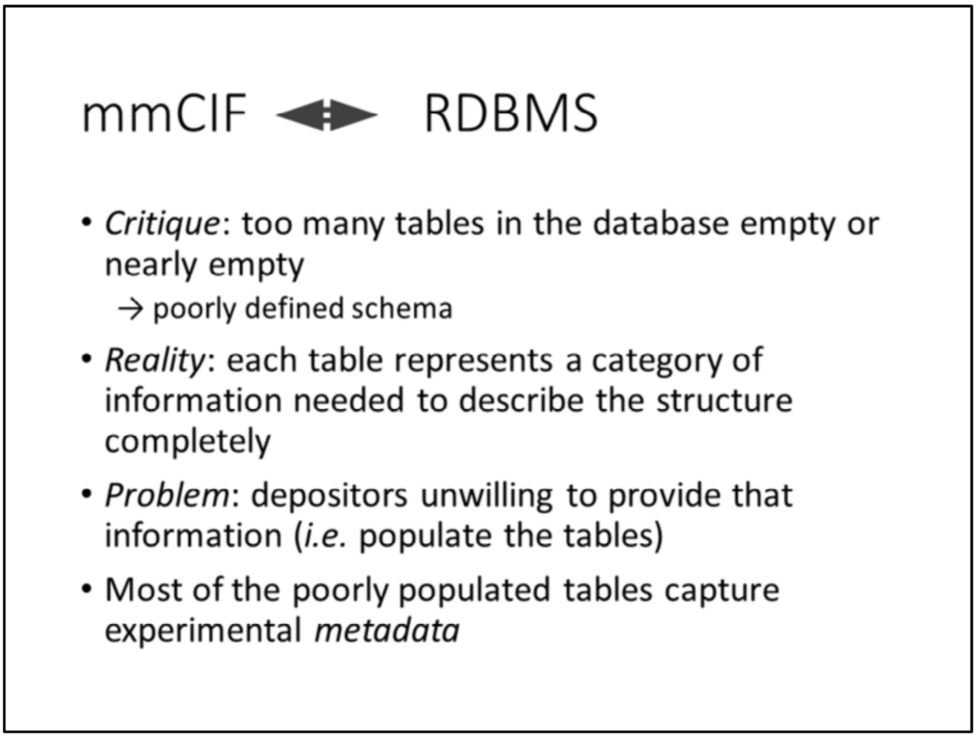
Several biological ontologies are now in use, including FMA (Foundational Model of Anatomy), which details the key concepts and relations in anatomy[3], and FuGO (The

© 2007 Nature Publishing Group http://www.nature.com/naturebiotechnology

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.

This slide is my favourite illustration of the need to bring different points of view to the analysis of a technical challenge. A 2007 critique published in *Nature Computational Biology* claimed the PDF was outdated and in need of an upgrade. The fact that the PDB continues to flourish more than a decade later suggests that it was able to survive this criticism.

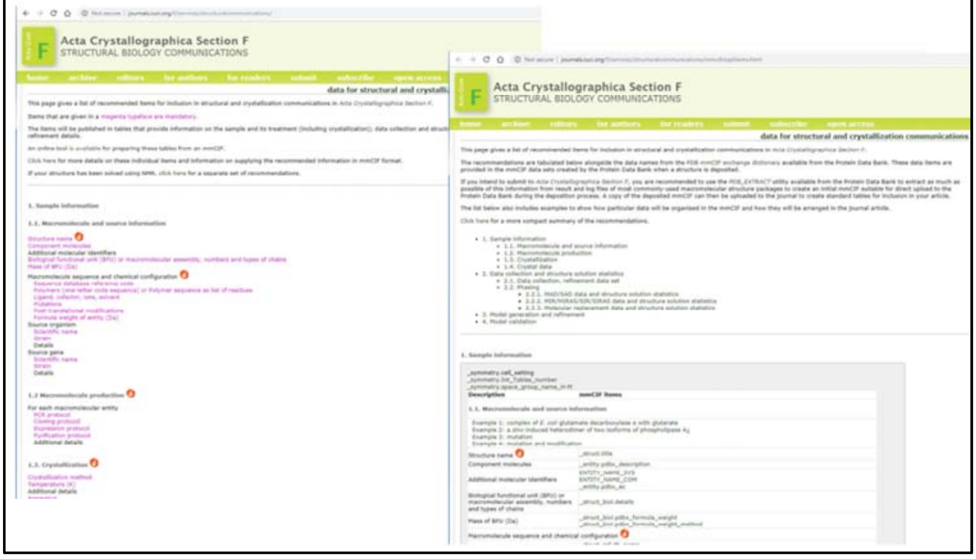


mmCIF ↔ RDBMS

- *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*

The criticism itself rested on the large number of nearly-empty tables, which the authors adduced as evidence for poor database design. Normally a well-normalised relational database would aim to have its separate tables well populated. While the poorly populated tables in the PDB might indeed have run against best practice in efficient space utilisation and data retrieval from a hypothetical database, I make the case that each table existed to capture specific information that structural biologists would like to make use of. The fact that they were poorly populated was a reflection that the community was unwilling to provide the desired information, either through lack of appreciation of the real value of having that data, or (more likely) because of the effort involved to do so, or the lack of tools to facilitate the desired data capture.

Metadata capture in macromolecular reports



Acta Crystallographica Section F
STRUCTURAL BIOLOGY COMMUNICATIONS

data for structural and crystallization communications

This page gives a list of recommended items for inclusion in structural and crystallization communications in *Acta Crystallographica Section F*.

Items that are given in a magenta typeface are mandatory.

The items will be published in tables that provide information on the sample and its treatment (including crystallization); data collection and structure refinement details.

An online tool is available for preparing these tables from an mmCIF.

Click here for more details on these individual items and information on supplying the recommended information in mmCIF format.

If your structure has been solved using NMR, click here for a separate set of recommendations.

1. Sample information

1.1. Macromolecule and source information

Structure name
Component molecules
Additional molecular identifiers
Biological functional unit (BFU) or macromolecular assembly, numbers and types of chains
Mass of BFU (Da)
Macromolecule sequence and chemical configuration
  Sequence database reference code
  Polymers (one-letter code sequence) or Polymer sequence as list of residues
  Ligand, cofactor, ions, solvent
  Mutations
  Post-translational modifications
  Formula weight of entity (Da)
Source organism
  Scientific name
  Strain
  Details
Source gene
  Scientific name
  Strain
  Details

1.2 Macromolecule production

For each macromolecular entity
  PCR protocol
  Cloning protocol
  Expression protocol
  Purification protocol
  Additional details

1.3. Crystallization

Crystallization method
Temperature (K)
Additional details

Acta Crystallographica Section F
STRUCTURAL BIOLOGY COMMUNICATIONS

data for structural and crystallization communications

This page gives a list of recommended items for inclusion in structural and crystallization communications in *Acta Crystallographica Section F*.

The recommendations are tabulated below alongside the data names from the PDB mmCIF exchange dictionary available from the Protein Data Bank. These data items are provided in the mmCIF data sets created by the Protein Data Bank when a structure is deposited.

If you intend to submit to *Acta Crystallographica Section F*, you are recommended to use the PDB_EXTRACT utility available from the Protein Data Bank to extract as much as possible of this information from result and log files of most commonly used macromolecular structure packages to create an initial mmCIF suitable for direct upload to the Protein Data Bank during the deposition process. A copy of the deposited mmCIF can then be uploaded to the journal to create standard tables for inclusion in your article.

The list below also includes examples to show how particular data will be organised in the mmCIF and how they will be arranged in the journal article.

Click here for a more compact summary of the recommendations.

- 1. Sample information
  - 1.1. Macromolecule and source information
  - 1.2. Macromolecule production
  - 1.3. Crystallization
  - 1.4. Crystal data
- 2. Data collection and structure solution statistics
  - 2.1. Data collection, refinement data set
  - 2.2. Phasing
    - 2.2.1. MAD/SAD data and structure solution statistics
    - 2.2.2. MIR/MIRAS/SIR/SIRAS data and structure solution statistics
    - 2.2.3. Molecular replacement data and structure solution statistics
- 3. Model generation and refinement
- 4. Model validation

1. Sample information

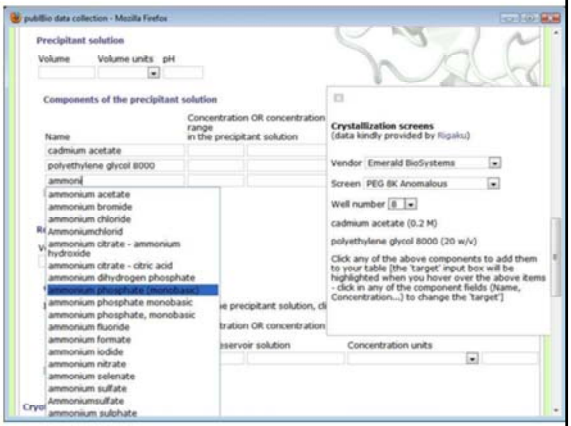
_symmetry.cell_setting
_symmetry.Int_Tables_number
_symmetry.space_group_name_H-M

Description	mmCIF items
1.1. Macromolecule and source information	
Example 1: complex of *E. coli* glutamate decarboxylase α with glutarate Example 2: a zinc induced heterodimer of two isoforms of phospholipase A Example 3: mutation Example 4: mutation and modification	
Structure name	_struct.title
Component molecules	_entity.pdbx_description
Additional molecular identifiers	ENTITY_NAME_SYS ENTITY_NAME_COM _entity.pdbx_ec
Biological functional unit (BFU) or macromolecular assembly, numbers and types of chains	_struct_biol.details
Mass of BFU (Da)	_struct_biol.pdbx_formula_weight _struct_biol.pdbx_formula_weight_method
Macromolecule sequence and chemical configuration	

IUCr journals sought to help out in this regard. While Howard Einspahr and Manfred Weiss were Editors of *Acta Cryst. F* they drew up a detailed list of the information that would be desirable in providing a complete description of a macromolecular structure, including details of the diffraction experiment, phasing strategy, sample preparation and crystallization, among others. The slide shows the two lists that they prepared, one as a textual listing [https://journals.iucr.org/f/services/structuralcommunications/], the other as a parallel list of mmCIF data items providing this information [https://journals.iucr.org/f/services/structuralcommunications/mmcifreqditems.html]. In principle this could form the basis for a *checkCIF* for protein structures. In practice, it is still difficult to persuade authors to provide this level of detail, and the existence of these lists is no longer explicit in the journals' *Notes for Authors*.

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*?



publBio data collection - Mozilla Firefox

Precipitant solution

Volume | Volume units | pH

Components of the precipitant solution

Name	Concentration OR concentration range in the precipitant solution
cadmium acetate	
polyethylene glycol 8000	
ammoni	

ammonium acetate
ammonium bromide
ammonium chloride
Ammoniumchlorid
ammonium citrate - ammonium hydroxide
ammonium citrate - citric acid
ammonium dihydrogen phosphate
ammonium phosphate (monobasic)
ammonium phosphate monobasic
ammonium phosphate, monobasic
ammonium fluoride
ammonium formate
ammonium iodide
ammonium nitrate
ammonium selenate
ammonium sulfate
Ammoniumsulfate
ammonium sulphate

Crystallization screens
(data kindly provided by Rigaku)

Vendor: Emerald BioSystems

Screen: PEG 8K Anomalous

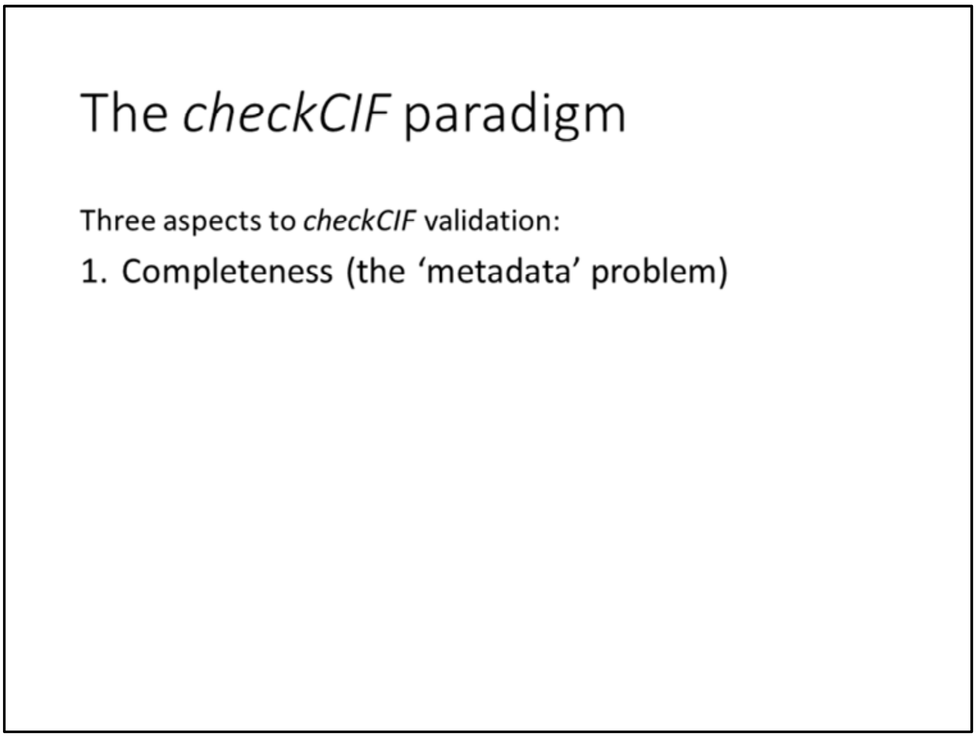
Well number: 8

cadmium acetate (0.2 M)

polyethylene glycol 8000 (20 w/v)

Click any of the above components to add them to your table [the 'target' input box will be highlighted when you hover over the above items - click in any of the component fields (Name, Concentration...) to change the 'target']

The *publBio* annotator is a tool for authoring macromolecular structure articles or crystallization papers. It is currently somewhat hidden from view (select “about” on the *publBio* home page to find a link to it, or use directly the URL http://publbio.iucr.org/publbio/publbio.php?version=1) and development has been frozen. However, it offers greater functionality in retrieving data from a fully populated mmCIF input file, and for populating crystallization detail by lookup of online databases. It involves building a publication by working through a number of distinct pages (‘macromolecule’, ‘crystallization’, ‘crystal data’, ‘data collection’, ‘data-collection statistics’, ‘phasing’ and ‘refinement’) which correspond to the underlying mmCIF dictionary categories or category groups. Once mastered, it offers an efficient and relatively painless way to populate a full PDB entry, but at present there is still too great an initial learning curve to attract authors used to the conventional way of preparing an article. I hope that the IUCr will continue to offer it, and in future renew its development efforts, as authors come to understand its benefits.

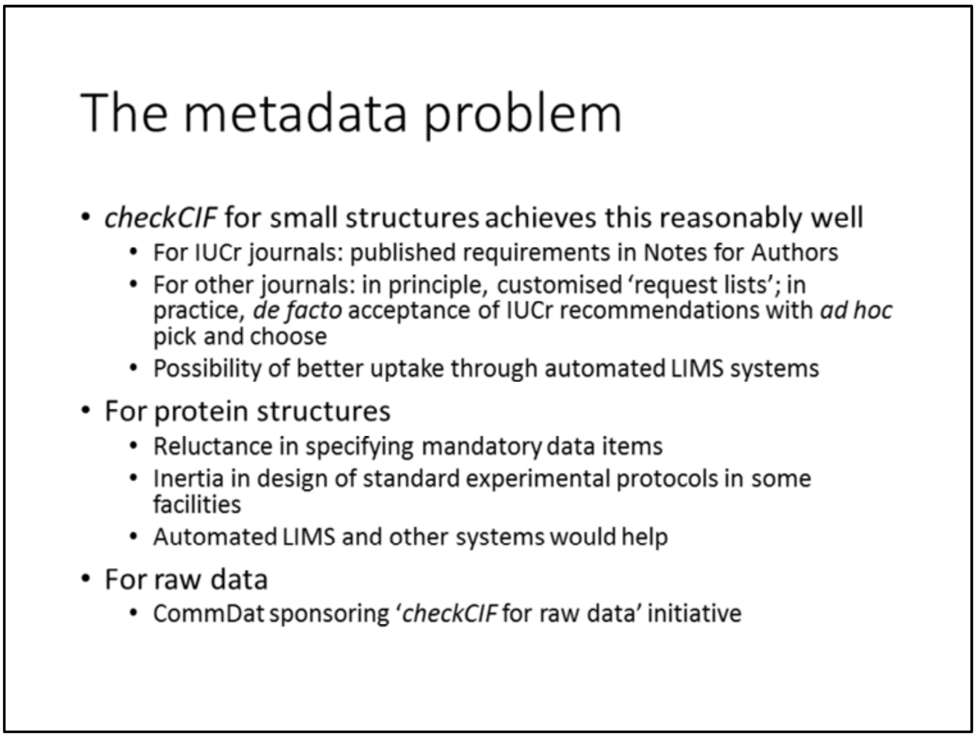


The *checkCIF* paradigm

Three aspects to *checkCIF* validation:

1. Completeness (the ‘metadata’ problem)

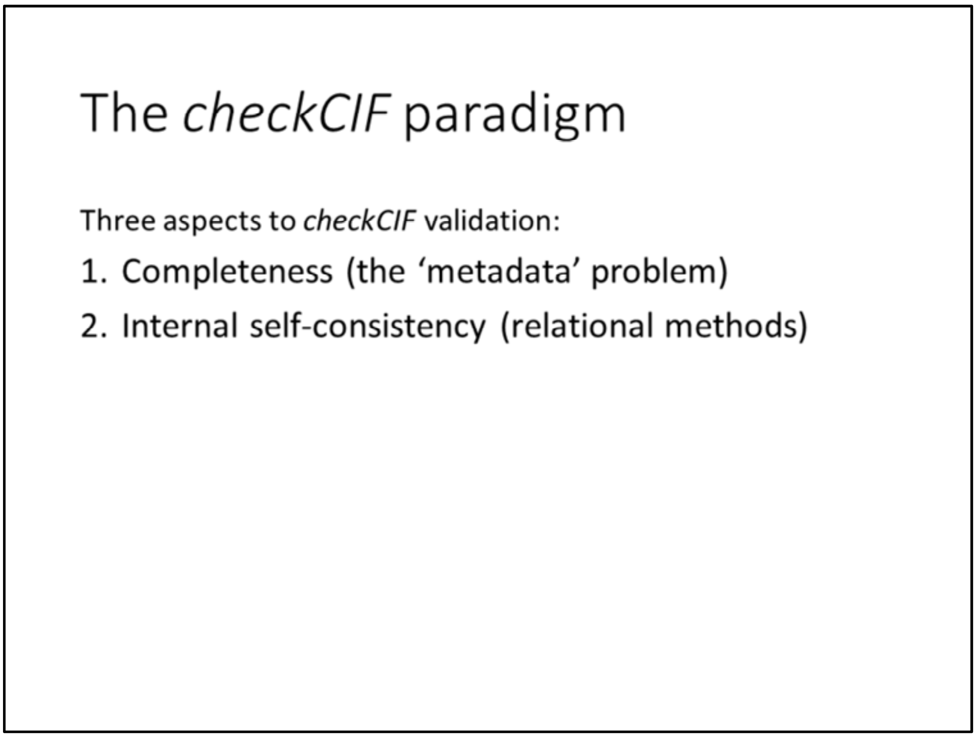
So we reiterate the threefold nature of the *checkCIF* paradigm in the light of the material covered in the presentation.



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

Metadata capture remains a significant challenge if one is to be able fully to reproduce an experiment or an author's treatment of the data collected from an experiment.

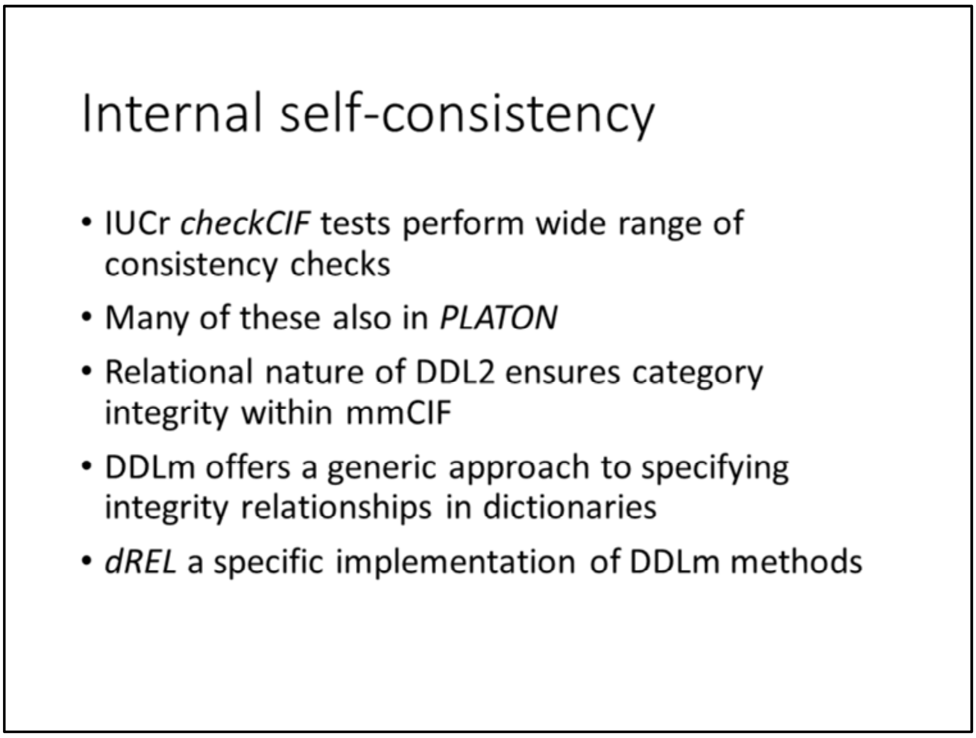


The *checkCIF* paradigm

Three aspects to *checkCIF* validation:

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

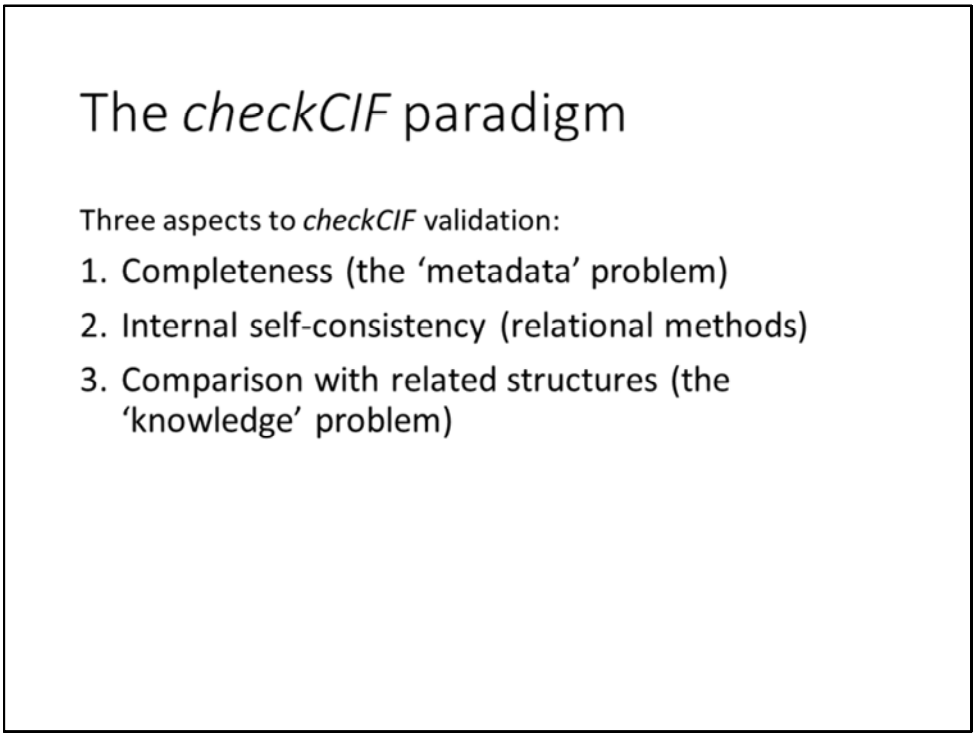
Internal consistency is what I really meant by the one word ‘correctness’ in the opening slide.



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

Checks on internal consistency are facilitated by the relational nature of CIF. Structures in the PDB (and therefore their mmCIF ‘mirror’ instances) are guaranteed to have the referential integrity of a relational database. Manual checks built from the definitions in the CIF dictionaries can, in future, be replaced by enhanced definitions that contain algorithmic methods stating the relationships beteen different data items in a manner that can be evaluated by generic mechanical programs.

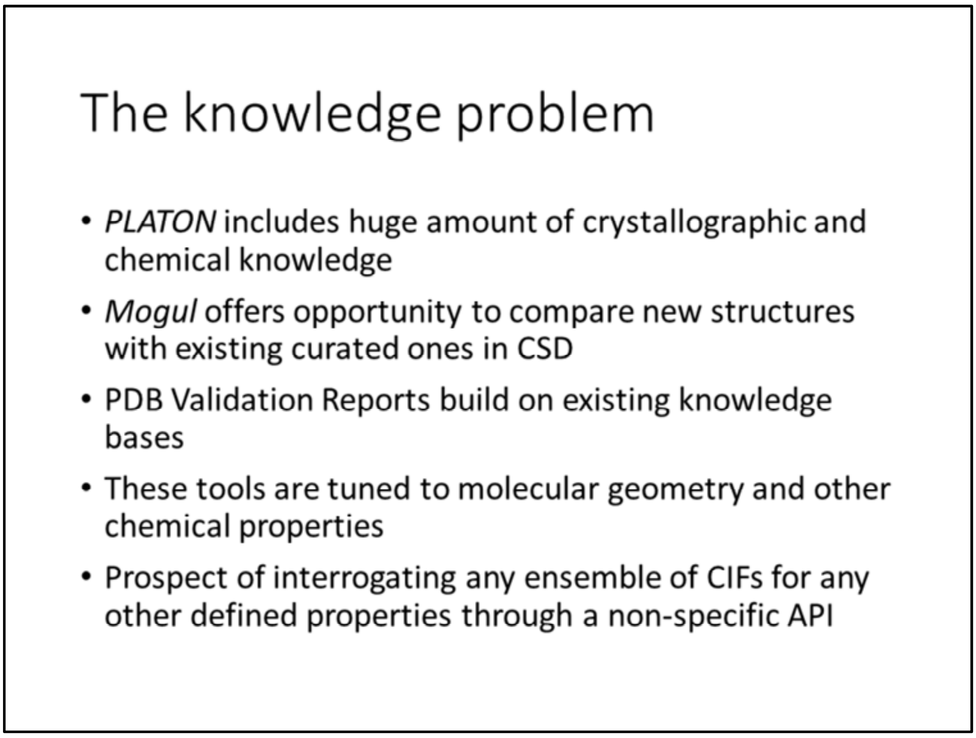


The *checkCIF* paradigm

Three aspects to *checkCIF* validation:

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

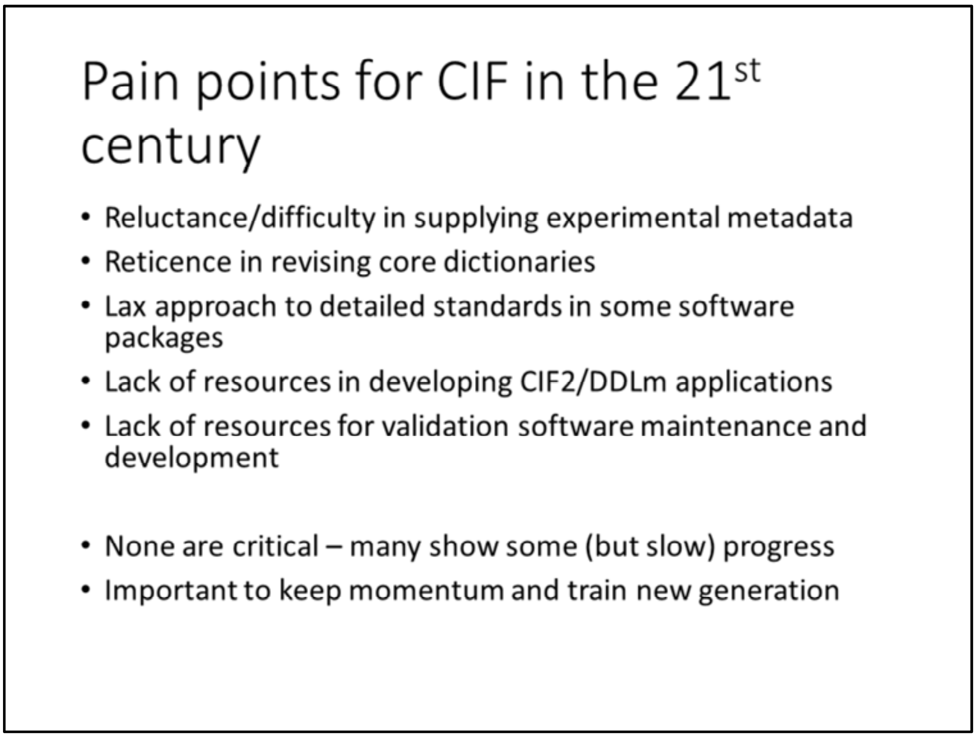
And my one-word ‘context’ resolves itself to a comparison with similar structures in the wider universe of knowledge.



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

This is where *checkCIF* is most contentious – many alerts refer to outliers from a statistical aggregate of values. They are not necessarily ‘wrong’ (though we can expect in future that machine-assisted learning in collaboration with the ‘complete’ and ‘correct’ tests of *checkCIF* might be better able to produce such a judgement, at least in some cases). I also raise here the notion that work on generic APIs interfacing with collections of CIF data files (the initial pilot will use the IUCr journals corpus) will allow *ad hoc* distributed databases to spring into life as CIFs become populated with new data items (*e.g.* describing non-crystallographic physical or chemical properties). Once again, machine learning and artificial intelligence could lead through this mechanism to novel research strategies.



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

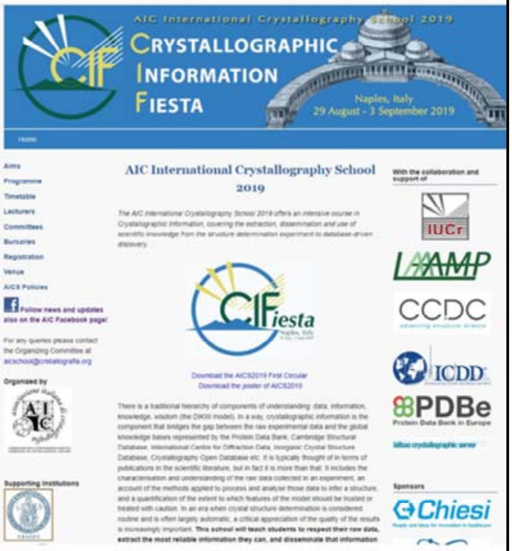
In a presentation to a general (*e.g.* CODATA) audience, I would end on that high note. But amongst friends, it’s worth reflecting on the shortcomings both within the CIF specifications and their practical implementation. Compared with many disciplines, crystallography has achieved a great deal. But it is in danger of reaching a plateau (especially as the generation that brought CIF to the community fades into retirement) unless new blood can be stirred.

“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.”



AIC International Crystallography School 2019

CRYSTALLOGRAPHIC INFORMATION FIESTA

Naples, Italy
29 August - 3 September 2019

Home

Aims
Programme
Timetable
Lecturers
Committees
Bursaries
Registration
Venue
AICS Policies

Follow news and updates also on the AIC Facebook page!

For any queries please contact the Organizing Committee at aicschool@cristallografia.org

Organized by

Supporting Institutions

AIC International Crystallography School 2019

The AIC International Crystallography School 2019 offers an intensive course in Crystallographic Information, covering the extraction, dissemination and use of scientific knowledge from the structure determination experiment to database-driven discovery.

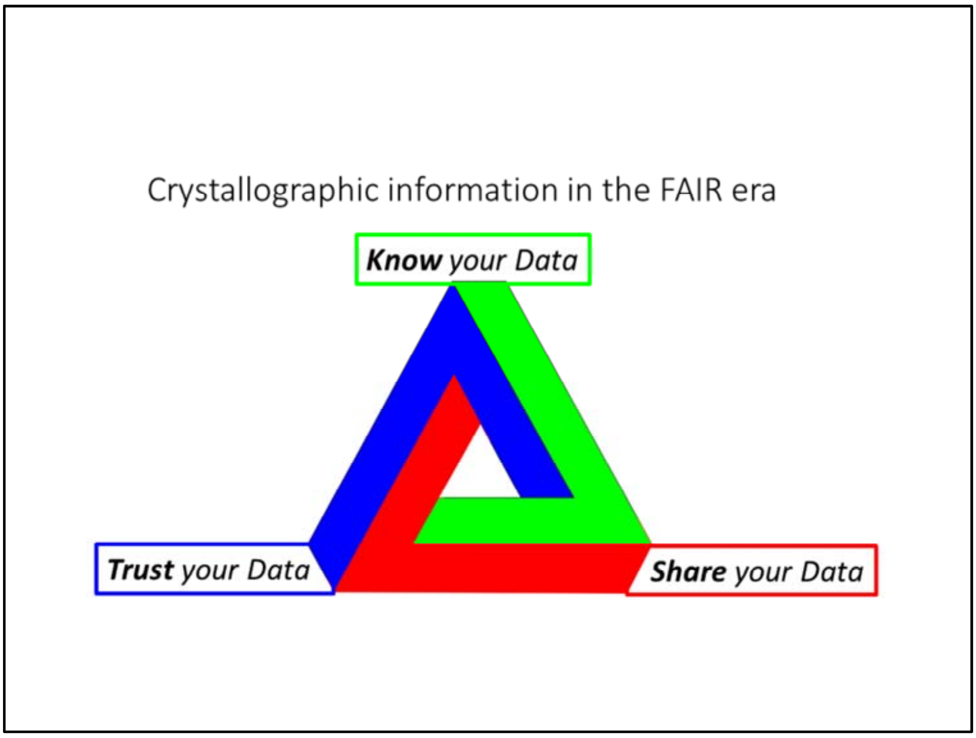
Download the AICS2019 First Circular
Download the poster of AICS2019

There is a traditional hierarchy of components of understanding: data, information, knowledge, wisdom (the DIKW model). In a way, 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. It is typically thought of in terms of publications in the scientific literature, but in fact it is more than that. It includes the characterisation and understanding of the raw data collected in an experiment, an account of the methods applied to process and analyse those data to infer a structure, and a quantification of the extent to which features of the model should be trusted or treated with caution. In an era when crystal structure determination is considered routine and is often largely automatic, a critical appreciation of the quality of the results is increasingly important. **This school will teach students to respect their raw data, extract the most reliable information they can, and disseminate that information**

With the collaboration and support of

Sponsors

And so we return to the starting quotation, and show that it is part of a prospectus for the first School on Crystallographic Information, which, if successful, might just spark the enthusiasm of the younger generation.



Crystallographic information in the FAIR era

***Know** your Data*

***Trust** your Data*

***Share** your Data*

And the take-home message of that School is encapsulated in the final graphic – an exhortation to understand what today’s ‘black-box’ instrumentation is actually doing with data, to be sure that you believe that analysis because you have worked through it critically with your own eyes, and that only then can you be confident of disseminating the results of your research as trustworthy. This philosophy should be held by all scientists. As crystallographers, we are fortunate in having CIF to make it a smoother process.