

# Crystallographic Information and Data Management

A Satellite Workshop to the 28th European  
Crystallographic Meeting

## A Coherent Information Flow in Crystallography

Brian McMahon



International Union of Crystallography  
5 Abbey Square  
Chester CH1 2HU  
UK  
[bm@iucr.org](mailto:bm@iucr.org)



# Science

From Latin *scientia*: 'knowledge'

Exploration

Discovery

Interpretation

Inspiration

**Information**

**Data**

Facts

Knowledge

Communication

Understanding



# International Union of Crystallography

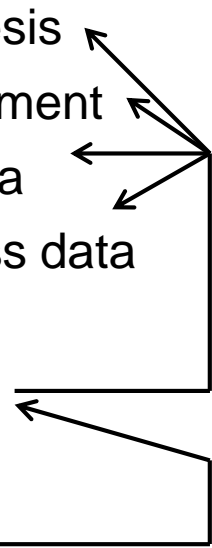
A commitment to scientific communication

1948 – IUCr founded; *Acta Crystallographica* launched; *Structure Reports* launched; Commission on Crystallographic Nomenclature; Commission on Crystallographic Data  
1952 – International Tables for X-ray Crystallography published  
1967 – Union Member of CODATA (founded in 1966)  
1970s – Commission on Crystallographic Data reports to Executive Committee on structural databases: Powder Diffraction File (founded 1938); Cambridge Structural Database (1965); Protein Data Bank (1971); Nucleic Acid Database (1991); Inorganic Crystal Structure Database (1978); CRYSTMET (1974) etc.  
1978 – Computing and Data Commissions call for a standard file structure (SCFS)  
1987 – Executive Committee calls for new standard to allow electronic submission to journals  
1991 – CIF format adopted; used in journal submissions and database (CSD) import  
1993 – COMCIFS founded to maintain the CIF standard  
2011 – Working Group convened by Executive to consider routine deposition of diffraction images and other raw data sets

# A paradigm for scientific communication

## General

Frame hypothesis  
Perform experiment  
Collect raw data  
Reduce/process data  
Derive model  
Validate model  
Submit paper  
Peer review  
Publish article  
(Archive/disseminate data)



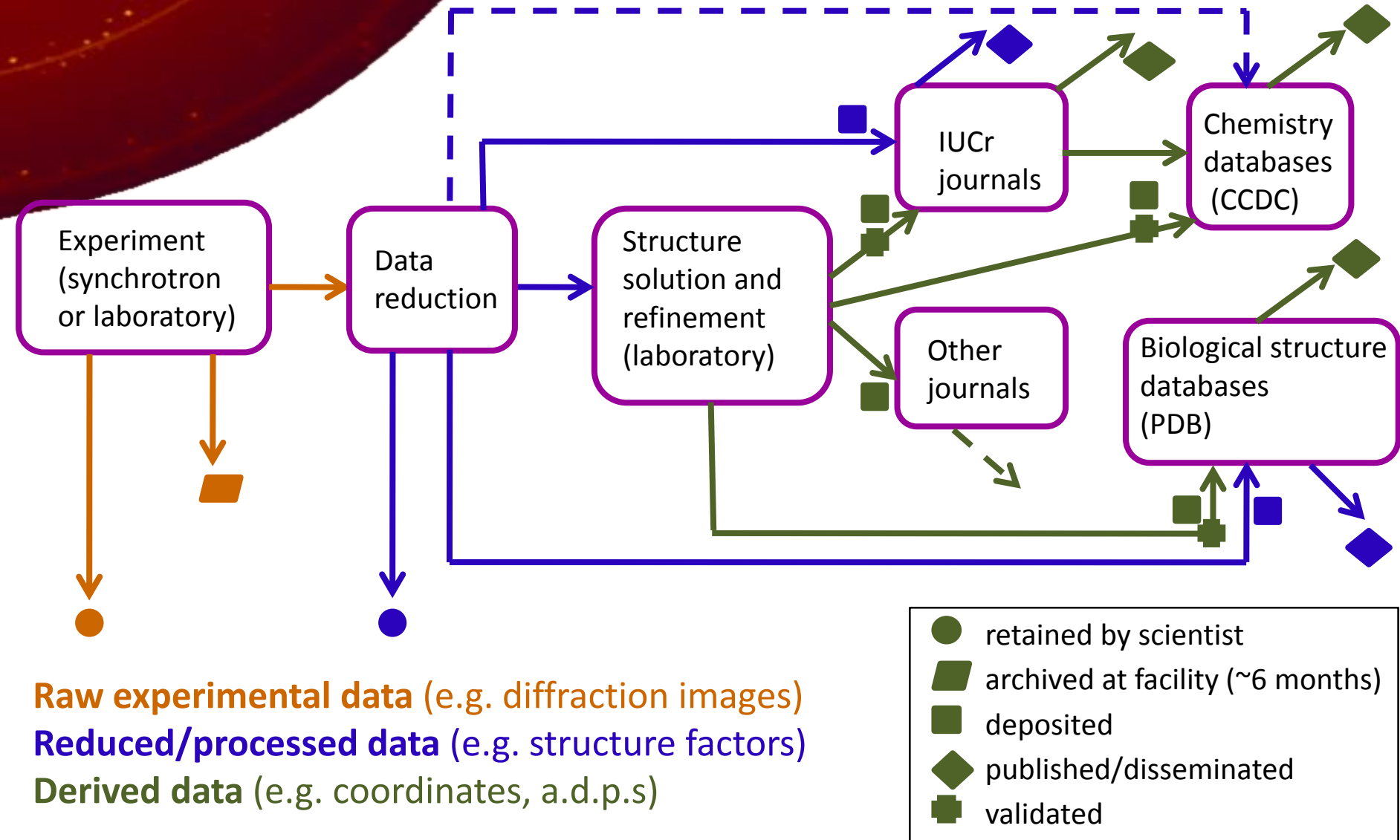
```
graph TD; A[Frame hypothesis] --> B[Perform experiment]; B --> C[Collect raw data]; C --> D[Reduce/process data]; D --> E[Derive model]; E --> F[Validate model]; F --> G[Submit paper]; G --> H[Peer review]; H --> I[Publish article]; I --> J["(Archive/disseminate data)"];
```

## X-ray crystallography

Structure/function e.g. pharma  
X-ray diffraction from single crystal/powder

- **X-ray diffraction images (~ 1 GB)**
- **Structure factors (~ 1-10 MB)**
- **Structure solution/refinement packages**
- ***PLATON/checkCIF***
- **Submit paper in CIF format**
- **Includes checkCIF/database searches**
- **As PDF/rich HTML, with CIF, s.f.s**
- **(imgCIF), coreCIF/mmCIF, PDB, CSD**

# Data flow in crystallography





# Crystallographic Information Framework

A unifying set of ideas for the definition and  
exchange of crystallographic data

1991 – Crystallographic Information File format

# Crystallographic Information Framework

## Crystallographic Information File format

655

*Acta Cryst.* (1991). **A47**, 655–685

**International Union of Crystallography**

**Commission on Crystallographic Data**

**Commission on Journals**

**Working Party on Crystallographic Information**

**The Crystallographic Information File (CIF): a New Standard  
Archive File for Crystallography\***

BY SYDNEY R. HALL

*Crystallography Centre, University of Western Australia, Nedlands 6009, Australia*

FRANK H. ALLEN

*Crystallographic Data Centre, University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, England*

AND I. DAVID BROWN

*Institute for Materials Research, McMaster University, Hamilton, Ontario L8S 4M1, Canada*

(Received 8 April 1991; accepted 28 June 1991)

### Abstract

The specification of a new standard Crystallographic Information File (CIF) is described. Its development is based on the Self-Defining Text Archive and Retrieval (STAR) procedure [Hall (1991). *J. Chem. Inf. Comput. Sci.* **31**, 326–333]. The CIF is a general, flexible and easily extensible free-format archive file; it is human and machine readable and can be edited by a simple

### Introduction

There is an increasing need in many branches of science for a uniform but flexible method of archiving and exchanging data in electronic form. Rapid advances in computer technology, coupled with the expansion of local, national and international networks, have fuelled the need for such a facility. The variety and relative inflexibility of existing data exchange formats have inhibited their effective use. This note presents a field-based technique

# Crystallographic Information Framework

## Crystallographic Information File format

```
data_I
_chemical_name_systematic      'Biphenyl-2,4,4',6-tetracarboxylic acid monohydrate'
_chemical_formula_moiety       'C16 H10 O8, H2 O'
_chemical_formula_sum          'C16 H12 O9'
_chemical_formula_weight       348.26
_symmetry_cell_setting         monoclinic
_symmetry_space_group_name_H-M 'P 21/c'
_symmetry_space_group_name_hall '-p 2ybc'
loop_
  _symmetry_equiv_pos_as_xyz
    'x, y, z'      '-x, y+1/2, -z+1/2'      '-x, -y, -z'      'x, -y-1/2, z-1/2'
_cell_length_a      5.638(4)
_cell_length_b      16.160(11)
_cell_length_c      16.798(12)
_cell_angle_alpha    90.00
_cell_angle_beta     92.524(12)
_cell_angle_gamma    90.00
_cell_volume         1528.9(19)
```





# Crystallographic Information Framework

A unifying set of ideas for the definition and  
exchange of crystallographic data

1991 – Crystallographic Information File format

1991 – Data dictionaries separating semantics from syntax

# Crystallographic Information Framework

Data dictionaries separating semantics from syntax

```
data_refl_n_phase_calc
  _name                '_refln_phase_calc'
  _category            refln
  _type               numb
  _list               yes
  _list_reference     '_refln_index_'
  _units              deg
  _units_detail       'degrees'
  _definition
;                      The calculated structure-factor phase in degrees.
;
```



# Crystallographic Information Framework

A unifying set of ideas for the definition and exchange of crystallographic data

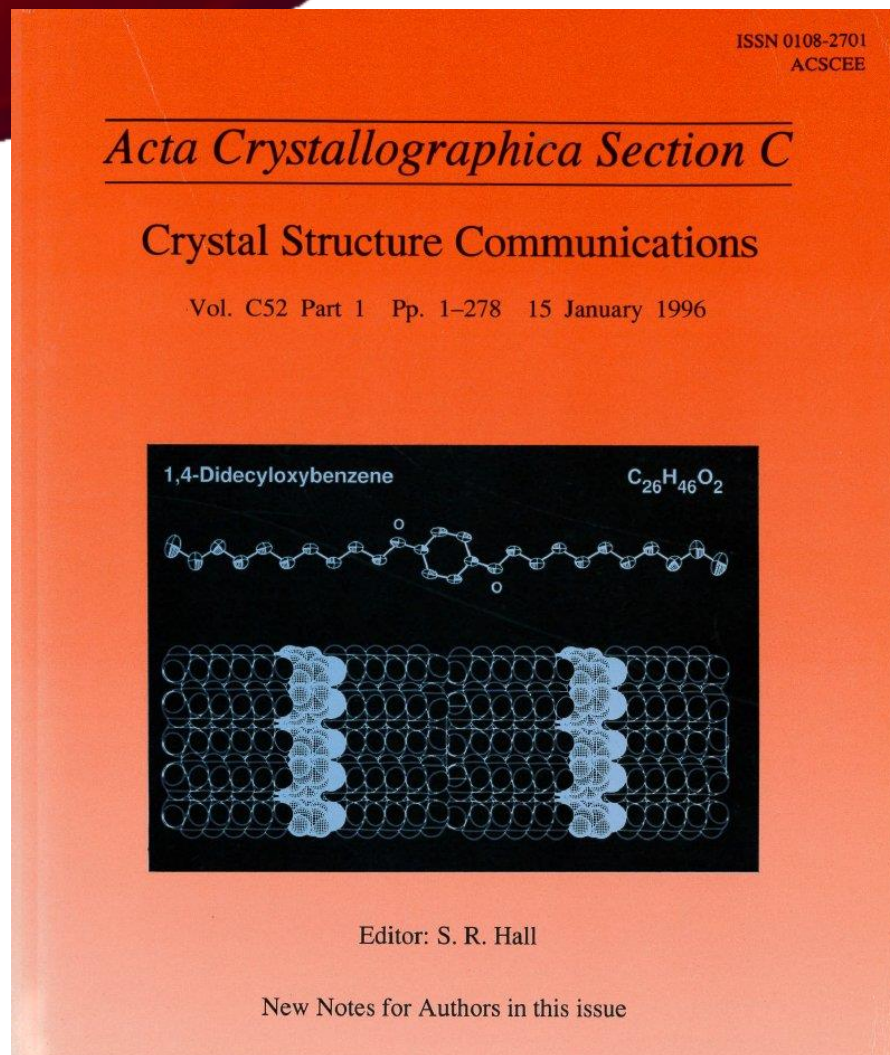
1991 – Crystallographic Information File format

1991 – Data dictionaries separating semantics from syntax

1996 – CIF mandatory submission format for *Acta C*

# Crystallographic Information Framework

CIF mandatory submission format for *Acta Crystallographica Section C*



*Acta Cryst.* (1996). **C52**, xvii–xviii

## Editorial

This year will see several important changes to the publication and presentation of *Acta Crystallographica Section C*. These include the mandatory requirement that submissions be electronic, the availability on the Internet of *Contents* information for each issue and the introduction of a listing of unpublished structures deposited in the Cambridge database. I would like to use this opportunity to give the reasons for introducing these changes, as they will significantly affect both readers and authors of this journal.



# Crystallographic Information Framework

A unifying set of ideas for the definition and exchange of crystallographic data

1991 – Crystallographic Information File format

1991 – Data dictionaries separating semantics from syntax

1996 – CIF mandatory submission format for *Acta C*

1997-8 – PDB management by RCSB; mmCIF

# Crystallographic Information Framework

Research Collaboratory for Structural Biology takes over management of Protein Data Bank and re-engineers database using the macromolecular CIF (mmCIF) schema

The screenshot shows the Protein Data Bank (PDB) website interface. The top navigation bar includes the PDB logo, a search bar, and a link to 'PDB-101'. The main content area features a chart titled 'Yearly Growth of Total Structures' which displays the number of structures deposited each year from 1996 to 2013. The chart shows a steady increase in the number of structures over time, with a significant jump in 2013. The left sidebar contains various links and resources, including 'PDB-101', 'MyPDB', 'Home', 'Deposition', 'Tools', and 'Help'.

| Year | Number of Structures |
|------|----------------------|
| 1996 | ~1,000               |
| 1997 | ~2,000               |
| 1998 | ~3,000               |
| 1999 | ~4,000               |
| 2000 | ~5,000               |
| 2001 | ~6,000               |
| 2002 | ~7,000               |
| 2003 | ~8,000               |
| 2004 | ~9,000               |
| 2005 | ~10,000              |
| 2006 | ~11,000              |
| 2007 | ~12,000              |
| 2008 | ~13,000              |
| 2009 | ~14,000              |
| 2010 | ~15,000              |
| 2011 | ~16,000              |
| 2012 | ~17,000              |
| 2013 | ~18,000              |



Since 2003 the Worldwide Protein Data Bank consortium has been synchronising data using the PDBML/XML format, which is also based on the mmCIF ontology





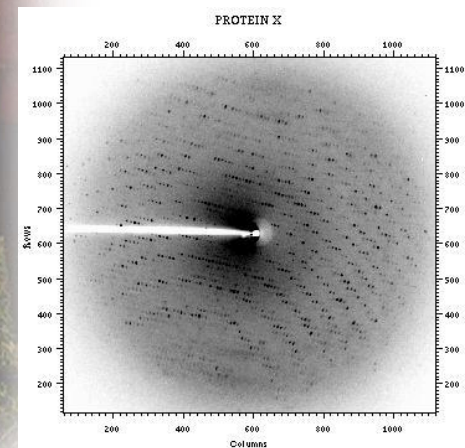
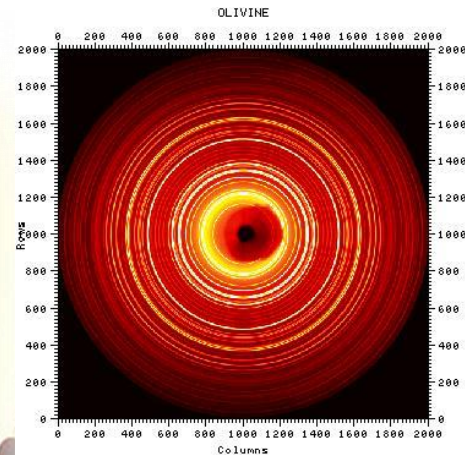
# Crystallographic Information Framework

A unifying set of ideas for the definition and exchange of crystallographic data

- 1991 – Crystallographic Information File format
- 1991 – Data dictionaries separating semantics from syntax
- 1996 – CIF mandatory submission format for *Acta C*
- 1997-8 – PDB management by RCSB; mmCIF
- 2000 – imgCIF/CBF formats for X-ray diffraction images

# Crystallographic Information Framework

imgCIF/CBF formats for X-ray diffraction images







# Crystallographic Information Framework

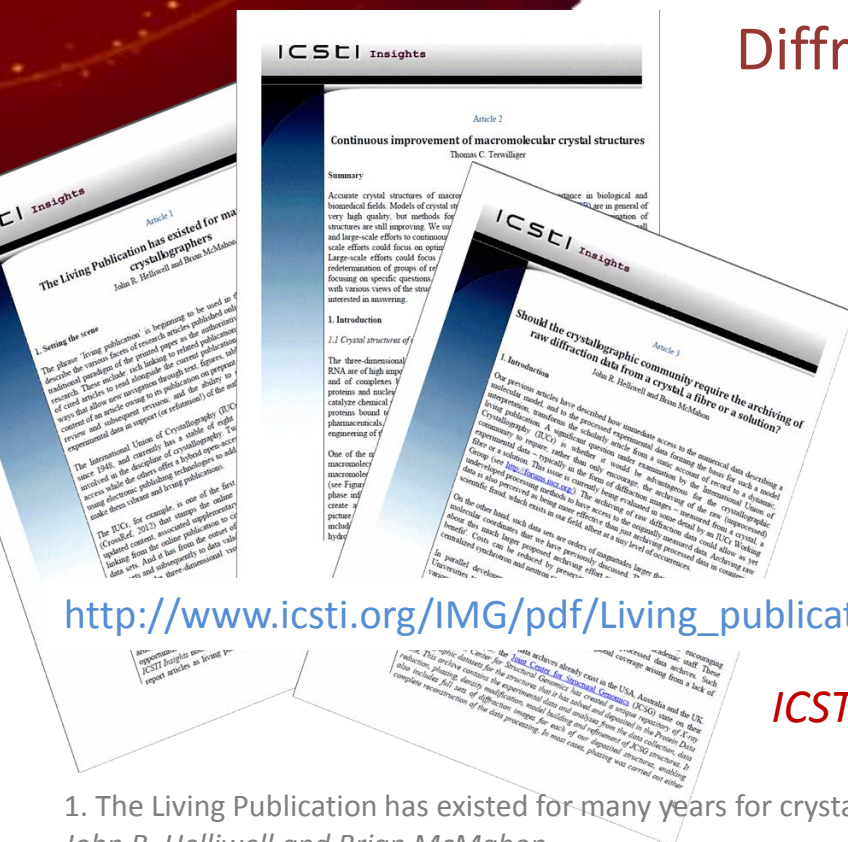
A unifying set of ideas for the definition and exchange of crystallographic data

- 1991 – Crystallographic Information File format
- 1991 – Data dictionaries separating semantics from syntax
- 1996 – CIF mandatory submission format for *Acta C*
- 1997-8 – PDB management by RCSB; mmCIF
- 2000 – imgCIF/CBF formats for X-ray diffraction images
- 2011 – Diffraction Data Deposition Working Group

# Crystallographic Information Framework

Diffraction Data Deposition Working Group.

Chair: John Helliwell



ICSTI Insights Series



1. The Living Publication has existed for many years for crystallographers

*John R. Helliwell and Brian McMahon*

2. Continuous improvement of macromolecular crystal structures

*Thomas C. Terwilliger*

3. Should the crystallographic community require the archiving of raw diffraction data from a crystal, a fibre or a solution?

*John R. Helliwell and Brian McMahon*

ECM27 Workshop, Bergen,  
Norway, 6 August 2012



# Crystallographic Information Framework

A unifying set of ideas for the definition and exchange of crystallographic data

- 1991 – Crystallographic Information File format
- 1991 – Data dictionaries separating semantics from syntax
- 1996 – CIF mandatory submission format for *Acta C*
- 1997-8 – PDB management by RCSB; mmCIF
- 2000 – imgCIF/CBF formats for X-ray diffraction images
- 2011 – Diffraction Data Deposition Working Group
- 2013 – DDLm Workshop, U. Warwick

# Crystallographic Information Framework

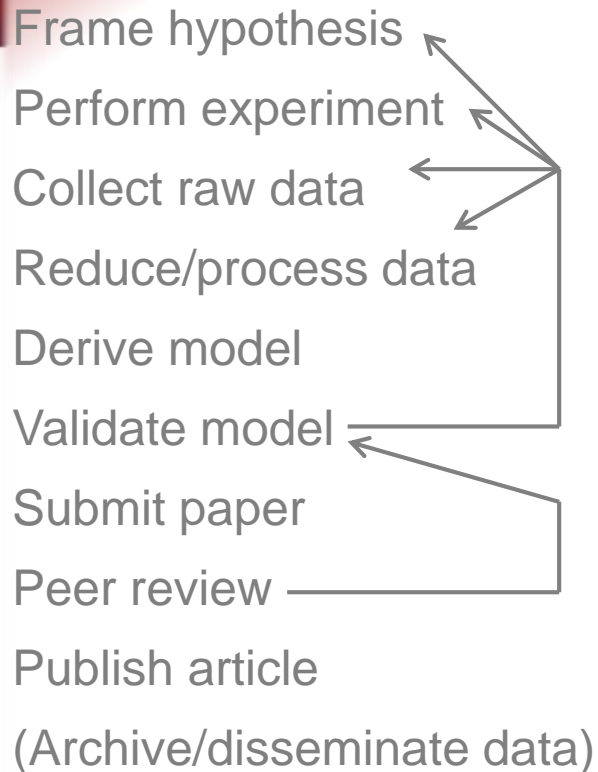
DDLm incorporates algorithmic *methods*

```
data_refl_n_phase_calc
  _name      '_refln_phase_calc'
  _category      refln
  _type      numb
  _list      yes
  _list_reference      '_refln_index_'
  _units      deg
  _units_detail      'degrees'
  _definition
; The calculated structure-factor
  phase in degrees.
;
```

```
save_refl_n_phase_calc
  _definition.id      '_refln_phase_calc'
  loop_
  _alias.definition_id      '_refln_phase_calc'
  _definition.update      2013-04-27
  _description.text
;
  The phase of the calculated structure-factor.
;
  _name.category_id      refln
  _name.object_id      phase_calc
  _type.purpose      Measurand
  _type.source      Derived
  _type.container      Single
  _type.contents      Real
  _enumeration.range      0.:360.
  _units.code      degrees
  loop_
  _method.purpose
  _method.expression
  Evaluation
;
  phase = Atan2d ( _refln.B_calc, _refln.A_calc )
  If(phase < 0.) _refln.phase_calc = phase + 360.
  Else      _refln.phase_calc = phase
;
  save_
```

# Pushing back the boundaries

## General



## *X-ray crystallography*

- **Structure/function e.g. pharma**
- **X-ray diffraction: single crystal/powder**
- **X-ray diffraction images (~ 1 GB)**
- **Structure factors (~ 1-10 MB)**
- **Structure solution/refinement packages**
- ***PLATON/checkCIF***
- **Submit paper in CIF format**
- **Includes checkCIF/database searches**
- **As PDF/rich HTML, with CIF, s.f.s**
- **imgCIF, coreCIF/mmCIF, PDB, CSD**



# Acknowledgements

The activities described in the talks in today's Symposium  
owe much to many collaborators over the years:

Alan Mighell, Alex Renshaw, Alexei Vagin, Allen Larson, Alun Ashton, Andre Authier, Andy Hammersley, Andy Howard, Arie Van Der Lee, Ashley Buckle, Ben Watts, Bill Clegg, Bob Hanson, Bob Sweet, Brian Matthews, Brian Toby, Charlie Bugg, Chris Nielsen, Colin Groom, Curt Haltiwanger, Dale Tronrud, Dave Duchamp, Dave Stampf, David Brown, David Watkin, David Watson, Doug Du Boulay, Doug Greer, Eldon Ulrich, Eleanor Dodson, Enrique Abola, Eric Gabe, Erica Yang, Ethan Merritt, Frances Bernstein, Frank Allen, George Ferguson, George Sheldrick, Gerard Bricogne, Gerard Kleywegt, Gotzon Madariaga, Greg Shields, Gunter Bergerhoff, Helen Berman, Herbert Bernstein, Howard Einspahr, Howard Flack, I. David Brown, Ian Bruno, James Hester, Jan Zelinka, Jean Richelle, John Huffman, Jim Kaduk, Joe Krahn, Joel Sussman, John Bollinger, John Helliwell, John Westbrook, Keith Watenpaugh, Kim Henrick, Lachlan Cranswick, Liz Lyon, Liz Potterton, Lynn Ten Eyck, Manfred Weiss, Mario Nardelli, Mark Koennecke, Martyn Winn, Matt Towler, Michael Scharf, Mike Dacombe, Mike Hoyland, Mike Hursthouse, Mois Aroyo, Nick Day, Nick England, Nick Spadaccini, Owen Johnson, Paul Edgington, Paul Mallinson, Paula Fitzgerald, Peter Grey, Peter Keller, Peter Murray-Rust, Peter Strickland, Phil Bourne, Phil Coppens, Ralf Grosse-Kunstleve, Richard Ball, Robert Downs, Sameer Velankar, Sandy Blake, Saulius Grazulis, Shoshana Wodak, Sidney Abrahams, Simon Coles, Simon Hodson, Simon Parsons, Simon Westrip, Sine Larsen, Steve Androulakis, Steve Bryant, Syd Hall, Ted Maslen, Tom Koetzle, Tom Terwilliger, Ton Spek, Tony Linden, Vicky Karen, Vivian Stojanoff, Weider Chang, Wolfgang Bluhm, Yvon Le Page

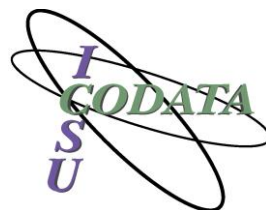
*... and many more besides*



# Crystallographic Information and Data Management

A Satellite Workshop to the 28th European Crystallographic Meeting

## Sponsors



DECTRIS®



WORLDWIDE  
**PDB**  
PROTEIN DATA BANK



WILEY

