Standards in scholarly crystallographic data publishing

Brian McMahon

Research & Development Officer
International Union of Crystallography
5 Abbey Square
Chester CH1 2HU

bm@iucr.org

Crystal Grid Workshop, U. Indiana
26 April 2007
“We are all publishers now”

But what does it take to make us *scholarly* publishers?

– Common interests
– Respect for provenance, ownership, rights, priority
– Quality control
– Peer review
– Effective organisation, discovery, archiving
Data that inform the discussion

Raw data
(image plate, diffractometer, film)

Primary data
(structure factors)

Derived data
(six-dimensional structural model)
Spectrum of crystal structure ‘publications’

1. Radical new knowledge of crystal/molecular structures (rare)
2. Radical new methods for solving/refining structures (rare)
3. Major contribution to understanding of chemical properties of one or a class of compounds (*Acta Cryst. Section B*)
4. Significant new information about a compound or group of related compounds (*Acta Cryst. Section C*)
5. Full description of a new structure (*Acta Cryst. Section E*)
6. Routine structure determination (*Acta E/eBank/chemistry journal*)
7. Partial/incomplete structure determination (chemistry journal/lab notebook)
Routine structure determination

- New format articles in *Acta Cryst. E*
  - High volume
  - Routine structures
  - Reduced peer review costs
  - Still subject to *checkCIF* analysis
  - Facilitate low-cost open-access publication
  - Full text and data available as supplementary document and in CIF
At the edge of the spectrum…

- Supplementary data in chemistry journals
- Voluntary deposit with curated databases
- Institutional repositories (universities)
- Voluntary open-access subject repositories
- Crystallography service facilities
- Industrial and pharmaceutical companies
- Deposit of structure factors
- Total loss
Data publication at source

Initiatives such as eBank are particularly valuable:

- Some prospect of longevity
- Use of common protocols/federation
- Address domain-specific concerns
- Large enough (as federated entities) to discuss special arrangements for archiving (including with publishers)
- Comprehensive within user base (does not rely on voluntary action)
- Quality control indicators
Extending the scholarly publication paradigm

- Standard data formats (CIF)
- OAI-PMH
- DOI, openURL
- Standard metadata
- Links to all data
- Links to publication
- Rights
- Quality (checkCIF)
Standards promoted by IUCr

• Data definition
  – CIF dictionaries

• Data context
  – Metadata, identifiers

• Data discovery
  – Metadata, openURL, OAI-PMH

• Data quality
  – checkCIF, DDLm
Data definition

• Small-unit-cell crystallography
  – Core CIF dictionary
  – *Acta Crystallographica C/E*
  – checkCIF
  – Links to CCDC, eBank

• Chemistry
  – coreCIFchem
  – InChI
  – MIF
  – CML
Data definition 2

• Biological macromolecular structures
  – mmCIF, pdbx, nmrSTAR
  – PDB
• Powder, modulated structures, electron density
  – pdCIF, msCIF, rhoCIF
• Small-angle scattering
  – sasCIF
• Neutron scattering
  – ?
Data definition 3

• CIF dictionaries
  – Crystallographic Information Framework
  – CIF/STAR file format
  – XML (choice of schemas)

```plaintext
data_pd_instr_beam_size_
loop_ _name       '_pd_instr_beam_size_ax'
                 '_pd_instr_beam_size_eq'
  _category        pd_data
  _type            numb
  _enumeration_range 0.0:
  _units          mm
  _units_detail   'millimetres'
  _definition ; Axial and equatorial dimensions of the radiation beam at the specimen
              position (in millimetres). The perpendicular to the plane containing the
              incident and scattered beam is the axial (_ax) direction.
;```

Crystallography
Journals
Online
Data context

• Metadata: existing standards
  – Dublin Core
    • Provenance, title, genre, topic
  – METS
  – MPEG-21 DIDL
    • Packaging standards
  – Rights
    • Creative Commons, Science Commons
Data context 2

• Metadata requirements for crystallographic data
  – Branch of science: crystallography, chemistry, materials
  – Nature of data set: crystal structure determination
  – Type of experiment: single-crystal, powder, electron diffraction
  – Experimental parameters: rotating-anode, area detector
  – Type of compound: inorganic, organic, metal-organic, protein
  – Purpose of study: full structure determination, phase transition, characterisation/identification, phasing, calibration
  – Identification: compound name, formula
  – Description of associated data files: images, structure factors, validation reports, refinement restraints, structure model
Data context 3

• Identifiers
  – PDB codes
  – CCDC refcodes
  – InChI

• DOIs (digital object identifiers)
  – Registration agencies
    • CrossRef
    • TIB Hannover
Data context 4

• Granularity
  – DOI should relate to a discrete data set
  – DOI may resolve to interface to data set

• Addressability
  – Subsets may be identified by passing parameters to DOI resolver
  – Parameter passing via openURL queries
Data context 5

- Parameter passing

&rfr_id=ori:rid:springer.com
&rft_id=doi:10.1107/9780955360206
&rfr_dat=cr%5FsetVer%3D01%26cr%5Fpub%3D10.1107%26cr%5Fwork%3DThe%20230%20space%20groups%26cr%5Fsrc%3D10.1007%26cr%5FsrvTyp%3Dpdf%26cr_rfr_dat=sgtable7o1o015c.

- Resolves to:
http://it.iucr.org/Ab/ch7o1v001/sgtable7o1o015c
Data discovery

- Identifiers (DOIs)
- Metadata
- Linking
- RSS
- Portals
- OAI-PMH
Data discovery

- Linking
- Portals
- eBank
- Reciprocal Net
- COD
- OAI-PMH
- OAIster
Data validation

- Syntax checking (vcif, ciftest, RDBMS)
- Consistency checking
  - mandatory/recommended data items
  - key indicators (checkCIF/Acta Cryst. E)
  - ‘norm’ versus ‘outlier’ values (PLATON, ProCHECK)
  - data self-consistency (checkCIF, DDLm)
Data validation 2

- Mandatory / recommended data items
  - varies by domain
  - journal editorial policies
  - useful for flagging by checkCIF
Data validation 3

- Key indicators
  - study type
  - conditions (temperature, pressure)
  - geometry precision [mean $\sigma$(C-C)]
  - completeness
  - disorder
  - $R$ factor
  - $wR$ factor

http://journals.iucr.org/services/cif/keyindicators.html
Data validation 4

- Norm / outlier
  - **PLATON**
  - Alerts C, B, A
  - Increasing deviation from expected norms
  - Alert A requires justification for publication in Acta Cryst. E
Data validation 5: macromolecule structures
Data validation 6

- DDLm
- Data definitions
- Relationships expressed by algorithmic methods
- Compatible with existing CIF data sets
- Extensible data types

```
save_cell.atomic_mass
_definition.id            '_cell.atomic_mass'
_definition.update        2006-06-20
_description.text
  ; Atomic mass of the contents of the unit cell. This is calculated from the atom sites present in the ATOM_TYPE list, rather than the ATOM_SITE lists of atoms in the refined model.

_description.common       'Cell Atomic Mass'
_name.category_id         cell
_name.object_id           atomic_mass
_type.purpose             Assigned
_type.container           Single
_type.contents            Real
 Enumeration.range         0.: daltons
_units.code               daltons
_loop_
  _method.purpose
  _method.expression
    Evaluation
  ; mass = 0.
  Loop t as atom_type {
    mass += t.number_in_cell * t.atomic_mass }
_cell.atomic_mass = mass
; save_
```
Data validation 7

- DDLm draft specifications

http://www.iucr.org/iucr-top/cif/ddlm/
The Way Ahead

• Data description/metadata standards
• Agreed quality standards
• Standard resource discovery protocols
• Federated resource holdings
• Distributed search interface
• Archives!