FAIR crystallographic data

Brian McMahon
International Union of Crystallography
5 Abbey Square
Chester CH1 2HU, UK
bm@iucr.org

This is an invited contribution to the first annual meeting of the PaNOSC (Photon and Neutron Open Science Cloud) organisation in Trieste, 4-5 November 2019. The meeting aimed to bring together not only PaNOSC project partners, but also other EOSC clusters with the aim of sharing information and increasing collaboration among different parts. Project partners presented the status and progress of the work packages.

The FAIR principles for crystallographic data

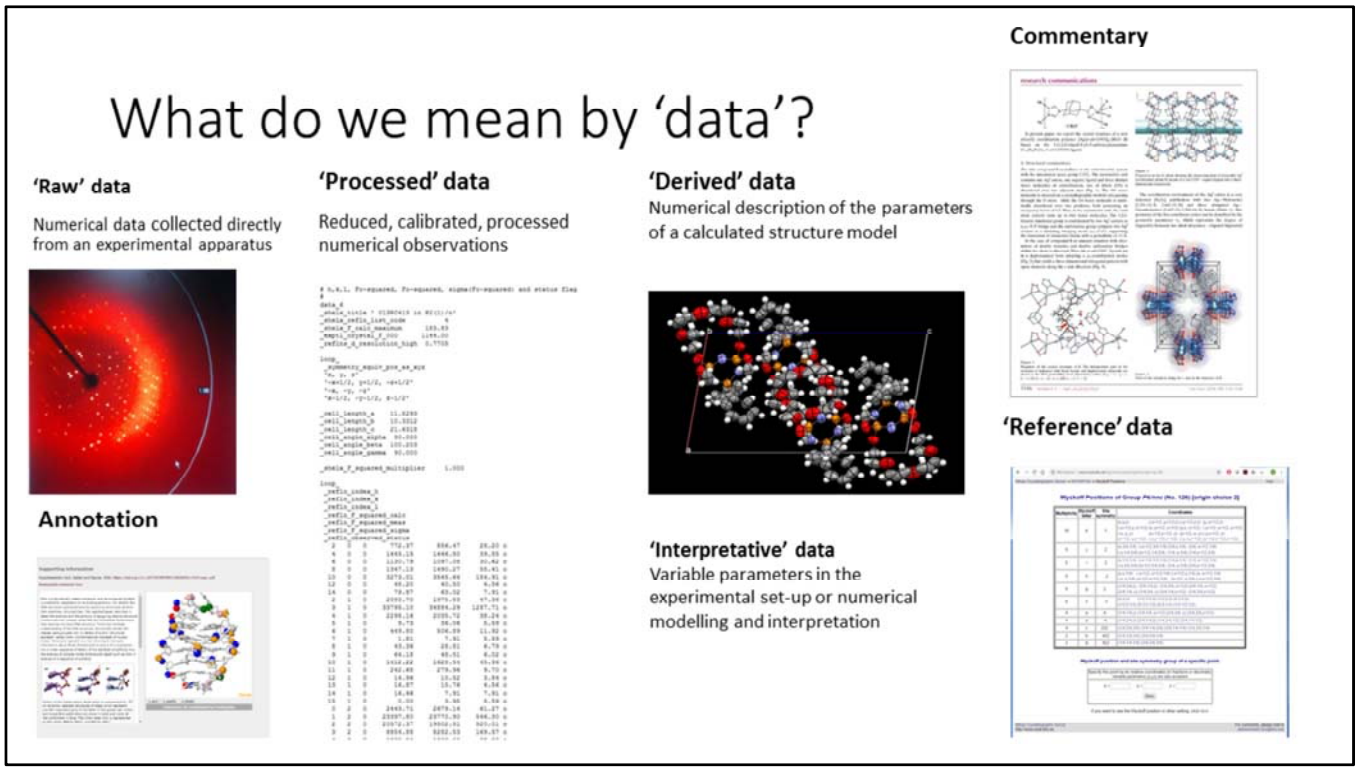
- Findable
  - Unique identifiers, descriptive metadata, e.g. DOI
- Accessible
  - Data stores addressable through identifiers, descriptive metadata or queries
- Interoperable
  - Common vocabulary for descriptive metadata and queries
- Reusable
  - Standard file format(s)

We enumerate the components of the acronym FAIR and describe in a coarse-grained way how they are applied in crystallography.
A unique identifier helps to establish that a data set you retrieve is indeed the one that you were looking for. If you have prior knowledge of the unique identifier, then a data set bearing that value is indeed what you wanted. More usually, you will know something about the data set, characterised by a well defined set of metadata terms. A registry system which provides a rich set of metadata terms associated with each unique identifier is a good mechanism for facilitating findability. DOI (digital object identifier) is an appropriate technology, particularly as there is considerable infrastructure for managing DOI registration, update and characterisation. For literature, this is well managed through CrossRef, which has been built on publisher standards for articles extending back many years (arguably centuries). The equivalent system for scientific research data, DataCite, is still evolving its metadata descriptors. Because IUCr journals have long required the deposition of supplementary data sets, structural CIFs and experimental data sets (structure factors, Rietveld profiles) are assigned DOIs within the CrossRef system. The same is true of PDB entries. However, most scientific data will probably be registered with DataCite. There is, however, interoperability between different DOI registration systems.
In crystallography, we are fortunate to have centralised databases of structural models (and to some extent experimental data sets) – *e.g.* CSD, COD, ICSD, PDB. The IUCr Diffraction Data Deposition Working Group looked at the prospects for establishing

similar repositories for diffraction images, but the size of such data sets makes that unfeasible. More likely is that diffraction images will be stored either on large-scale public repository facilities (*e.g.* Zenodo) or in experimental facilities or institutional repositories. This will make the findability criteria more important. Note that here I have introduced 'queries' alongside 'metadata', to suggest that many of the data sets will sit in, or behind, database query engines. To locate data sets spread across federated repositories, the queries will need to be standardised. Actually, 'queries' in this sense really are not fundamentally different from metadata descriptors.

So, for interoperability, there needs to be a standard vocabulary for the metadata or query terms. Within a domain, this is achievable – it was the original driver for CIF. Across domains it is more of a challenge, but the existence of particular sets of standards (*e.g.* the CIF dictionaries) provides a good starting point for identifying convergences or divergences of terminology between similar domains.

Reusability absolutely demands access to the standard vocabularies, so that the new user/application is confident of the meaning of any data item. It is also greatly facilitated by having a relatively small set of well-defined file formats. Provided there is a standard vocabulary, format conversion is usually relatively straightforward. Within a closed ecosystem, a single file format would be the ideal; but often there are pragmatic reasons why multiple formats need to be supported. Nevertheless, mappings between different formats must be well defined, and are much easier to achieve with standard vocabularies in place.



It's generally helpful in data-related meetings to define what one means by 'data', because there are of course many types of data, and their volume and characteristics require somewhat different approaches in practice. Here I quickly run through the various types of crystallographic data that one might encounter, for two reasons. One is to make the point that they are all relevant to the full scientific record of a piece of research – parts, as it were, of a knowledge continuum. The other is to highlight that the Crystallographic Information Framework, which is the data exchange system developed and promoted by the IUCr, seeks to include all of these types of data, making it particularly useful as a unified model of knowledge representation. The raw data is probably the component that this meeting is most interested in, but it's worth pointing out that synthetic approaches like CIF are beneficial in reducing the friction in moving along this knowledge spectrum, and that greatly helps the interoperability and reusability aspects of FAIR.

Typically, processed data (in this sense) are more compact than the raw data captured at the instrument, and thus may lend themselves more readily to archiving for FAIR re-use. IUCr journals provide access to structure factors and Rietveld profiles for non-biological structures. Each data set has a unique DOI, though is not treated as a first-class digital object (i.e. they are "part of" a structural article). The PDB archives structure factors and NMR restraints for each biological structure, though – as far as I am aware – these do not have individual DOIs. The protein structure entries do.

Derived data in this context are the molecular or crystal structure models, that in their

turn are incorporated into the structural databases. Note that CCDC and the Crystallography Open Database (COD) encourage deposition of structure factors, because these are not always required by non-IUCr journals.

Typically, what I am calling 'interpretative' data are things like constraints and restraints applied during refinement. The CIF project has some difficulty in capturing these in a complete and objective way, since they are often ad hoc devices within particular software implementations. This illustrates the desirability (perhaps need) for archiving software alongside data and literature – fraught with problems, of course, because of machine dependencies, languages, supporting libraries – but also encourages research into capturing and/or defining algorithmic methods. [In a very modest way, DDLm/dREL may be the start of thinking about this in a general way.]

Annotation includes descriptive relationships between subsets of the included structural and experimental data (automatically or manually-generated; the example shown is a Protopedia molecular tour served as supporting information to a journal article). Providing this type of functionality is much easier for the sort of comprehensive data description systems like CIF that can treat numerical data and classical 'metadata' on a similar footing.

Commentary refers in our context to a publication in the traditional scientific literature. It's an extension of the 'annotation' idea, where links to related literature, data sets or other external objects are incorporated into the host data set. This is a good model of where you want FAIR practice to be able to take you to.

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.* Many reference databases are, or include, compilations of discrete data sets. A good example of the way this all comes together is that the database schema for the PDB itself is, essentially, the mmCIF dictionary, which was developed as a mechanism for describing a *single* structure and its component parts.

Benefits of Crystallographic Information Framework

- Interoperable across all types of crystallographic data
- Treats 'metadata' and 'data' on same footing
- Establishes a common file format

  (more or less: DDL1/DDL2/DDLm 'dialects'; CBF as binary equivalent of imgCIF; CIF1 vs CIF2 syntax)

  (but that's not too important because the data structures, types *etc*. are well defined)

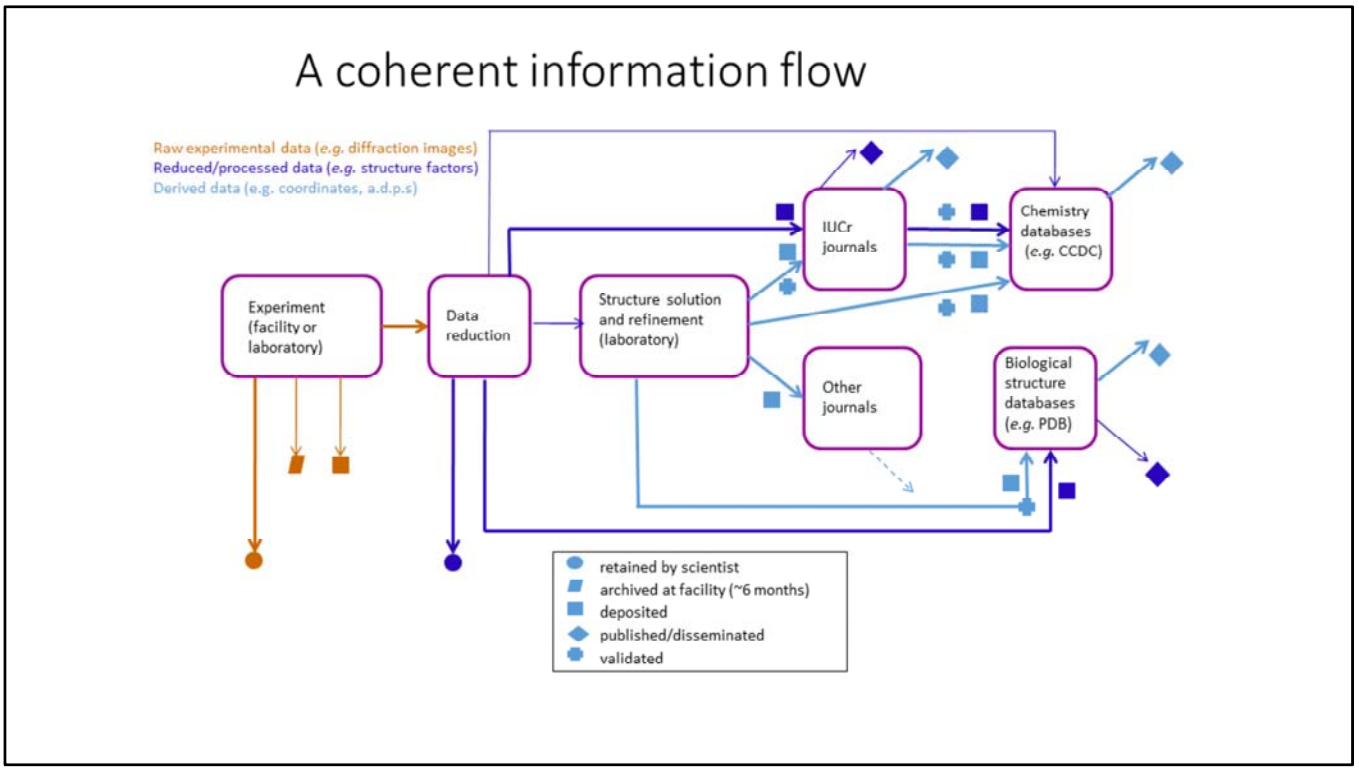
Here we describe some of the architectural features that have made CIF a solid basis for developing FAIR practices in crystallography. First is that it was designed to be generic and extensible, so that it could be applied across any type of data involved in crystallographic research.

There is no formal distinction between items that might be characterised as metadata and those that are categorised as data. This facilitates the flexibility and widespread applicability of CIF. For applications that do use the CIF format consistently, it can facilitate software development, because a single parser can handle any content in a CIF file (even if just by ignoring items not relevant to the current application).

The use of a standard file format along the 'coherent information flow' can make software systems (and workflows) easier to write and maintain. It's not essential – as we have seen in practice, HDF5/NeXus is increasingly common at the raw data end. Even so, it raises the possibility say of retaining a subset of collected images to be converted to imgCIF/CBF format for archive purposes. Anyway, the point to be made is that the existence of a common file format is a useful facilitator, but not an essential requirement for applications to interoperate within the broader CIF framework. Too much time and effort has been spent in the past in format 'holy wars'.

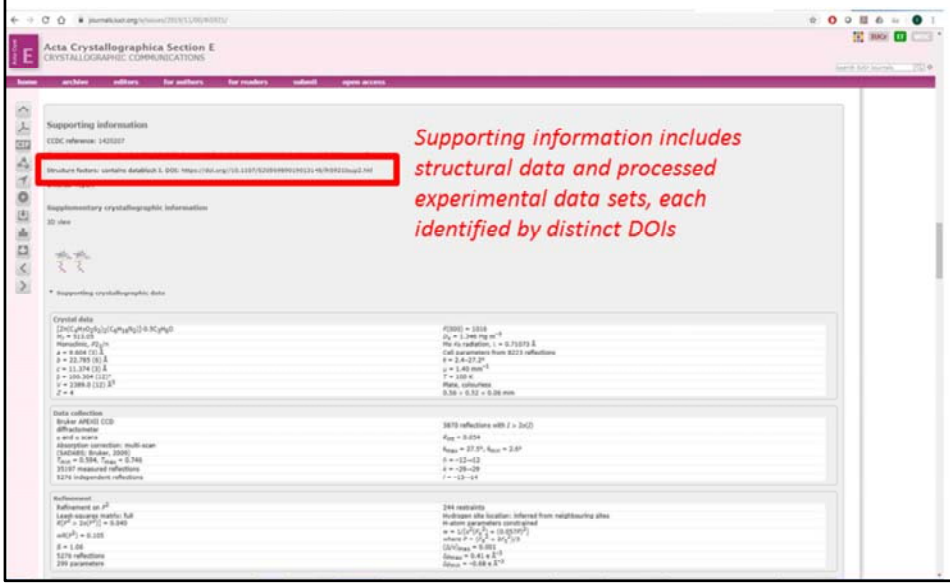
Even within the CIF syntax specifications, there is room for some variation in format, either to accommodate strict relational data models or to allow binary data compression. Nevertheless, controlled variation within a small compass allows for the development of generic software tools and libraries.

And, again, format interconversion, though often a nuisance, is not really a killer if there are wider benefits to be gained from maintaining specialised formats for certain purposes.



This rather well known schematic demonstrates the use of CIF at every stage from experimental station through structure solution, validation, publication and database deposition. In practice CIF **format** files are not involved at every point along this trajectory (*e.g.* HDF5/NeXus in the experimental facilities, Word documents in some variants of the publication process), but all the necessary data items can be characterised by terms defined in CIF dictionaries.

Varieties of data managed by CIF in IUCr publications



Small unit-cell structures:

- submitted as CIF
- include article text (commentary)
- structural model (derived data)
- annotation (visualise individual positions, bonds)
- experimental details ('metadata')
- experimental (processed) data

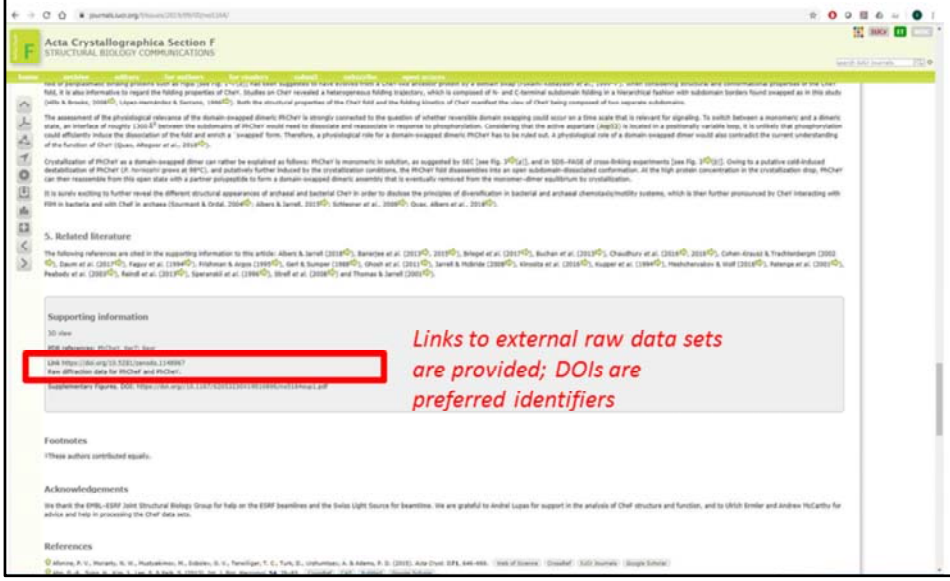
Here are a few slides simply to illustrate how well CIF handles the various types of data in the particular field of publication. In some IUCr journals (*Acta Crystallographica* Sections *B*, *C*, *E* and *IUCrData*) the entire text of an article can be submitted as a CIF file (it is actually the only format supported by *Acta E* and *IUCrData*). This is the 'commentary' type of data in our earlier slides.

Structural data supporting any small-unit-cell structure reported in IUCr journals must be sent in a CIF file (either the submission file for structure reports journal, or an accompanying 'supporting information' file). In some of our journals there is limited application of the notion of 'annotation' data (internal hyperlinks relating displayed text with additional information, as illustrated here), and there is also the ability to interact directly with the numeric data in an article (*e.g.* 3-D visualisation using an application like *JSmol*). There is also a prototypical *JSmol* editor to allow authors to create annotated 3-D views of aspects of the structure.

As emphasised before, 'metadata' in the sense of descriptions of the experimental procedures are embedded integrally within the CIF, and so experimental tables can be formatted automatically for publication (that relieves authors of a lot of labour!).

Processed experimental data, such as structure factors or powder profiles, are also handled easily in CIF format. For IUCr journals (though not necessarily for journals of other publishers), such supporting data are mandatory for small-unit-cell structures. Structure factors are also mandated for PDB deposition of biological macromolecular structures.

Varieties of data managed by CIF in IUCr publications



Macromolecular structures:

- most of the same features as small-cell structures
- links to raw diffraction data sets

Following the IUCr DDDWG/CommDat efforts, public deposition of raw image data sets is actively encouraged for biological macromolecular structures (and supported in the case of small-unit-cell structures, though this community does not yet regularly do so).

Benefits of Crystallographic Information Framework

- Interoperable across all types of crystallographic data
- Treats 'metadata' and 'data' on same footing
- Establishes a common file format

  (more or less: DDL1/DDL2/DDLm 'dialects'; CBF as binary equivalent of imgCIF; CIF1 vs CIF2 syntax)

  (but that's not too important because the data structures, types etc. are well defined)
- **Has precisely defined data items**

However, underlying all these benefits and extensible to other data transmission mechanisms is the fact that CIF data items are formally and precisely collected and defined by community-based expert groups. These machine-readable 'ontologies' (controlled vocabularies, defined physical units and acceptable ranges of values, interrelationships) form the basis for functional interoperability between diverse file formats and software implementations, within any crystallographic or related field.

Data definitions in CIF 'dictionaries'

Managed by COMCIFS

- Crystallographic Core (coreCIF) – 1991 and ongoing
- 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

Managed by 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

This is the current list of distinct dictionaries openly available to the crystallographic community. There are also some local dictionaries (for specific software applications or database implementations), and a growing number of comparable dictionaries in other similar fields (materials science, biological NMR structures).

Example of a data item definition

```
save__diffrn_radiation.polarizn_source_norm

    _item_description.description
;       The angle in degrees, as viewed from the specimen, between the normal to the polarization
        plane and the laboratory Y axis as defined in the AXIS category.

        Note that this is the angle of polarization of the source photons, either directly from a
        synchrotron beamline or from a monchromator.

        This differs from the value of _diffrn_radiation.polarisn_norm in that
        _diffrn_radiation.polarisn_norm refers to polarization relative to the diffraction plane
        rather than to the laboratory axis system.

        In the case of an unpolarized beam, or a beam with true circular polarization, in which
        no single plane of polarization can be determined, the plane should be taken as the XZ
        plane and the angle as 0.

       See _diffrn_radiation.polarizn_source_ratio.
;
    _item.name                  '_diffrn_radiation.polarizn_source_norm'
    _item.category_id             diffrn_radiation
    _item.mandatory_code          no
     loop_
         _item_range.maximum
         _item_range.minimum      90.0   90.0
                                  90.0  -90.0
                                 -90.0  -90.0
    _item_type.code               float
    _item_units.code              degrees
    _item_default.value           0.0
     save_
```

This is an example, chosen pretty much at random from the imgCIF dictionary, to demonstrate the machine-readable attributes associated with a data item definition, and also to indicate the level of precision that is present in many definitions. The goal is to eliminate (or at least minimize) scope for ambiguity or error in sharing data between different applications.

How this might appear in a data file

```
data_image_1

    # category DIFFRN
    _diffrn.id P6MB
    _diffrn.crystal_id P6MB_CRYSTAL7

    # category DIFFRN_SOURCE
    loop_
    _diffrn_source.diffrn_id
    _diffrn_source.source
    _diffrn_source.type
     P6MB synchrotron 'SSRL beamline 9-1'

    # category DIFFRN_RADIATION
    loop_
    _diffrn_radiation.diffrn_id
    _diffrn_radiation.wavelength_id
    _diffrn_radiation.monochromator
    _diffrn_radiation.polarizn_source_ratio
    _diffrn_radiation.polarizn_source_norm
    _diffrn_radiation.div_x_source
    _diffrn_radiation.div_y_source
    _diffrn_radiation.div_x_y_source
     P6MB WAVELENGTH1 'Si 111' 0.8 0.0 0.08 0.01 0.00
```

Just a very brief example of how the data appear in a CIF format file. The data items are clearly labelled with the data name (as defined in the associated dictionary). The format is lightweight, easily machine-readable and can be laid out in a way that makes it easy for humans to read (though the layout is not rigid, and the contents could be packed more densely if desired). Conversion of this format to any other **tag**, **value** paradigm (*e.g.* XML) is very straightforward.

The CIF-NeXus concordance

- Use of imgCIF dictionary as basis for a NeXus MX profile
- Collaboration between COMCIFS and NIAC
- Facilitation meeting at COMCIFS workshop, U. Warwick, 2013
- Herbert J. Bernstein, Tobias Richter *et al*.
- Ongoing project
- HDRMX meetings to establish essential metadata for high-data rate macromolecular crystallography

A very specific example of interoperability with the CIF framework is the concordance established between a COMCIFS Working Group and the NeXus International Advisory Committee over the last few years, to equivalence CIF data names with tags in the NeXus macromolecular crystallography profile, and to help extend the tag sets in both formalisms as required by the community.

CBF, NeXus/HDF5 Interaction

- CBF remains CBF, NeXus remains NeXus
- Interoperablity lets either be used when needed
- New dictionaries and extensions to existing dictionaries help in documenting mappings
- Applications gain from extension to the APIs, starting with CBFlib
- HDF5 and NeXus users gain CBFlib compressions
- CBF users gain HDF5 compressions

Acknowledgement: H. J. Bernstein, DDDWG Workshop, Rovinj, 22-23 August 2015

Major features of the CIF-NeXus interaction, as listed by Herbert Bernstein in his presentation to the DDDWG Workshop on Metadata.

IUCr bodies related to data standards

- COMCIFS (Committee for the Maintenance of the CIF Standard)
  1993-
- DDDWG (Diffraction Data Deposition Working Group)
  2011-2017
- CommDat (Standing Committee on Data)
  2017-

In recent years three IUCr bodies have taken particular interest in data standards. There have also been transient committees charged with oversight of the curated crystallographic databases, and a previous Commission on Crystallographic Data; these responsibilities now fall under the purview of CommDat.

DDDWG 2011-2017

- IUCr Working Group on Diffraction Data Deposition
- Chair: John R. Helliwell
- Terms of reference: 'lead the development of standards for the representation of data and associated metadata that can lead to the routine deposition of raw data'
- Workshops:
  - Workshop at ECM-27, Bergen: August 6 2012
  - **Workshop at ECM-29, Rovinj: August 22-23 2015** (metadata)
  - Workshop at ACA 2017, New Orleans: May 26 2017
- http://www.iucr.org/resources/data/dddwg

The DDDWG was particularly effective in establishing the role of raw data deposition (and to some extent standardisation) particularly in structure solution by X-ray diffraction methods, but with some awareness of other experimental data types and methodologies.

DDDWG approach

- Careful analysis of technical requirements (articles in *Acta Cryst.*)
- Canvassing of community views
- Engagement of IUCr Commissions
- Biological structures community initially conservative
  - Structure factors mandatory for PDB depositions
  - Possible interest in unmerged structure factors
- However, over lifetime of DDDWG, interest grew
  - Working repositories (Store.Synchrotron; IRRMC, SBGrid)
- Chemical crystallographers now being consulted

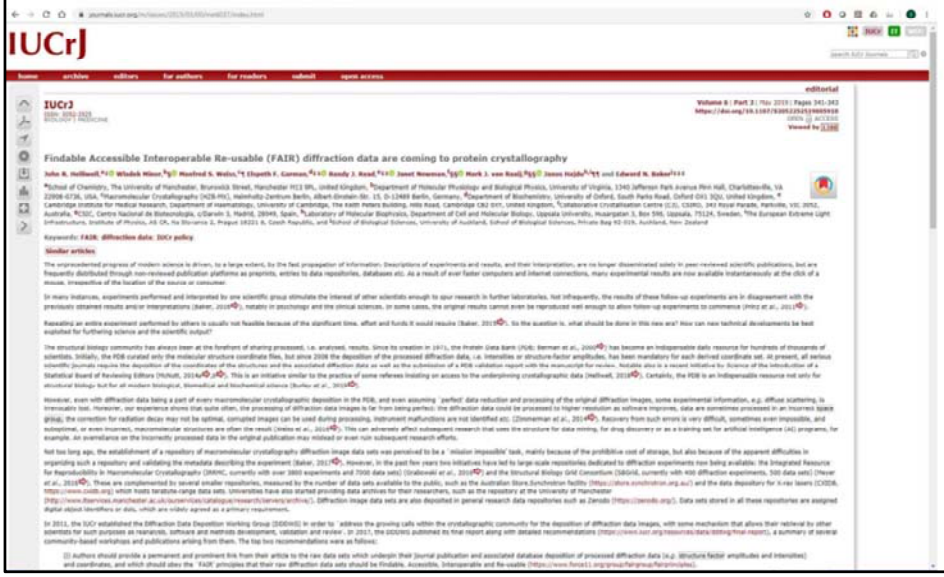
Without going into too much detail, this slide emphasises that the working group both formulated its own analysis and took account of community opinions, both within existing mechanisms (the ccp4bb bulletin board was an effective medium for hearing about the evolving concerns of macromolecular crystallographers) and by establishing new ones (a public discussion forum at https://forums.iucr.org/viewforum.php?f=21). The analysis papers commissioned for *Acta Cryst. D* in 2014 were a valuable part of this feedback loop; since they were published, the five articles have collectively had over 20,000 downloads and 85 citations in Web of Science.

DDDWG final recommendations (17 in all!)

- **Authors should provide a permanent and prominent link from their article** to the **raw data sets** which underpin their journal publication and associated database deposition of **processed diffraction data** (*e.g.* structure factor amplitudes and intensities) **and coordinates**, and should obey 'FAIR' principles (Findable, Accessible, Interoperable and Re-usable https://www.force11.org/group/fairgroup/fairprinciples)
- A registered Digital Object Identifier (**DOI**) **should be the persistent identifier of choice** (rather than a URL) as the most sustainable way to identify and locate a raw diffraction data set.
- An **archive of raw diffraction data** sets for **currently unsolved crystal structures** should be pursued.
- An **archive of raw diffraction data** sets **showing significant diffuse scattering** should be pursued.
- Workshops for **research data management training** for the community should continue and be sponsored and organised by the IUCr.
- There should be continued regular checking by the IUCr Executive Committee of the progress of the **IUCr Commissions** logging of their **raw diffraction data metadata**.

These are (perhaps) the most significant of the DDDWG's final recommendations. The first two have resulted in recommendations in IUCr biological structure journals encouraging authors to deposit and provide access to raw data sets.

CommDat/Commission on Biological Macromolecules journals initiative



"IUCr Journals are now taking the lead by encouraging authors to provide a DOI for their deposited original raw diffraction data when they submit an article describing a new structure or a new method tested on unpublished diffraction data. In the case of methods developed or tested with raw diffraction data, these data must be available to referees, and deposition of such data will eventually become compulsory. Permanent and prominent links will be provided from articles to the underpinning experimental data of each published research study."

Here is the relevant text from the Editorial in *IUCrJ*, which was also published in other IUCr journals. As this was a recommendation of the Commission on Biological Macromolecules, it does not extend explicitly to data sets collected for non-biological macromolecular structures; but all IUCr journals will provide links to any raw data sets deposited in accordance with these principles.

CommDat 2017-

- IUCr Standing Committee on Data
- Subsumes DDDWG (and other interests/activities)
- Formal relationship with COMCIFS on data representation standards
- Undertaking survey of chemical crystallography community
- Joint project with COMCIFS: '*checkCIF* for raw data'
  - Define minimal set of 'mandatory' metadata for effective (re)use of an image
  - Define computational strategies for identifying and retrieving specific metadata
  - Determine an approach to validation of image data

Another ongoing activity of CommDat is the effort to encourage instrument vendors to provide complete and consistent metadata allowing reuse of diffraction images. This has been labelled *checkCIF for raw data*, by analogy with the IUCr journals' successful *checkCIF* validation approach to derived structural data sets and publications.

What is on the horizon/what would you like to see elsewhere?

- More ontologies
- More collaborations with other related communities (interoperability)
- Further development of DDLm (enhanced automated validation)

A few thoughts in response to Tobias Richter's invitation to discuss this topic.
We are keen to see continuing work to characterise areas of structural science not adequately covered by existing CIF dictionaries.
We are keen to work with other communities to define areas of overlap or contiguous data descriptions.
We are developing the tools to improve internal validation of data values and inter-relationships. For example, dREL is a very specific programming language developed to optimise such validation methods in CIF dictionaries constructed using the DDLm formalism. Note that it is perfectly feasible to maintain a dictionary (*i.e.* ontological schema) in CIF/DDLm formalism, defining terms whose use is not confined to CIF format files.

Contacts

- CommDat: Chair John R. Helliwell john.helliwell@manchester.ac.uk
- COMCIFS: Chair James Hester jamesrhester@gmail.com
- *checkCIF* for raw data: Loes Kroon-Batenburg l.m.j.kroon-batenburg@uu.nl

And here are the current Chairs and lead investigators of these current activities.