Draft COMCIFS Triennial Report for the period 2011-2014

Introduction

COMCIFS is responsible for maintaining and developing the Crystallographic Information Framework (CIF) on behalf of the IUCr. COMCIFS activities include development and approval of new dictionaries as well as development and support of the underlying standards for syntax and dictionary construction. Work in the previous triennium (2008-2011) was primarily devoted to incremental improvements in existing dictionaries and development of a new set of standards underpinning CIF.

Dictionaries

A key aspect of the CIF project over the last two decades has been codification of crystallographic knowledge into machine-readable dictionaries. Once approved, dictionary management passes to a 'Dictionary management group' (DMG) which now has full autonomy in managing dictionary updates. The rate at which these dictionaries have appeared has slowed markedly since the publication of Volume G in 2005. The current triennium has seen the publication of one long-awaited dictionary defining data items for describing twinning.

Given that it is now almost a decade since the publication of Volume G, COMCIFS has recently initiated a review of the core CIF dictionaries to determine what new datanames might be necessary and which existing definitions require updating.

Exploiting DDLm

For historical reasons, CIF dictionaries are currently written using one of two “dictionary definition languages” (DDLs). Recognising that a single DDL would be preferable, around a decade ago COMCIFS commissioned work on a replacement DDL that would incorporate the advantages of the earlier DDLs. The new DDL, dubbed “DDLm”, was accepted in draft form at the Osaka IUCr meeting and the required enhancements to CIF syntax were largely agreed to at the Madrid IUCr meeting in 2011. The recent triennium saw the publication of two key papers describing the basis of the new standards: “DDLm: A new dictionary definition language”, Spadaccini, N. and Hall, S.R., *J. Chem. Inf. Model.*, 2012, **52**(8) pp 1907-1916 and “dREL: A relational expression language”, Spadaccini, N. Castleden, I. R., du Boulay, D. and Hall, S.R., *J. Chem. Inf. Model.*, 2012, **52**(8), pp 1917-1925. Work during the current triennium has been focussed on developing tools and dictionaries capable of exploiting these new standards. One ground-breaking aspect of these standards is that they allow machine-readable description of the mathematical relationships between datanames. As a demonstration of the power of this approach, web-browser software has now been developed that allows interactive calculation

and verification of dataitem values (for example, structure factors), based solely on the values already present in the data file and the contents of an arbitrary dictionary loaded into the web page. This software is immediately useful, as the new DDLm dictionaries include equivalent datanames from previous dictionaries, which means that this software works also for archival CIF data files written using datanames from current dictionaries. This software and related topics were presented and discussed at an intense and fruitful two-day workshop prior to the 2013 ECM meeting in Warwick. Efforts in the coming triennium will focus on improving DDLm tools, converting dictionaries to DDLm, and supporting moves to use DDLm technology in production within the IUCr offices.

Macromolecular developments

The macromolecular CIF dictionary (mmCIF) is by far the largest CIF dictionary, and is usually augmented by the even larger PDBx dictionary, resulting in close to 10,000 definitions in total. The latter dictionary is actively maintained by the wwPDB project (http://wwpdb.org). Due to the long history of the original PDB format, community uptake of the more flexible CIF format has been slow. The difficulty of encapsulating ever-larger structures within the confines of the PDB format has seen a concerted effort this triennium to establish mmCIF/PDBx as standard for data exchange and archiving in structural biology. This was largely achieved by 2013, with changes in wwPDB deposition systems, two workshops to facilitate the transition, and support for mmCIF/PDBx included in major structural biology software packages.

Software

Uptake of CIF is heavily dependent on the availability of CIF-conversant software. In the first decades of CIF, end-user software would usually include custom code for reading and writing CIF files, which increased barriers to CIF adoption, as well as multiplying opportunities for incorrect CIF reading and writing due to coding errors. This proliferation of custom CIF solutions usually arises from a mismatch between the needs of a given program and the functionality and coding language of available libraries. One way of improving the match between CIF libraries and application software is to design an Application Programming Interface (API) and accompanying reference implementation which would meet the needs of the majority of CIF programs. A draft API has been developed and a standard C library conforming to this API is being prepared as a test of the design and as the reference implementation.

Interaction with other data management initiatives

At the Madrid IUCr meeting in 2011, the IUCr executive committee created the Diffraction Data Deposition Working Group (DDDWG). As a result of the recommendations of this working group, IUCr commissions were tasked with determining the metadata needs of their fields. While no explicit mention of CIF (or COMCIFS) was made, COMCIFS' fundamental role within the IUCr has been to define and manage metadata and several COMCIFS members have decades of experience in doing this. A group of COMCIFS members have therefore agreed to advise any IUCr bodies that wish to take advantage of this expertise. I urge the Executive to encourage and facilitate such communication between COMCIFS and IUCr commissions so

that the many false starts that dogged pre-CIF efforts at data standardisation in crystallography are not repeated.

As other scientific fields initiate their own metadata projects, there will be increased need for harmonisation between these efforts and CIF definitions. Many large-scale neutron and X-ray facilities are now storing their raw data in the NeXus framework, which is controlled by the NeXus International Advisory Committee (NIAC). In 2013 the NIAC met with COMCIFS to explore ways to harmonise the NeXus definitions for raw images produced by protein crystallography experiments with the definitions contained within the imgCIF dictionary. As a result of these discussions and concerted efforts of a number of researchers, a harmonised set of definitions is available making it possible to convert freely between NeXus and imgCIF files. It is hoped that this effort will lead to further harmonisation between NeXus and other fields covered by CIF dictionaries.

Membership

COMCIFS participants include a large number of advisers/observers and a small number of voting members. John Bollinger became a voting member following the Madrid congress, and R. Grosse-Kunstleve resigned partway through the triennium. The other voting members are J. Hester (chair), B. McMahon (secretary), H. Bernstein, and J. Westbrook.

James Hester, Chair.