Adoption of a next generation dictionary definition language: DDLm

CIF has now been in active use as a data exchange framework for about fifteen years. The exchange, archiving and deposition of CIF data are supported by data dictionaries that define the items most widely-used in the different crystallographic applications. Dictionaries are CIF documents organised according to a specific set of rules referred to as the *dictionary definition language* (DDL). In crystallography two DDL's have evolved to satisfy an expanding definition need; DDL1 is the simplest and is used for the core, powder, modulated and precision density dictionaries; DDL2 has more relational attributes and is used for the macromolecular, image and symmetry dictionaries. While the DDL1 and DDL2 attributes are similar, the differences add an extra level of complexity to the application and the maintenance of the CIF dictionaries. Knowledge management technologies have also advanced considerably since either DDL was introduced and it was for these reasons COMCIFS decided at the Florence Congress to charge a working group (Syd Hall, Nick Spadaccini and John Westbrook) with the task of proposing a "next generation" DDL for use with future crystallographic definitions.

The brief for this group was that a new DDL should, at a minimum, : meet the attribute capabilities of existing DDLs; increase the semantic richness and precision in definitions, ; and provide mechanisms that will enable common data definitions to be shared across domain-specific dictionaries (i.e. the same item need not appear in more than one dictionary). An overarching requirement was that definitions be simple to write, to maintain and to apply in CIF exchange processes.

A DDL proposal has been received from the working group for our consideration. It has been called DDLm to reflect the inclusion of methods attributes for relating defined items. The name of the new DDL is not important at this stage, and can be discussed later prior to adoption.

COMCIFS now needs to discuss and assess this proposal. Details of DDLm are available on the IUCr web at http://www.iucr.org/iucr-top/cif/ddlm/index.html. This page provides four URLs.

- *README* summarises the files available in the *File Distribution* URL.
- *Descriptive Documents* contains RTF and PDF files describing the new DDL.
- *File Distribution* provides individual files for downloading.
- *Zip Distribution* provides an automatic download of all files as a .zip file.

ddl.dic contains the formal definitions of the DDLm attributes and is the main document for consideration by COMCIFS. The other files describe the DDLm attributes, and to show their typical application in crystallographic definitions. The TEST domain dictionaries cif.dic, cif_core.dic, core_*.dic and com_*.dic have been used to trial the DDLm attributes, and these files are provided here to illustrate the typical application of the attributes. They do not constitute a proposal to replace the current core dictionaries.

It is recommended you read the descriptive documents **DDLm_spec_08.rtf**, **ddl_attr_08.pdf**, **ddl_import_08.pdf** and **dREL_spec_08.rtf** in that order. It certainly helps to appreciate the role "importation" plays in the new DDL; this is the mechanism that ensures that an item is defined in only one file and is imported only when individual dictionaries are used in applications. To best understand the use of the DDLm attributes in definitions, look at the test domain dictionaries **core_diffr.dic**, **core_struc.dic**, etc. Note that the division of the core definitions into these particular dictionary modules is not important at the moment, and has been done simply to illustrate how definitions may now be organized into their natural groupings (and be maintained independently) without any need to duplicate common definitions. Importation attributes will expand these dictionaries automatically as they are applied and thus ensure that the latest unique definitions are sourced and used.

Look carefully at the use of the symbolic methods expressions in the definition of derivative data items (i.e. items that are related to other data items). These inbuilt methods serve a number of functions; foremost they permit a much more precise definition of an item in terms of its relationship to other items, but methods may be also applied actively to specific data instantiations for the purposes of evaluation and validation. This is possible because the symbolic language dREL is executable using a Jython engine that will be supplied to software developers. It is also recommended that developers be encouraged to develop their own "methods expressions" (written in a computer language of choice) and link these to the dictionaries via the item tags. Such external "methods systems" would not be overseen by COMCIFS, though it may wish to provide standard data files for conformance testing of derivative calculations.

All of the new definition features of DDLm will not be elaborated on here, however COMCIFS members are asked in particular to carefully review the new TYPE attributes. These are much more comprehensive than DDL1 and DDL2 and provide for more precise definition and validation. Note that container types List, Array, Tuple and Table depend on the acceptance of new multi-line string delimiters based on matching square, round and curley brackets. It is proposed that List and Array strings be bounded by square [ ] brackets; Tuples by round ( ) brackets and Tables by curly { } brackets. Bracketed strings may be nested and extend over multiple lines. This syntax change has been advocated in COMCIFS discussions for several years now and it is will be applied to the Star File specifications. These string delimiters are the only change to affect CIF data files (all existing files will be backwards compatible) and the proposed syntax is unlikely to pose problems for future parsers.

The **_type.purpose** attribute identifies the function and origin of a data item; and in particular if its value is a number deemed to be "Measured" (either directly from the experiment or by derivation) and that therefore has an associated standard uncertainty value. This is important in order to cope with the current practice of allowing the SU value to be appended to the measured value of a single item (e.g. _blat 0.622(6)) or as two items with separate tags (e.g. _blat 0.622 _blat_su 0.006). It is proposed with DDLm that the existence of the item <tag>_su be *implicit* if the item <tag> is defined as Measured; and that both the concatenated and the separated value-su constructions be accepted interchangeably. That is, it is not necessary to define <tag>_su in a dictionary but its presence in an instance document is understood and acceptable.

It should be stated that DDLm, as a next generation definition language, is intended principally for use in future dictionaries. Because of its added semantic richness, application to existing dictionaries will be encouraged but not be insisted on, and it is expected that the existing dictionaries will be supported as long as they are widely used. The motivation for the adoption of DDLm must be its inherent advantages in meeting current and future knowledge management challenges, and its attractiveness to software developers so that dictionaries are used routinely for handling data.