COMCIFS Triennal Report to the IUCr General Assembly for 2005-2008.
DRAFT

Introduction:
Eighteen years have passed since the Union adopted CIF (Crystallographic Information Framework, formerly Crystallographic Information File) as a standard for the submission of crystal structure reports to the Union journals. We have learned much during that time and COMCIFS, which has responsibility for managing CIF, has been reviewing past progress and planning future directions. Priority in the early years was given to preparing the required CIF dictionaries which now contain an impressive two thousand definitions of crystallographic items. Until recently few other scientific disciplines had a comparable set of dictionaries with such a wide community acceptance though extensive work is now underway to provide the biomedical sciences with a comparable ontology. The CIF dictionaries are used in conjunction with the STAR file syntax as the format for the extensive current archive of CIF-based structure reports. In the field of small-cell crystallography CIF is widely accepted as the standard for the submission of structure reports to many scientific journals, for their archiving and downloading as well as for transferring crystallographic information between users. In the macromolecular field CIF is used to archive the Protein Data Bank, and while the file structure does not yet have the same wide community acceptance, the wwPDB exchange dictionary and the official mmCIF dictionary on which it is based provide the definitive definitions of terms in this field.

Dictionary Definition Language (Methods) - DDLm:
Eighteen years of experience has shown us how CIF can be enhanced, and at the Florence IUCr Congress in 2005 COMCIFS agreed to develop a new dictionary definition language (DDLm) with advanced capabilities. DDLm will make it easier to keep dictionaries up-to-date, to assemble virtual dictionaries customized to individual CIFs, to allow CIFs to include vectors, matrices and tables, and to include in the dictionaries machine-readable expressions that will allow items not present in particular CIFs to be calculated from the data that is present and, more importantly, will allow for consistency checks. The dictionaries will thus not only define individual crystallographic concepts in human readable form, but will describe how the concepts are related in machine-readable format. Programs written for DDLm dictionaries will be able to read all existing CIFs, and will bring a considerable added value to the task

During the triennium a final version of DDLm has been created and is currently being evaluated. Work has begun on converting the existing dictionaries to the DDLm standard. It will take some time to complete this conversion, but each new dictionary can be brought on line as soon as it is approved. Software that takes advantage of the DDLm features is already being written.

Microsymposium 96 at the Osaka Congress has been arranged jointly with the Commission on Crystallographic Computing to introduce DDLm and describe the way it will impact on the design of crystallographic software.

Dictionaries:
Work continues on the evolving CIF dictionaries. The imgCIF dictionary was designed to record images, specifically those of raw diffraction patterns from two dimensional detectors. Changes to the imgCIF dictionary to support its use at SLS, Diamond and ESRF are being

discussed on the imgCIF list and are the subject of a series of workshops. imgCIF has been adopted as the output format for the new Dectris Pilatus 6M detector at the Swiss Light Source in Villigen Switzerland and as of early January 2008, ADSC had prepared software to produce imgCIF from all its detectors for use at the Diamond Light Source in Chilton, England.

The imgCIF group has held a number of workshops during the triennium including one at the ACA meeting in Hawaii in 2006, one at BNL and another in the UK in connection with BSR in 2007. Further work and workshops will continue in 2008 and the results of this effort will be discussed at the IUCr Congress in Osaka.

In addition to the construction of DDLm dictionaries, version 2.4 of the core CIF dictionary has been released during the triennium and further revisions are planned. Work on dictionaries for small angle scattering and reflectivity has resumed after a hiatus.

Software:
A notable addition to the suite of CIF programs is the Python package PyCIFRW which not only reads CIFs using the CIF dictionary, but reads CIF dictionaries using the DDL and even validates DDL against itself [Hester, J. R. (2006). J. Appl. Cryst. 39, 621-625]. This work leads naturally to the provision of DDLm compatible software which will work in the same way.

With financial help from the IUCr, Herbert Bernstein and students have produced updated versions of standard software libraries and tools (the CIFTEST parser test and validation suite, CIFtbx3, cyclops, vcif2, and a new utility to fold and unfold long-line CIFs) that are compliant with the version 1.1 CIF specification. They have also been working on DDLm compliant software for publications.

The appearance of a number of CIF editors: enCIFer, CIFedit and PublCIF, which validate CIFs against the relevant dictionaries means that definitions of items can be displayed on the screen and syntax errors can be detected to ensure that editing produces a conformant CIF.

PublCIF, produced by he IUCr editorial office, combines this dictionary-based CIF validation with a sophisticated collection of utilities that will assist prospective authors. These include active links to the checkCIF service, the ability to incorporate validation report forms (VRFs) generated by checkCIF, data entry wizards, table editors, previews of articles formatted in the styles of the different IUCr structural journals, citation sorting and checking, private databases of authors and citations, and dictionary browsing facilities. A tool has also been developed for creating enhanced figures in all IUCr journals; these figures are three-dimensional visualizations of molecular structure with associated animations, specified views and schematic representations, and use the CIF (or mmCIF) directly for atomic coordinates and crystallographic symmetry information.

PublCIF software is being further developed to support publication of biological macromolecular structures. Howard Einspahr has worked closely with John Westbrook, of the PDB, and IUCr editorial staff to develop streamlined procedures for incorporating structural data from PDB deposits in associated publications. A set of recommended data items has been drawn up describing information on: the sample and its treatment (including crystallization); data collection and structure solution; and structure refinement details. These can be harvested automatically from an mmCIF and will generate a table for publication in the article.

Macromolecules:
The Protein Data Bank continues to the extend the content of the wwPDB Exchange Data Dictionary (PDBX). PDBx a superset of mmCIF, which in addition to macromolecular X-ray methods, includes structure and experimental representations of NMR, 3D electron microscopy, homology modeling, and experimental details of protein production (http://mmcif.pdb.org) A translated version of this dictionary is maintained as an XML schema (PDBML) (http://pdbml.pdb.org). During the triennnium significant extensions have been added to represent large molecular assemblies and more detailed chemical description of both polymer and non-polymer molecular components in macromolecular structures. This year the BMRB partner of the wwPDB released an integrated data deposition tool for NMR and experimental and structure data NMRADIT using a consolidate mmCIF dictionary. In other words there is now a lossless translation between NMR STAR and the PDB Exchange dictionary. The wwPDB has released a remediated version of the PDB archive which takes advantage of these latter extensions [K. Henrick; et al. Nucleic Acids Research 2008 36(Database issue):D426-D433; doi:10.1093/nar/gkm937]. Software related to this work can be found at http://sw-tools.pdb.org/ which contains a broad selection of mmCIF dictionary and data management tools. This includes our editor tool ADIT, dictionary validating tools, database loading tools, and data harvesting tools.

Publicity and housekeeping
A complete documentation of CIF concepts and associated data dictionaries was completed in 2005 and published as Volume G of the IUCr *International Tables* series.

COMCIFS runs a number of discussion lists that can be publicly accessed from the IUCr web site. These include, among others, the main COMCIFS discussion list used for general announcements and discussions of matters relating to CIF itself, and the cif-developers list that has proved popular among software developers for obtaining advice on some of the more esoteric aspects of the standard that are, however, vital for programming.

Now that CIF is firmly established in the crystallographic community and the nature of the work of COMCIFS is moving from the production of dictionaries to the coordination of software, it is time to bring younger members with fresh expertise into COMCIFS. Together with many of the members of COMCIFS who helped to establish the original CIF standard at the end of the last century, I will be handing over responsibilities at the Osaka Congress to a renewed team. I would like to take this opportunity of thanking Helen Berman, Syd Hall and Gotzon Madariaga for the great work they have done in guiding COMCIFS through its first critical years.

I. David Brown
Chair