MSWK.CF.09 A SUPPORTING LIBRARY OF SOFTWARE TOOLS FOR CIF AND DDL 2. John Westbrook and Shu-Hsin Hsieh, Department of Chemistry, Rutgers University, Piscataway, NJ 08855-0939

A Data Description Language (DDL) provides the framework for the construction of comprehensive dictionaries of terminology such as the macromolecular CIF (mmCIF) dictionary. The DDL also provides features which may be used by software to perform rigorous validation of CIF dictionaries and CIF data files. A software library (CIFLIB) that is based on DDL 2 and provides simple and convenient access to information in CIF dictionaries and data files will be described. This library contains a collection of high level functions to process and check data stored in CIF format using the full detail of the mmCIF dictionary.

CIFLIB provides a C language application program interface which encapsulates all I/O operations and integrity checking on CIF dictionaries and data files from a calling application. Some of the functions provided by CIFLIB include: read and write operations on CIF format data files and dictionaries; read and write access to individual data items; detailed integrity checks on CIF data files and dictionaries; robust error handling; and methods to efficiently navigate the CIF data model. Examples of the use of the library will illustrate how CIFLIB has been applied as a data validation tool in processing crystal structure data at the Nucleic Acid Database Project.

Support for this work has been provided by the NSF (BIR 9510703).

MSWK.CF.10 TRANSLATING MMCIF DATA INTO PDB ENTRIES . Frances C. Bernstein, Protein Data Bank, Chemistry Dept., Brookhaven National Laboratory, Upton, NY 11973-5000, USA and Herbert J. Bernstein, Bernstein + Sons, 5 Brewster Lane, Bellport, NY 11713-2803, USA.

The major steps needed to translate mmCIF data into a "pseudo-PDB" format (a format sufficiently similar to standard PDB format to be accepted by most applications) are presented, with examples drawn from the program cif2pdb. The objective is to help application developers and people writing CIFs understand uses of mmCIF which will hinder translation to PDB format and to help users familiar with PDB format understand new mmCIF constructs.

The Protein Data Bank format has been used for over 20 years to archive macromolecular data, is produced by many refinement programs and is used as an input format by many applications. Adoption of the mmCIF dictionary by the IUCr will lead to the creation of a significant pool of mmCIF data sets. However, it may be some time before existing application programs can handle mmCIF input. Therefore it is important to have facilities to translate mmCIF data into PDB format to facilitate the use of CIFs with existing programs. The PDB is developing a CIF-based AutoDep program,which uses the WWW. In those areas where there is a one-to-one correspondence, AutoDep will use mmCIF tokens and produce the appropriate PDBrecords. However, there are areas where more complex transformations are needed and in user labs it is not always possible, or even desirable, to make a perfect PDB entry from an mmCIF data set. If the only purpose of the translation is to display a molecule, then there may be no reason to reorganize residues and het groups and change atom names to match PDB standards. We discuss both the production of a "pseudo-PDB" format which can be converted into a rigorous PDB format with further processing, and the capabilities of the PDB AutoDep program. We consider ways to construct a valid PDB ATOM/HETATM/TER list, and approaches to deal with mmCIF tag values for chain identifiers and atom names which may not fit into PDB field sizes.

Work supported in part by US NSF, PHS, NIH, NCRR, NIGMS, NLM and DOE under contract DE-AC02-76CH00016.

MSWK.CF.11 A MMCIF TOOLBOX FOR CCP4 APPLICATIONS. Peter A. Keller, CCP4, Daresbury Laboratory, Warrington WA4 4AD, UK

A programmers' toolbox of routines has been written, for manipulation of CIF's. It has been designed to integrate with the CCP4 protein crystallography suite[1], and enables existing CCP4 applications (written in Fortran77) to be converted to use appropriate categories of CIF information. New applications which conform to existing CCP4 practice, are also easily written.

Input CIF's are parsed according to the STAR syntax rules[2] (implemented using a PCCTS[3] grammar), and an abstract syntax tree (AST) representation of the data is built. Information on the location of each data item is maintained internally, for rapid retrieval. CIF data is generated or modified by constructing or manipulating an AST. Data verification is performed against a dictionary conforming to the Macromolecular DDL version 2.1[4]. The dictionary itself is converted to a binary representation which incorporates a hashed lookup on data names.

The Fortran interface shields the programmer from the more complex aspects of accessing CIF's, by keeping all data structures internal to the toolbox. To access data, application programs specify files, data blocks and data names, with the values being handled as the Fortran data type appropriate to that defined in the CIF dictionary. A system of contexts makes the data access routines effectively re-entrant, which simplifies multiple accesses to data categories, and only a minimum of information needs to be maintained by the application itself between calls to toolbox routines.

[1] CCP4. Acta Cryst. D50, 760-763.(1994)
[2] S.R. Hall, N. Spadaccini. J. Chem. Inf. Comput. Sci. 34, 505-508 (1994)
[3] T.J. Parr. 'The Purdue Compiler Construction Tool Set'. See http://www.igs.net/~mtr/software-development/pccts.htm
[4] J. Westbrook, S.R. Hall. See http://ndbserver.rutgers.edu/DDL/index.html

MSWK.CF.12 IMAGENCIF: AN INITIATIVE TO STANDARDISE IMAGE FORMATS. A P Hammersley,ESRF, BP 220, 38043 Grenoble Cedex, France

A working group, self-named as "imageNCIF", has been formed to extend the CIF concept to cover the storage of area detector data. It has developed out of discussions which took place at a computing workshop held at the Brookhaven National Laboratory in March 1995 [1]. The principle goals are to provide a header that describes the data and may contain auxiliary information needed by analysis software, and to produce a simple efficient storage format.

E-mail discussions have led to the conclusion that a true CIF (ASCII-text based) format is not appropriate for raw area detector data, simply because of the quantity of data involved (Mbytes for an image and maybe Gbytes for a data-set). A CIF-compatible format is developing, where a binary file would consist of a text header followed by the data. The header section would conform to all normal CIF rules except, perhaps, for the manner in which "lines" of text were separated. The image data would follow in a binary section. Items such as the size of the binary data and the manner in which individual elements were stored would be described in the header section through additional CIF data names. A simple utility program would be provided to extract the CIF-compatible section from the binary file, and output a true CIF.

[1] R Sweet, "A Workshop on Graphical User Interfaces for Synchrotron Protein Crystallography", Synchrotron Radiation News, Vol. 8, No. 5, pp 6-7 (1995).