MSWK.CF.05 INTRODUCTION TO THE CIF POWDER DEFINITIONS. Brian H. Toby, Reactor Radiation Division, National Institute of Standards and Technology, Gaithersburg MD 20899 USA.

Definitions have been added to the CIF dictionary to accomodate powder diffraction measurements and results. These extensions support data from nearly all types of instruments, including energy-dispersive, multidetector and conventional xray and neutron diffractometers and cameras. While CIF was initially designed to document a crystallographic determination, the powder diffraction definitions allow for complete documentation of an experiment, as is required to exchange or archive raw data. In contrast to conventional single crystal CIFs, there are many circumstances where several CIF blocks will be needed for description of a single experiment or structure determination, so the powder CIF dictionary provides for pointers between CIF blocks and files.

MSWK.CF.06 THE MMCIF DICTIONARY: COMMUNITY REVIEW AND FINAL APPROVAL. Paula M. D. Fitzgerald, Merck Research Laboratories, Helen Berman, Department of Chemistry, Rutgers University, Philip Bourne, San Diego Supercomputing Center, Brian McMahon, International Union of Crystallography, Keith Watenpaugh, Physical and Analytical Chemistry, Pharmacia & Upjohn, and John Westbrook, Department of Chemistry, Rutgers University

The Crystallographic Information File (CIF) was developed by the IUCr Working Party on Crystallographic Information, in an effort sponsored by the IUCr Commission on Crystallographic Data and the IUCr Commission on Journals. The result of this effort, a dictionary of data items sufficient for archiving the small molecule crystallographic experiment and its results, was formally adopted by the IUCr in 1990.

In 1990, the IUCr formed a working group to expand the dictionary to include data items relevant to the macromolecular crystallographic experiment. As this effort progressed, we realized that the complex nature of the macromolecular experiment demanded a more rigorous data model than was provided for by the original CIF dictionary and its syntax laws (the Dictionary Description Language, DDL), and so a new DDL was developed and the mmCIF data model was recast as a flat-file representation of a relational database schema. This data model provides for the storage of information concerning all aspects of the macromolecular crystallographic structure determination process, beginning with the source of the material, and proceeding through crystallization, data collection, phasing, model fitting, model refinement and analysis, and description of the structure.

After five years of work and development, the macromolecular extensions to the CIF dictionary (mmCIF) were completed and presented to the community for review in August of 1995. The review process has resulted in a large number of changes, corrections, and additions, and we are extremely grateful to the many dedicated people who have looked carefully at the data model and its representation in the mmCIF dictionary, and who have made such cogent and thoughtful suggestions.

In March, 1996, the dictionary was opened to a wider audience for review and comment via announcements on several major bulletin boards. The dictionary itself, related documentation and examples, and related software and DDL information are publicly available on the World Wide Web at http://ndbserver.rutgers.edu/mmcif. Formal adoption of the dictionary by the IUCr is expected in mid-1996.

MSWK.CF.07 READING, WRITING AND VALIDATING CIFS USING CIFTBX2 AND CYCLOPS . Herbert J. Bernstein, Bernstein + Sons, 5 Brewster Lane, Bellport, New York 11713-2803, USA and Sydney R. Hall, Crystallographic Centre, University of Western Australia, Nedlands 6009, Australia.

The basic steps needed to adapt existing Fortran applications and write new applications which will manipulate CIFs are explained. We emphasize techniques needed to make applications compatible with both DDL1 and DDL2 CIFs. We discuss validating CIFs and more general STAR documents against dictionaries.

CIFs are becoming the standard for presentation of small molecules, and the pending adoption of the mmCIF dictionary by the IUCr is encouraging increasing use of CIFs for macromolecules. It is critically important for existing applications, such as molecular display programs, to be adapted to accept small molecule and macromolecule CIFs for input and to be able to produce CIFs as output in order to ensure a common interchange format among programs. Application programmers need to become familiar with the dictionary-based definition of CIF tokens and to design their applications so that the addition of new layered dictionaries will not require a redesign of code. We use the experience in adapting the DDL1 versions of several programs to a compatible DDL1/DDL2 environment to illustrate some to the practical issues involved. We show how tools, such as the extended version of CIFtbx2, a new version of a Fortran subroutine library for programmers developing CIF applications, and CYCLOPS 2, a new version of the very effective STAR data name checking program, can be used to make the transition to CIF and between DDL1 and DDL2 more manageable.

Issues that are addressed include managing large dictionaries efficiently with the use of hash-tables, the use of layered dictionaries, the implications of categories, and the implications of the more precise data types of mmCIF.

MSWK.CF.08 TRANSLATING PDB ENTRIES INTO MMCIF Philip E. Bourne, San Diego Supercomputer Center, PO Box 85608, San Diego, CA 92186-9785, USA, Herbert J. Bernstein, Bernstein + Sons, 5 Brewster Lane, Bellport, NY 11713-2803, USA and Frances C. Bernstein, Protein Data Bank, Chemistry Dept., Brookhaven National Laboratory, Upton, NY 11973-5000, USA.

The essential steps needed to map Protein Data Bank (PDB) entries into valid mmCIF data sets are discussed. Examples of converting both routine and complex structures using actual PDB entries with the program pdb2cif are given.

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. The pending adoption of the mmCIF dictionary by the IUCr, in response to the need to explicitly represent a larger amount of data which can be parsed by computer, (necessary as the number of structures continues to grow exponentially), has made translation from PDB format to mmCIF format a pressing issue.

In this talk we review the techniques needed to move from structures represented in PDB format to mmCIF format. Some data items have direct mapping with minor syntactic adjustment, such as for author names and journal references. Other data items, however, require us to recast our thinking along new lines. For example, the PDB format works with chains and HET groups, while mmCIF uses entities (discrete chemical components). Proper identification of entities in a PDB entry may require looking for sequence homologies. As another example, consider beta sheets. The PDB format treats a bifurcated sheet astwo distinct sheets which happen to have certain strands in common, while mmCIF allows all the strands involved to be represented as a single sheet. This requires strand matching and alignment to go from PDB format to mmCIF. What has currently been automated in pdb2cif and what still requires human intervention will be discussed.

Work supported in part by US NSF grant no. BIR 9310154 (for PEB), US NSF, PHS, NIH, NCRR, NIGMS, NLM and DOE under contract DE-AC02-76CH00016 (for FCB).