A3.6.2.2. Extensions for structural genomics

An International Task Force on Deposition, Archiving, and Curation of Primary Information for Structural Genomics was formed under the auspices of the International Structural Genomics Organization (ISGO) in 2001 (Berman, 2001) and was asked to develop specifications for data from structural genomics projects to be deposited with the PDB. The recommendations from this working group are summarized at http://deposit.pdb.org/mmcif/sg-data/xstal.html and http://deposit.pdb.org/mmcif/sg-data/nmr.html. For data from crystallography-based projects, the content extensions are largely focused on a more detailed description of phasing, tracing and density modification. All of the ISGO recommendations have been incorporated into the PDB exchange dictionary.

A3.6.2.3. Noncrystallographic methods

The IUCr-sponsored development of data dictionaries has been focused exclusively on crystallographic methods. As the repository for all three-dimensional macromolecular structure data, the PDB accepts structures determined using noncrystallographic techniques such as NMR and cryo-electron microscopy. The description of noncrystallographic methods is beyond the remit of the IUCr, so the PDB has worked with the NMR and cryo-electron microscopy communities to develop data dictionaries that describe these techniques within the mmCIF framework.

A3.6.2.3.1. *NMR*

The PDB exchange dictionary includes a description of NMR sample preparation, structure solution methodology, refinement and refinement metrics. These extensions were developed in collaboration with the BioMagResBank (BMRB; Ulrich *et al.*, 1989). The BMRB is the archive for experimental NMR data for biological macromolecules and has played an active role in the development of the mmCIF data dictionary. In selecting a format for archiving NMR data, the BMRB opted to use the STAR syntax (Hall, 1991) rather than the more restrictive CIF syntax. Despite this difference in syntax, the conceptual representation of macromolecular structure in the NMR dictionary (NMRStar) has remained semantically very close to the mmCIF representation. This has facilitated the exchange of data and dictionaries between the BMRB and the PDB, the sharing of software tools, and the development of a common platform for depositing data.

A3.6.2.3.2. *Cryo-electron microscopy*

Cryo-electron microscopy (as a technique for the determination of the structure of large molecular assemblies) is also described in the PDB exchange dictionary. The data extensions for cryo-electron microscopy include a description of the sample preparation, raw volume data (Henrick *et al.*, 2003), structure solution and refinement. These extensions have a prefix of `em_` (http://mmcif.pdb.org/dictionaries/mmcif_jims.dic/Index/).

A3.6.2.3.3. *Protein production*

The International Task Force on Deposition, Archiving, and Curation of Primary Information for Structural Genomics (Section A3.6.2.2) has also provided recommendations for the deposition of information about protein production. These recommendations are summarized at http://deposit.pdb.org/mmcif/sg-data/protprod.html. These data extensions have been used as the foundation for the Protein Expression Purification and Crystallization database (PEPCdb, http://pepcdb.pdb.org/) and for the protein production process model developed to support the Structural Proteomics in Europe initiative (SPINE; http://www.spineurope.org/).

A3.6.2.4. Supporting software

The RCSB/PDB has developed a set of software tools which support the PDB exchange dictionary framework (Chapter 5.5). These include *PDB_EXTRACT*, a tool to extract data from the output files of structure determination applications; *ADIT*, a web-based editor for data files based on the PDB exchange dictionary; and *CIFTr*, a translator from mmCIF to PDB format. These applications and other supporting utilities can be downloaded from http://sw-tools.pdb.org/.

The development of the mmCIF dictionary and DDL2 has been an enormous task, and any list of contributors to the effort will certainly be incomplete. Still, we must try. We have so appreciated the people that have taken the time to think carefully and constructively about all of this, and we would like to recognize their efforts. We begin by recognizing Syd Hall, David Brown and Frank Allen, who began the entire CIF effort and who recruited us to do the extensions for macromolecular structure.

Chapter 1.1 describes the formation of the original mmCIF working group, chaired by Paula Fitzgerald and including Enrique Abola, Helen Berman, Phil Bourne, Eleanor Dodson, Art Olson, Wolfgang Steigemann, Lynn Ten Eyck and Keith Watenpaugh. However, the number of people who contributed to the original design of the mmCIF data structure is much larger. We would like to thank Steve Bryant, Vivian Stojanoff, Jean Richelle, Eldon Ulrich and Brian Toby.

There are also the people who realized the shortcomings of the original DDL and worked hard to convince us that a more rigorous underpinning for the dictionary would be needed. Among them are Michael Scharf, Peter Grey, Peter Murray-Rust, Dave Stampf and Jan Zelinka.

Writing the dictionary and developing the new DDL were just the starting points for evaluation and critique, and this effort has been greatly aided by the input from COMCIFS, the IUCr committee with oversight over this process (David Brown, Chair). But the real process of review, after the dictionary was released to the public for comment in August 1995, has involved a much larger number of people. We cannot say enough about the valuable input we have received from Frances Bernstein, Herbert Bernstein, Dale Tronrud and Peter Keller.

Our efforts have been greatly enabled by the staff of the Nucleic Acid Database at Rutgers University, who have dealt with many of the technical issues of the implementation of mmCIF with real data. So we would also like to thank Anke Gelbin, Shu-Hsin Hsieh and Christine Zardecki.

Without the three CIF workshops described in Chapter 1.1, this effort would never have taken the shape and focus it now has, and we are eternally gratefully to Eleanor Dodson (York), Phil Bourne (Tarrytown) and Shoshana Wodak (Brussels), who organized the workshops, and also to Helen Berman and John Westbrook for hosting the subsequent workshop at Rutgers following the publication of the mmCIF dictionary. We thank the European Science Foundation (ESF), the European Union (EU), the National Science Foundation (NSF) and the US Department of Energy (DOE), who provided the funding.

The RCSB/PDB is operated by Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the Center for Advanced Research in Biotechnology/UMBI/NIST. RCSB/PDB is supported by funds