2.4. Specification of the Molecular Information File (MIF)

By F. H. Allen, J. M. Barnard, A. P. F. Cook and S. R. Hall

2.4.1. Introduction

This volume is primarily concerned with methods for the exchange of crystallographic information, such as experimental conditions and measurements, computational procedures and results, and the geometrical description of three-dimensional (3D) chemical structures. Such information, now available for some 400 000 compounds (Allen & Glusker, 2002) is, of course, vitally important in chemistry and in many other branches of science. However, it must be appreciated that two-dimensional (2D) chemical structural diagrams are available for over eight million compounds, and are fundamental components of the language of chemistry at all levels. Two-dimensional graphical representations indicate atomic connectivities, formal bond types and residual atomic charges, and provide the universal formalism through which chemists communicate with each other on a daily basis and document their results.

In common with scientists in other disciplines, chemists were early users of computer technology. They solved their information needs through the creation of major databases of chemical compounds and the development of methods for searching these databases for complete structures or substructural fragments. From these data, software can compute the 3D structures and properties of molecules, and the resulting molecular images can be displayed and manipulated. Consequently, 2D chemical diagrams are at the heart of many computerized documentation systems, and are the basis for computational chemistry applications that form part of the routine armoury of the modern chemist.

Computationally, the 2D diagram is treated as a mathematical graph (Harary, 1972). The nodes of the graph represent atoms and the edges of the graph represent bonds. Each of these primary components can have additional attributes: element type, valency, charge *etc.* in the case of the atomic nodes, and bond type, cyclicity indicators *etc.* in the case of the bonded edges. Within this formalism, 2D display coordinates and 3D crystallographic or computed coordinates are additional atom attributes, while interatomic distances can further qualify the bonded edges. Using the concepts of graph theory, it is then possible to write algorithms for the analysis of chemical graphs, *e.g.* for the detection of chemical rings and ring systems (*e.g.* Wippke & Dyott, 1975), for the analysis of functional groups and their relationships *etc.* Most importantly, procedures have also been developed for the matching of complete chemical graphs (full graph isomorphism), and for the location of chemical substructures within a complete chemical graph (sub-graph isomorphism) (*e.g.* Feldmann *et al.*, 1977). In this way it is possible to achieve graphical substructure searches of very large collections of 2D chemical diagrams.

As with early crystallographic data exchange, structural chemistry applications use their own specialized formats for input, manipulation and output. The ready exchange of chemical data is often inhibited by specific data formats and by the enormous variation in methods used to represent 2D structures, stereochemical descriptors and certain 3D structural attributes. These are computational 'bottlenecks' that detract from an effective use of the large financial and intellectual investment in proprietary software and database systems. They have also contributed to the major need for in-house format conversion software, which must be continually upgraded and maintained to accommodate developmental changes within imported systems.

The need for a universal interchange format for chemical information became apparent in the late 1980s, at almost exactly the same time as crystallographers recognized a similar need. Data standards in structural chemistry involve many international organizations and individuals, and consequently a number of proposals for exchanging data were initially put forward. From these the Standard Molecular Data (SMD) format (Bebak *et al.*, 1989; Barnard, 1990) emerged as the leading contender. In the early 1990s, discussions between the SMD and CIF developers led to a re-expression of the SMD data items within the Self-defining Text Archive and Retrieval (STAR) File syntax (Hall, 1991; Hall & Spadaccini, 1994).

This chapter describes the initial core data definitions of a universal exchange format for chemistry, the Molecular Information File (MIF: Allen *et al.*, 1995), that arose from this coalescence of concepts and ideas. MIF is a complementary approach to CIF (Hall *et al.*, 1991). Because SMD was fundamental to the development of MIF, we begin this chapter with a brief history of this project.

2.4.2. Historical background

The Standard Molecular Data (SMD) format was initially developed by a group of European pharmaceutical companies in the mid-1980s. Draft documents were made available from 1987 and the specification was published (Bebak *et al.*, 1989). A meeting in Frankfurt in 1988 established a series of technical working groups under the auspices of the Chemical Structure Association (CSA) to examine the format specifications in detail and to make recommendations for any revision. As a result, a draft form of a revised format, described as SMD Version 5.0, was published in February 1990 (Barnard, 1990). A document describing the core format, *i.e.* those data items regarded as essential in any exchange file, was prepared by one of us (JMB) for consideration by Subcommittee E49.51 of the American Society for Testing and Materials (ASTM).

In December 1993, the ASTM subcommittee E49.51 approved a standard specification for the content (*i.e.* recommended data items) of computerized chemical structural files (ASTM, 1994), although the subcommittee did not publish any proposals for a format specification. Later, the *Chemical Abstracts* Service (CAS) circulated a draft proposal for a connection-table-based exchange format for chemical substances and queries. It used some ideas that are similar to the 1990 SMD proposal and is expressed within the framework of the Abstract Syntax Notation 1 (ISO, 2002*a*,*b*). MDL Information Systems Inc. has also published a description of their proprietary formats (Dalby *et al.*, 1992) and a number of other software systems now provide interfaces to these formats.

Affiliations: Frank H. Allen, Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, England; John M. Barnard and Anthony P. F. Cook, BCI Ltd, 46 Uppergate Road, Stannington, Sheffield S6 6BX, England; Sydney R. Hall, School of Biomedical and Chemical Sciences, University of Western Australia, Crawley, Perth, WA 6009, Australia.