During this period, the IUCr Working Party on Crystallographic Information had commissioned one of us (SRH) to coordinate the development of a universal file to replace the existing fixed-format Standard Crystallographic File Structure (SCFS: Brown, 1988). As documented in Chapter 1.1, the CIF approach was adopted as the international standard in 1990 and published by Hall *et al.* (1991). Although the small-molecule CIF is able to store a representation of 2D chemical topology, its data definitions do not meet all the needs of the chemical community. In 1991, the IUCr became interested in further extending CIF into the chemical arena and discussions took place between representatives of the CIF project and of the SMD Technical Working Group. These meetings decided that an integration of the SMD format and the STAR syntax was desirable because it provided a number of advantages over the existing SMD specifications (Barnard & Cook, 1992). In particular, SMD/STAR provides for a clearer separation of the data structure and the data content roles, together with more flexible data extensibility in future versions. In addition, automated data validation of STAR/SMD files is possible using electronic data dictionaries. In a wider context, there were obvious opportunities for integrating with other applications of the STAR File.

2.4.3. MIF objectives

Molecular information embraces a broad spectrum of data related to chemical and molecular structure. It includes both individual and linked data items, *inter alia* spectroscopic measurements, thermochemical data, electrochemical properties, crystal structure information and so on. These items represent the data descriptors of molecular chemistry and it is intended that all of these will eventually be accommodated in the MIF approach. However, the initial MIF implementation (Allen *et al.*, 1995), summarized in this chapter, treated only the important core information: the data items needed to specify the connectivity and stereochemistry of molecules and their 2D and 3D spatial representations. The MIF data items needed for more extensive applications must, in the future, involve the collaborative efforts of informatics and database experts from chemical industry and academia.

A dictionary of the initial MIF core data items described in this paper is given in Chapter 4.8. This is the abbreviated text version of the definition attributes contained in the electronic dictionary file. The core MIF data items provide descriptors for representing the 2D connectivity of a molecule or substructure, the conventions for relative or absolute stereochemical relationships, and the coordinates and conventions used for the generation of 2D and 3D graphical depictions. These data items apply to complete molecules, or to substructures with incomplete or variable attributes. As a consequence they are well suited for query definitions in substructure search systems, a feature that will be discussed later in this chapter.

2.4.4. MIF concepts and syntax

The syntax of the Molecular Information File is based on that of the STAR File (Hall, 1991; Hall & Spadaccini, 1994). A MIF is an ASCII text file that can be read or amended using a standard text editor, and that can be processed computationally without conversion to another format. The organization and expression of MIF data is summarized in Table 2.4.4.1. Each file consists of a series of data blocks and each block consists of a series of individual data items. There may be any number of items within a block and any number of blocks within a file. A data block represents a logical grouping of data items and, in most MIF applications, a data block will usually specify a complete chemical entity, *i.e.* a fully defined molecule or a query substructure.

Table 2.4.4.1. *Brief overview of the MIF syntax*

A text string is a string of characters bounded by white space, single or double quotes, or semicolons in column 1.
A data name is a text string bounded by white space starting with an underline.
A data value is a text string not starting with underline, preceded by an identifying data name.
A list is a sequence of data names, preceded by '`loop_`' and followed by a list of data values.
A save frame is a collection of data within a data block, preceded by '`save_`*`framecode`*' and closed with '`save_`'.
A data block is a collection of data, preceded by '`data_`*`blockcode`*'.
A global block is a collection of data, preceded by `global_`, that is common to all subsequent data blocks.
A file may contain any number of data blocks or global blocks.
A data name must be unique within a data block.

The MIF syntax, unlike that of a CIF, places no restrictions on line lengths or nested loop levels. For a detailed understanding of the differences between a MIF and a CIF, the reader should compare this chapter with Chapter 2.2 or refer to the published details of the STAR syntax (Hall & Spadaccini, 1994), the specification of the CIF core data items (Hall *et al.*, 1991) and the Dictionary Definition Language (Hall & Cook, 1995) used to define data items in the electronic version of a STAR dictionary.

CIF data, described by over a thousand items in the current dictionaries (see Part 4), encompass the fields of crystallographic structure and diffraction techniques, and these data items could readily be incorporated into a MIF. It should be noted, however, that currently the reverse is not possible because the current CIF syntax does not support nested loops or save frames.

2.4.4.1. Data identification

The fundamental principle that underpins MIF is exactly as for CIF: every data item is represented by a unique data tag followed by its associated data value. These combinations are referred to as tag–value pairs or tuples. Data names must start with an underscore (*i.e.* underline) character and data values may be any type of string, ranging from a single character to many lines of text. Here are some simple examples of MIF data items:

```
_atom_mass_number       79
_atom_type              Se
_display_colour         blue_medium
```

The complete list of MIF core data items is given in Chapter 4.8.

2.4.4.2. Looped lists

Repetitive data are stored in a MIF as lists of values, as they are in a CIF. Each list is prefaced by a `loop_` statement and a sequence of data names that identify the data values that follow in 'packets' of equal length. The values in each packet match the order and number of the data names. Any number of packets may appear in a looped list.

Atom and bond properties are typical of the information to appear in a looped list. The atoms and bonds of thiabutyrolactone in MIF format are shown in Fig. 2.4.4.1. The description of each data item in this example is given in Chapter 4.8, although the meanings are clear from the self-descriptive data names. The number of data values in each list is an exact multiple of the number of data names at the start of each loop structure. Looped lists are terminated by the next list or by any other data name, data block or end of file. Comments may be included in a MIF and are preceded by a `#` character, as illustrated in Fig. 2.4.4.1.

Hierarchical data may require the use of nested loop structures (see the `_display_*` loop in Fig. 2.4.4.2). Note that the packet for `_display_id` of 7 has two sets of `_display_conn_` values giving