This was one of the issues confronting the Working Party on Crystallographic Information formed by the IUCr in 1988 (Section 1.1.6) to decide on the most appropriate universal file data language for crystallography from those under development (McCarthy, 1990). It is interesting historically to note that one of these was HGML – not the web markup language we know today, but the *Human Genome Mapping Library* language. Another language considered by the working party was ASN.1 (ISO, 2002) used by the National Institute of Standards and Technology and several US Government departments. It is an accepted ANSI and ISO standard for data communication and is supported by software, such as NIST's OSI Toolkit. ASN.1 possesses a rich set of language constructs suited to representing complex data, but suffers from data identifiers that are encoded and not human-readable, and a syntax that is verbose (particularly for repetitive data items such as those common in crystallography). These characteristics mean that a typical Protein Data Bank (PDB) file expressed in ASN.1 notation increases in size by up to a factor of 5. This was hardly an attraction in the 1980s when storage media were very expensive. Moreover, the ASN.1 syntax is not particularly intuitive, and is difficult to read and to construct. In contrast, the STAR File proposed at the first WPCI meeting had a relatively simple syntax, was human-readable and provided a concise structure for repetitive data. It also proved suitable for constructing electronic dictionaries, as will be discussed in later chapters. However, its serious and well recognized weakness in 1988 was that any recording approach using a simple syntax to encode complex data must involve sophisticated parsing software, and at that time only the prototype software (*QUASAR*; Hall & Sievers, 1990) was available. It was therefore not a straightforward decision for the WPCI to decide to recommend the STAR File syntax as a more appropriate data language for crystallographic applications. It was this decision that led to the development of the CIF approaches described in this volume.

2.1.3. The syntax of the STAR File

The syntax of the STAR File (Hall, 1991; Hall & Spadaccini, 1994) has been used to develop a number of discipline-specific exchange and archival approaches, including the Crystallographic Information File (CIF) (Hall *et al*., 1991), the Molecular Information File (MIF) (Allen *et al*., 1995), the dictionary definition language (DDL1) (Hall & Cook, 1995), the macromolecular dictionary definition language (DDL2) (Westbrook & Hall, 1995) and the STAR dictionary definition language (StarDDL) (Spadaccini *et al*., 2000). The details of the CIF, MIF, DDL1 and DDL2 approaches are given in Chapters 2.2, 2.4, 2.5 and 2.6, respectively.

A STAR File is a sequential file containing lines of standard ASCII characters. A file may be divided into any number of discrete sets of unique data items. Sets may be in the form of data blocks, global blocks or save frames. The syntax rules for these sets are given below in descriptive form. A more rigorous description of the STAR File syntax is given in Appendix 2.1.1 in extended Backus–Naur form (McLennon, 1983).

The STAR File is a free-form language in which spaces (ASCII 32), vertical tabs (ASCII 11) and horizontal tabs (ASCII 9) are collectively referred to as <blank>, and newlines (ASCII 10), form feeds (ASCII 12) and carriage returns (ASCII 13) are collectively referred to as <terminate>. White spaces <wspace>, used to separate lexical tokens within the file, are all characters in the joined set of <blank> and <terminate>.

2.1.3.1. Text string

A text string is defined as any of the following.

(*a*) A sequence of non-white-space characters on a single line excluding a leading underscore <_> (ASCII 95).

Examples:

```
5.324
light-blue
```

(*b*) A sequence of characters on a single line containing the leading digraph <wspace><'> and the trailing digraph <'><wspace>. <'> is a single-quote character (ASCII 39) and <wspace> is white space.

Examples:

```
'light blue'
'classed as "unknown"'
'Patrick O'Connor'
```

Note that the use of the <'> character in the text string that is bounded by a <'> character is not precluded unless it is immediately followed by <wspace>. The leading and trailing digraphs serve to delimit the string and do not form part of the data. In the above example the value associated with the text field 'light blue' is **light blue**.

(*c*) A sequence of characters on a single line containing the leading digraph <wspace><"> and the trailing digraph <"><wspace>. <"> is a double-quote (ASCII 34) character and <wspace> is white space.

Examples:

```
"low melting point"
"Patrick O'Connor"
"Doug Collins' crystal"
"classed as "unknown""
```

The use of the <"> character in the text string that is bounded by a <"> character is not precluded unless it is immediately followed by <wspace>. The leading and trailing digraphs serve to delimit the string and do not form part of the data.

The text strings of type (*a*), (*b*) and (*c*) cannot contain line-breaking characters, and therefore cannot span multiple lines. There can be more than one text string per line if each value is preceded by a data name, or the values are part of a looped list (see Section 2.1.3.5).

(*d*) A sequence of lines starting with <terminate><;> and finishing with <terminate><;>, where <;> is the semicolon character (ASCII 59).

Example:

```
; School of CSSE
  UWA
;
```

The requirement that the <;> character be the first character on the line does not preclude the presence of other characters on the same line, in as much as it does not violate the STAR File structure.

The leading and trailing digraphs delimit the text field and do not form part of the data. The character sequence between the digraphs, including any line-breaking characters, constitutes the value of the text field. The value associated with the above example is <blank>**School**<blank>**of**<blank>**CSSE** <terminate><blank><blank>**UWA** (note in particular that the <terminate> character preceding the final ; delimiter is *not* part of the value).

Copyright © 2006 International Union of Crystallography