SUPPORTING INFORMATION

Appendix A

Extensions to the 1991 STAR File format

The changes to the STAR File specification relative to the published version of 1991[1] are:

- The STAR character set has been extended to include all UNICODE characters except for a small number of non-printable or conflicting characters. The allowed character set for the STAR File is U+0009, U+000A, U+000D, U+0020 to U+D7FF, U+E000 to U+FFFD and U+10000 to U+10FFF of the UNICODE set. The use of U+0007 is allowed under restricted circumstances detailed in the paper.
- With the 1991 syntax, only save frames could be referenced, and then only within their parent data block. The extended syntax enables data items, individually or collectively, to be referenced across any file or cell partition. The encoding of the reference-value is detailed in the paper.
- With the 1991 syntax, save frames could not be nested. The extended syntax permits save frames to reside within save frames.
- The allowed character for whitespace delimited values has been restricted to additionally exclude; the left curly bracket `{` (U+007B), the right curly bracket `}` (U+007B), the left square bracket `[` (U+005B), the right square bracket `]` (U+005D), and the comma `,` (U+002C) and as leading characters the underscore _ (U+005F), the apostrophe `'` (U+0027), the quote `"` (U+0022), and the semicolon `;` (U+003B).
- Triple-quote or triple-apostrophe delimited values are introduced and may span more than one line and contain embedded `<newline>`s.
- Quotes and apostrophes (single or triple) may be now included within their respective delimited values by a preceding BEL character (U+0007). This *escapes* a quote or apostrophe character(s) from interpretation as the terminating delimiter.
- The *List* data type, as an ordered set of data elements, is introduced. A list starts with an left square bracket character `[` (U+005B) and ends with its *pair-matched* right square bracket character `]` (U+005D). The list elements are comma separated (whitespace between values has no meaning), and the lists may be nested. A list is a recursive data type.
- The *Table* data type, as an unordered set (associative array) of data elements each indexed by a string label, is introduced. A table starts with a left curly bracket character `{` (U+007B) and ends with its *pair-matched* right curly bracket character `}` (U+007D). The table elements are comma separated (whitespace between values has no meaning). Each element consists of a quote- or apostrophe-delimited index label, a separating colon character `:` (U+003A), followed by a STAR data value. A table is a recursive data type.

(1) Hall, S.R. The STAR File: A New Format for Electronic Data Transfer and Archiving *J Chem. Inf. Comput. Sci.* **1991**, *31*, 326-333.