

IEEE Standards Interpretations for IEEE Std 1003.2™-1992 IEEE Standard for Information Technology--Portable Operating System Interfaces (POSIX®)-- Part 2: Shell and Utilities

Copyright © 1996 by the Institute of Electrical and Electronics Engineers, Inc. 3 Park Avenue New York, New York 10016-5997 USA All Rights Reserved.

Interpretations are issued to explain and clarify the intent of a standard and **do not** constitute an alteration to the original standard. In addition, interpretations are not intended to supply consulting information. Permission is hereby granted to download and print one copy of this document. Individuals seeking permission to reproduce and/or distribute this document in its entirety or portions of this document must contact the IEEE Standards Department for the appropriate license. Use of the information contained in this document is at your own risk.

IEEE Standards Department, Copyrights and Permissions, 445 Hoes Lane, Piscataway, New Jersey 08855-1331, USA

Interpretation Request #42

Topic: Regular Expression Issues **Relevant Clauses:** 2.8.3

The scheme used for defining REs and their semantics is based on the historical notion that primitive REs match a single character. This was broken with the introduction of CEs, as they may be multiple characters in length. These were added through bracket expressions so as to cause as little impact to the rest of the specification. This had two side-effects. The first is interactions with the rest of the semantics. The second is that the historical and careful difference in semantics between BREs and EREs, which generally allowed BREs to be implemented with polynomial-time algorithms, was erased. Both BREs and EREs now in principle require algorithms with runtimes that are exponential in the size of the pattern. For example, on line 2860 (2.8.3.2), a bracket expression is defined to match a character or CE. Unfortunately, if a duplicated bracket expression contains multibyte CEs, then (exponential runtime) backtracking could be necessary to find the longest match.

The main problem with 2.8.3.1 is that it is unclear how to interpret matching a character or CE against a string. How is the string interpreted? As a sequence of characters, a sequence of CEs, or as a mixture? [9] On line 2860 (2.8.3.2), a bracket expression is defined to match a single CE. Yet, line 2890 allows matching any "character or CE".

Proposed solution: Change the phrase "character or CE" to "CE". Rationale The wording suggests a nonexistent distinction. [10] On line 2860 (2.8.3.2), a bracket expression is defined to match a single CE. Yet, on lines 2949-2952 and lines 2957-2969, frequent reference is made to "characters".

Proposed solution: Change lines 2949-2952 and lines 2957-2969 to refer only to CEs,

and not to characters. Rationale Same as previous rationale. [11] Within 2.8.3.2, the term “expression” is used often without a qualifier, and in these cases its meaning is unclear. Examples include lines 2886 and 2891.

Proposed solution: The apparent intended meaning, namely CEs, should be used.

Interpretation Response

Since the standard is unclear as to how strings are broken up into a series of collating elements (see interpretation #40) it is also unclear how to interpret matching a character or collating element against a string, and as such no conformance distinction can be made between alternative implementations based on this. This is being referred to the sponsor. Part10 The standard is unclear on this issue, and no conformance distinction can be made between alternative implementations based on this. This is being referred to the sponsor. Concerns about the wording of this part of the standard have been forwarded to the sponsor. Part 11 The reference to “expressions” on page 80, line 2886 refers to the definition on page 79, lines 2864- 2866, and conforming implementations must conform to this.

Rationale for Interpretation

None.