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

Interpretation Request #27
Topic: LC_COLLATE **Relevant Clauses:** 2.5.2.2 **Classification:** Ambiguous situation

(Subclause 2.5.2.2, LC_COLLATE, lines 1654-1658 in Draft 12)
"User- defined ordering of collating elements. Each collating element shall be assigned a collation value defining its order in the character (or basic) collation sequence. This or- dering is used by regular expressions and pattern matching and, unless collation weights are explicitly specified, also as the collation weight to be used in sorting."

Given this passage, assume there are two similar LC_COLLATE fragments. The frag- ments include lowercase letters only to simplify the examples.

Here is the first fragment: <a <a>;<a>;<a> <a-grave<a>;<a-grave>;<a-grave> <a-acute<a>;<a-acute>;<a-acute> <b <b>;<b>;<b> <c <c>;<c>;<c> <d <d>;<d>;<d> . . . <z <z>;<z>;<z> . . .

Here is the second fragment: <a <a>;<a>;<a> <b <b>;<b>;<b> <c <c>;<c>;<c> <d <d>;<d>;<d> . . . <z <z>;<z>;<z> <a-grave<a>;<a-grave>;<a-grave> <a-acute<a>;<a-acute>;<a-acute> . . .

Suppose a user wanted to find all words that begin with a letter in the range a-c. At the XoJIG meeting, we agreed that a locale built using the first fragment returns words that begin with <a>, <a-grave>, <a-acute>, <b>, and <c>. However, there were varying opinions about whether the second fragment would return the same results, or would exclude <a-grave> and <a-acute>.

So the question is this: Should an RE run against a locale built using the second frag-

ment include the accented a)s in the range because they are defined as being in the same equivalence class as <a>, or should it exclude the accented a's because they are listed outside the range of a-c?

Interpretation Response
The standard is ambiguous in this area, since it is not clear what the phrase "collation sequence order" means or is. The two possibilities are "the order in locale file", or "the order determined by the weights in the locale file". The standard allows either behavior. Concern over the wording of this area has been forwarded to the Sponsors of the standard.

Rationale for Interpretation
None.