UCA DUCET: change CE for U+FFFD

Author: Markus Scherer
Date: 2012-sep-14

Proposal

Map U+FFFD to the second-highest primary weight: FFD ; [FFE.0020.0002]

This is at the top of the Trailing Weights range. It leaves the highest primary weight for CLDR’s tailoring of U+FFFF.

Alternative

Remove the mapping for U+FFFD from the DUCET (allkeys.txt).

This gives U+FFFD an unassigned-implicit primary weight. (See UCA section 7.1.3 Implicit Weights.)

Rationale

U+FFFD REPLACEMENT CHARACTER sorts among symbols and is thus variable by default.

FFEE ; [*0862.0020.0012] # HALFWIDTH WHITE CIRCLE
FFFC ; [*14D7.0020.0002] # OBJECT REPLACEMENT CHARACTER
FFFD ; [*14D8.0020.0002] # REPLACEMENT CHARACTER
10100 ; [*02E0.0020.0002] # AEGEAN WORD SEPARATOR LINE

With the default alternate-handling=shifted, U+FFFD becomes tertiary ignorable. This contradicts UCA section 7.1.1 Handling Ill-Formed Code Unit Sequences:

An implementation of the Unicode Collation Algorithm may also choose not to treat ill-formed sequences as an error condition, but instead to give them explicit weights. [...] However, to avoid security issues when using this strategy, ill-formed code sequences should not be given an ignorable primary weight. [...] The first approach is to weight each maximal ill-formed subsequence as if it were U+FFFD REPLACEMENT CHARACTER. [...] It would be possible to give U+FFFD an explicit primary weight that is unlikely to be “variable”, that is, at least as high as the digits range. However, it is not obvious which primary to assign. It is easiest to give it a fixed value, predictably higher than other explicit values.

Removing its mapping would make it default to an implicit weight, still higher than explicit and Han-character primary weights but lower than supplementary-unassigned and trailing weights.