L2/19-065

Compatibility Ideographs and UAX #45

John H. Jenkins

12 February 2019

While reviewing and updating the UAX #45 database for release with Unicode 12.0, I noted an inconsistency. We have currently redefined the Status values of “U” and “A” through “F” to refer to characters encoded within one of the CJK Unified Ideographs blocks.

Unfortunately, we don’t have a status value to indicate that a character is encoded within one of the Compatibility Ideographs blocks. These currently have a status of “U.” There are 36 such characters encoded in the CJK Compatibility Ideographs block and an additional eight encoded within the CJK Compatibility Ideographs Supplement block.

Technically, since these are all by definition variations of encoded ideographs, we could set the status of all these characters to “V” and set the value of the Unicode field to the character for which these are compatibility variants.

E.g., UTC-00915 (丌) has a Status value of “U” and a Unicode value of “U+FA0C.” U+FA0C (丌) is a compatibility variant of U+5140 (丌). We could therefore set the Status to “V” and the Unicode value to “U+5140.”

Doing this, however, would lose information, namely that the character is formally encoded within the Standard. I think it would be useful to maintain the current Unicode values but add a new Status value of “Compatibility” (or simply “Comp”) to indicate that the character is encoded as a compatibility ideograph.

In this case, UTC-00915 would have its Status changed to “Compatibility” and no other alterations would be necessary.