7 May 2009

Proposal to Change the Syntax for the kIRG_*Source Fields in the Unihan Database

John H. Jenkins
jenkins@apple.com

Michel Suignard has finished work on the final Han data file for Amendment 6. This file is ultimately the source for the data published in the Unihan database via the various kIRG_*source fields.

One step we follow every time these data are updated for 10646 is cross-check with the Unihan database, reconciling differences and making additions.

This process is complicated by the fact that we use different formats for the individual fields. These format differences are generally not substantial. For example, the current 10646 source data for U+3400 is G_KX007801 for the G-source, T6-222C for the T-source, and JA-2121 for the J-source. The corresponding source data in the Unihan database are KX, 6-222C, and A-2121, respectively.

Although the differences are usually trivial, this does add an extra step to the reconciliation process, and every additional step is a potential source of trouble.

I recommend that we revise the contents of theses fields in the Unihan database (and UAX 38) to match the 10646 sources. This will mainly mean revising the official regular expressions giving the value syntax for these fields. It could conceivably break existing parsers.

I note particularly that the U-source syntax should be changed in any event. The current regular expression is "U\+2?[0-9A-F]{4}". Now that UTF 45 has been published, we should switch to using references to that instead of Unicode code points, and the regular expression should become "UTC[0-9]{5}".