L2/10-129R

Title: Response to N3819 "Preliminary Proposal for Encoding Special Scripts and Characters in UCS for Uighur, Kazakh and Kirgiz" (submission by Chinese NB, =L2/10-129)

Date: 2010-08-11 (revised date)
Source: US National Body & Unicode Consortium
Authors: Roozbeh Pournader and Deborah Anderson (SEI, UC Berkeley)
Status: Joint NB, Liaison contribution
Action: For review by WG2 experts
Distribution: WG2

Note: In this document "07XX" refers to the characters proposed in N3819, whereas "U+XXXX" are characters already in ISO/IEC 10646 (/Unicode). Although it is shown below that all the proposed characters are covered by existing characters in 10646/Unicode, the code points are incorrect in the proposal, because the code points U+076B..U+0772 are already taken, see http://www.unicode.org/charts/PDF/U0750.pdf.

I. All of the characters proposed in N3819 (=L2/10-129) are either already encoded (A, below) or capable of being handled as sequences of existing characters (B, below).

A. Characters Already Encoded

1. The document's "076C" is covered by U+06D5 ARABIC LETTER AE, which has the behavior as required by "076C." Note that U+06D5 has an annotation identifying its use for Uighur, Kazakh, and Kirghiz.

2. "076E" is covered by U+0649 ARABIC LETTER ALEF MAKSURA, which has all the required properties as requested. Older software handles this character improperly; originally the character was right-joining, but it has been changed to be dual-joining. Not all fonts or software reflect this change, which might have been the cause of the confusion.

B. Characters Covered by Sequences of Existing Characters

The remainder of characters proposed in N3819 are digraphs, which are already represented by sequences of two existing characters. There is existing text that uses these sequences. Separately encoding the characters as proposed In N3819 would mean there would be a duplicate way of representing the entities. Also, precomposed forms are no longer encoded for entities that can be represented by a sequence of existing characters, because to do so would break normalization rules, such as NFC.

"076B" = U+0626 ARABIC LETTER YEH WITH HAMZA ABOVE and U+0627 ARABIC LETTER ALEF
"076D" = U+0626 ARABIC LETTER YEH WITH HAMZA ABOVE and U+06D5 ARABIC LETTER AE
"076F" = U+0626 ARABIC LETTER YEH WITH HAMZA ABOVE and U+0648 ARABIC LETTER WAW
"0770" = U+0626 ARABIC LETTER YEH WITH HAMZA ABOVE and U+06C7 ARABIC LETTER U
"0771" = U+0626 ARABIC LETTER YEH WITH HAMZA ABOVE and U+06C6 ARABIC LETTER OE

> “0772” = U+0626 ARABIC LETTER YEH WITH HAMZA ABOVE and U+0649 ARABIC LETTER ALEF MAKSURA

II. Other comments
On page 2, the text reads:

> For example, under serial number 3, Arabic letter YEH (U+0649) has two variants, but its corresponding character in Kazakh and Kirghiz has four variants. In Uighur, it has Eight variants.

The text refers to U+0649, which is ARABIC LETTER ALEF MAKSURA (and not “Arabic letter YEH”). It mentions it two variants, but this character is dual-joining, so this comment doesn’t reflect the current standard.