L2/08-123

Date: March 3, 2008

Title: Comments accompanying the US negative vote on FPDAM5 to ISO/IEC 10646:2003
Source: INCITS/L2
Action: Forward to INCITS

The US National Body is voting *no* with technical and editorial comments on the following SC2 ballot:

> SC2 N3982: ISO/IEC 10646: 2003/FPDAM 5, Information technology - Universal Multiple-Octet Coded Character Set (UCS) - AMENDMENT 5: Tai Tham, Tai Viet, Avestan, Egyptian Hieroglyphs, CJK Unified Ideographs Extension C, and other characters

Acceptance of technical comments T3-T8 will change our vote to *yes.*

T1. U Source reference

The amendment currently adds the Unicode U source "UTC" with the reference "The Unicode Standard 5.1-2008." We recommend to replace this reference by "Unicode Technical Report #45, U-source Ideographs", as this new technical report provides more details about the origin of those characters.

T2. Ideographic Variation Database

The end of the third paragraph of Clause 20.5 and the following note currently read:

> Variations sequences composed of a unified ideograph as the base character and one of VARIATION SELECTOR-17 to VARIATION SELECTOR-256 from the Supplementary Special-purpose Plane (SSP) are registered in the Ideographic Variation Database defined by Unicode Technical Standard #37.
>
> > NOTE 2 - The Ideographic Variation Database is currently empty. When entries are registered, these variation sequences will be referenced by this standard.

Following the procedure defined by Unicode Technical Standard #37, a new version of the Ideographic Variation Database has been accepted on December 12, 2007. This version is identified as '2007-12-14', and contains 14,651 sequences, covering the repertoire of the Adobe Japan1 repertoire. We suggest to replace NOTE 2 by the following paragraph or similar text:

> This version of the standard incorporates by reference the variation sequences listed in version 2007-12-14 of the Ideographic Variation Database, as described at <http://www.unicode.org/ivd/data/2007-12-14>.

T3. CJK Compatibility ideographs

While we entirely agree with the goal of establishing round-tripping between the ARIB character set and ISO/IEC 10646, we believe that there is now a better solution than encoding compability ideographs.

Consider the case of 恵 and 恵 which are distinct in the ARIB character set. In the model of ISO/IEC 10646, those two forms are unified as U+6075. To achieve round-tripping, the two forms must be mapped to different sequences of ISO/IEC 10646 characters.

The usual solution is to encode a compatibility ideograph, U+FA6B in this case, and to establish a canonical decomposition of that compability ideograph into the unified form, U+6075 in this case. The purpose of the decomposition is to account for the unification. Under that solution 恵 is mapped to the sequence <U+6075>, and 恵 is mapped to the sequence <U+FA6B>. However, this approach imposes a very severe constraint on implementations, as they can never normalize data; any normalization transforms <U+FA6B> into <U+6075>
and prevents round-tripping. Essentially, the canonical decomposition defeats the purpose of the compatibility ideograph.

With the advent of variation sequences, we have a better solution at our disposal. Indeed the variation sequence <U+6075, U+E0100> is targetting the form 恵 and the variation sequence <U+6075, U+E0101> is targetting the form 恵, so the ARIB characters can be mapped to those sequences and support round-tripping. Unlike the sequences of the usual solution, these variation sequences remain unchanged by normalization. This gives a much greater freedom to implementations.

The Ideographic Variation Database also contains sequences for the other three pairs of ARIB characters which are unified in ISO/IEC 10646.

In conclusion, we believe that the proposed characters U+FA6B..FA6E fail to effectively achieve the goal of round-tripping the ARIB character set, and that this goal can be achieved today using variation sequences already in the Ideographic Variation Database. We propose to not encode those four characters.

T4. Conflicting sources

The editor's note at the bottom of page 2 mentions that there is unresolved conflicting information concerning KangXi source references. We would like these conflicts to be resolved before further progression of the Amendment.

T5. Names of Hangul jamos

The names of the three characters 11FD, A96E and A973 should have an additional "H" at their end. The correct names are:

11FD HANGUL JONGSEONG KIYEOK-KHIEUKH
A96E HANGUL CHOSEONG RIEUL-KHIEUKH
A973 HANGUL CHOSEONG PIEUP-KHIEUKH

(Apparently, the original version of the proposal WG2 N3168 had incorrect names, which in turn led to incorrect names in WG2 N3242, which is what was accepted by motion M50.34. As described in WG2 N3257, a revision of WG2 N3168 included the correct names as above.)

T6. Avestan separation point

The US NB remains opposed to the encoding of yet another middle dot punctuation at position 10B38 (AVESTAN SEPARATION POINT).

T7. Archaic Sinhala numerals

The US NB has received information that indicates that more investigation is needed for the Sinhala archaic digits and numbers (0DE7-0DEF and 0DF5-0DFF). The US NB would like those characters to be moved to a future amendment

T8. Tai Tham

The US NB supports the recommendations of the Tai Tham ad-hoc meeting as documented in WG2 N3379, as well as the inclusion of the two additional characters requested in WG2 N3384.

E1. Incorrect U Source header

On page 5, the header of the additional code chart fragment for the new characters 9FC4 and 9FC5 is "U Unicode". Those characters only have a J source, so the header should be "J".