ISO/IEC JTC1/SC2/WG2/IRGN2109

Title: Character Duplication in CJK Unified Ideographs
Source: Henry Chan
Status: Individual Contribution

Background

Currently, there are wide-spread unifiable character duplication in the CJK Unified Ideographs. Besides those intentionally disunified by the Source Separation Rule, there are numerous cases of character duplication, which include exact duplicate, minor glyph variation, and “Disunified” characters identified in the UCV.

Furthermore, for CJK Extension D, many “new characters” added were actually variants of existing frequently used characters. The variations in these characters could be uniformably applied to many other characters containing the same component. For example, U+2B779 𫝹 which is a variant of U+5ff5念. There are also over 50 characters containing the component念. Could the characters with U+5FF5念component swapped with U+2B779 𫝹 all be added to CJK Unified Ideographs? That would cause the number of CJK Unified Ideographs to increase.

Problems

There are over 80,388 CJK “Unified” Ideographs but in fact many characters are not unified. The UCS character standard seeks to encode abstract characters, not the glyph forms. As the historical compatibility issue has been addressed with the Source Separation Rule, there is no further reason to deviate from the nature of the UCS character standard and encode an increasingly large amount of new code-points for characters which look similar to existing characters.

Technically, an increasing number of “Unified” Ideographs makes the full implementation of pan-CJK generic-use font very costly if not impossible to implement. Font foundries often have to reference to using only specific versions or planes of national encoding standards to limit the number of characters they need to cover.

Also, text processing becomes a big challenge. When searching for a certain character, the user may not be aware that another similar character in the text would have a different code-point. The user may also be confused why he can see a particular character, but when he/she input the character into the word processor search function via a specific IME, the word processor cannot find the character. The word processor would need to maintain a table of what different code-points the user may expect them to be the same character.

It is not meaningful to encode different glyph forms of the character at different code-points. For information exchange, the encoded character should be representative of its semantic meaning, while use of Ideographic Variation Selector to choose its exact glyph form if necessary. When using Ideographic Variation Selectors, word processor can simply ignore the Selector when doing comparison instead of matching every single character against a list of duplicates.

Proposal

I propose that the IRG maintain a list of confirmed duplicate characters in a separate section in the errata report. Also, a procedure could be established for member bodies and individuals to submit character duplication "suggestion". After verification or confirmation by other member bodies, the character duplication pair would be marked as "confirmed" and added to the errata report. These duplicate characters should cover exact duplicates (glyph exactly the same), and disunified-in-error characters.

In the past, due to taboo certain strokes were omitted from characters in printing the Kangxi dictionary. Many of these characters have been unnecessarily disunified in Extension B – they exist in Extension B for no other purpose but to show an exact form, which would be better handled now by Ideographic Variation Selectors. These character variations have yet to be added to the UCV. For example, u+248e5 𤣥 is the Kangxi Dictionary's taboo substitution for U+7384玄. There is no semantic difference between the two characters.

Currently, we cannot find a consolidated list of possible duplicate characters on the IRG website nor find the official list in the Unicode charts provided by the UTC. It is hard to track which characters have been identified and proven.

By providing such a list, word processor can use this list to do a "smart search" by finding the similar characters the user expect to find. Furthermore, font foundries can issue a general-purpose fonts that uses the same glyph for the different code-points. This is useful for commercial printing as the editor need not care about the specific code-point used; it will always display consistently. Content management systems can also use this list to canonicalize all variant characters to a single character for searching.

Second, my opinion is that member bodies should decide to unify more. If it is known that certain new submitted characters are cognate, semantically equivalent and in similar shape to existing characters, it would be better to unify them, and encode them via Ideographic Variation Sequences and nominate immediately for UCV. For example, U+ 2B7D2 𫟒 is a variation of U+8449葉. In fact, U+4E17 丗 is a known semantic variant of U+4e16世. In this case, it should be suggested that U+4E17 丗 and U+4e16 世 be added as unifiable pairs.

Furthermore, if the character is a Song-ification of Running-script forms which are not well standardized, it is better to unify it and add it via the IVS. At the worst case, the encoded character will not display in the intended form, however there will be no information loss and it can be converted to the new code-point if proven non-cognate.

In the past, there have been a few incident of erroneous unification. However, it is only a very small number compared to the extraneous disunification. Given that much of the submitted entries are dictionary entries, my opinion that member bodies should may choose unify with greater confidence that there will not be over-unification.

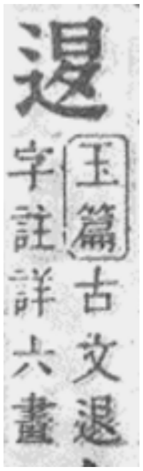
Duplicates Identified (Example)

1. 𨓤 (U+284E4) and 𨓆 (U+284C6) (New pair)

𨓤 (U+284E4)
kIRG_TSource: T6-4553

𨓆 (U+284C6)
kIRG_GSource GKX-1257.12

However, according to the KangXi Dictionary,



𨓤
玉篇古文退字註詳六畫

The bottom right component should be 夂 (foot) instead of 夕 (night). U+284E4 and U+284C6 are the same character.

2. 𣁋 (U+2304B) and 𢼸 (U+22F38) (Existing Pair)

𣁋 (U+2304B)

kIRG_GSource	GKX-0477.03
kIRG_HSource	H-FDC2
kIRG_TSource	T6-382E
kIRGKangXi	0477.030

𢼸 (U+22F38)

kIRG_GSource	GHZ-21457.04

In mainland china, most character containing component of "儿" would be replaced with "几". U+22F38 should have been unified with U+2304B. This character is also listed under the UCV as "Disunified", but in fact would be better classified as "duplicate".