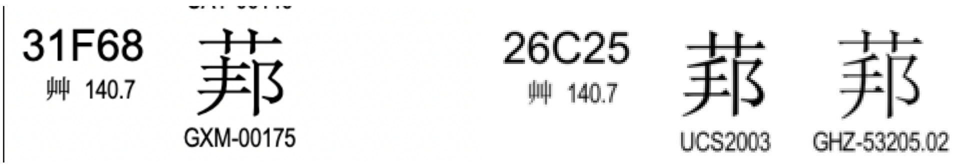
Title: **Errata of IRG WS2017 v7.0 (CJK Unified Ideographs Extension H)**
Type: Individual Contribution
Source: Henry Chan
Date: 2022-05-02 (updated 2022-05-08)
Action: For consideration by JTC1/SC2/WG2/IRG
Pages: 2

Summary

There are two characters in IRG WS2017 v7.0 which are unifiable with existing characters and should be removed.

In case re-sorting all characters would upset the stability of Extension H, which is due to be published with Unicode 15, it is suggested to leave the codepoints as-is and leave two gaps. This is identical to the treatment of U+2E9A which is a gap introduced into the CJK Radicals Supplement block because a character was unified with a character in the Kangxi Radicals block.

1. U+31F68 ⿱艹邦 (GXM-00175) and U+26C25 ⿱艹⿰手阝


31F68
艸 140.7
GXM-00175
26C25
艸 140.7
UCS2003
GHZ-53205.02

The character GXM-00175 (⿱艹邦) is unifiable and duplicated with U+26C25 (⿱艹⿰手阝).

A comment was posted on the IRG review tool by HKSAR (#4549) and also me (#4828) in version 3.0 suggesting unification of GXM-00175 to U+26C25, but was not discussed during the IRG 52, most likely due to lack of time due to the sheer number of comments on version 3.0.

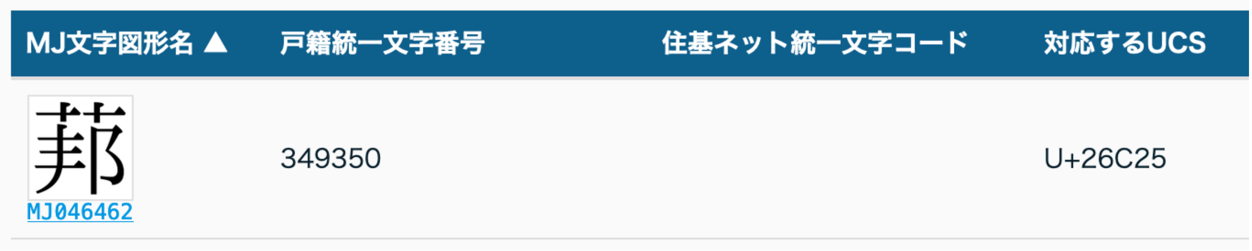
Tao Yang (China) for WS2017 v5 (#9090) replied that it was agreed that the bottom of U+26C25 is indeed etymologically 邦, asserted that GXM-00175 is the “correct” form, and asked for the solution where the “wrong” form is already encoded.

By IRG conventions, normally the character is (1) withdrawn (if there is an existing UCV rule) or unified to the existing character (if there is no existing UCV rule), and (2) the source reference and/or glyph of the existing coded character is replaced.

This was initially targeted for IRG 55 discussion but IRG 55 was cancelled due to COVID-19, and comments not re-posted for v6.0 were not discussed by IRG 56.

I posted a follow-up comment (#10050) in Nov 2021 which shows a source from《字彙補》with the same content as that from Hanyu Dazidian and has an identical form with GXM-00175, which furthers solidifies that the two forms are the same character without a doubt.

Also, an identical glyph in the Moji-Joho database has been coded at 349350 and mapped directly to 𦰥 (U+26C25):



MJ文字図形名 ▲	戸籍統一文字番号	住基ネット統一文字コード	対応するUCS
𦰥 MJ046462	349350		U+26C25

The entry MJ046462 is currently mapped to both the Daikanwa Jiten at entry 31125 and the Koseki character database at entry 349350. The changelog indicates that the mapping to U+26C25 was added in Ver.002.01, which was released in June 2012.

The glyph is mapped directly to U+26C25 and is not registered as an ideographic variation sequence. The existing G-source form for U+26C25 does not appear in the Moji-Joho database at all.

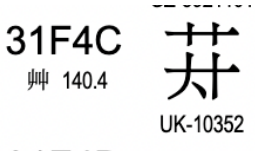
A glyph for Koseki character database entry 349350 was added to Glyphwiki even earlier in April 2012, and it was already mapped to U+26C25 by then. A glyph for Daikanwa Jiten entry 31125 was added in Glyphwiki in Sept 2018, also mapped to U+26C25.

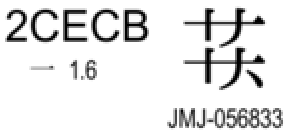
Unfortunately comment #10050 was missed when discussing WS2017 in IRG 58, because it was tagged as a comment on WS2017 v6.1 instead of v7.0 due to a misconfiguration in the online review tool.

A discussion record “not unified to U+26C25, irg50.” exists, but in IRG 50 this was the first review of WS2017, and all unification comments from Kawabata-san were resolved to be not unified if the glyph shapes of the proposed unification characters were not exactly identical. The fact that there is an existing mapping in the Moji-Joho database was not yet known at that time. The comments in v3.0 and later were left out of discussion by IRG by accident.

Encoding GXM-00175 separately in Extension H would be an unwanted de-facto disunification, and also upset existing encoded data based on the Moji-Joho database. It is unfortunate that this character existing in the Moji-Joho database was not horizontally extended by Japan in ISO10646, otherwise this issue would have been caught much earlier.

2. U+31F4C ⿱卄⿻廾丶 (UK-10352) and U+2CECB = ⿱サ⿻サ丶





Both characters are short form of 菩提 used in Buddist scriptures. It is better for UK-10352 to be unified with U+2CECB and encoded in the IVD instead of as a separate character.

Thanks to extc for alerting me to this issue.