Universal Multiple-Octet Coded Character Set

International Organization for Standardization

Organisation Internationale de Normalisation

Международная организация по стандартизации

Doc Type: Working Group Document

Title: Proposal to De-Unify One Obsolete Simplified Chinese Character

Source: Alexander Zapryagaev

Status: Individual Contribution

Date: 2019-09-30

Contents

1. Background....2
2. Character 1.1.136 (simplification of U+96EA)....3
3. The Matter of Other Unifications....7
4. Conclusion....8

Appendix 1: Font Data....9

Appendix 2: References....9

1. Background

Second stage simplifications (第二次　汉字简化方案一草案, SSS) were an abortive project of PRC government in late 1970s. They were supposed to become a continuation of the highly successful campaign of simplifications implemented throughout the 1950s-60s.

The campaign was split in two parts:

- Part One was released on 20th December 1977 and was consistently used in all the publications in 人民日報 *Rénmín Rìbào* until July 1978. It was widespread during that period and gained mass currency, but, announced a failure, soon practically dropped out of usage, though an official withdrawal (but not declaration of abandoning any simplification plans) was postponed until 24th June 1986. Many of the forms proposed either were previously existing popular forms of characters and/or remained in everyday use, such as in private letters and in signage, even after stopping the project.
- Part Two was published simultaneously with Part One but declared not for immediate use but rather for implementation in case the first one succeeds. Currently, unlike the first part, these characters are not recognized by the majority of the literate Chinese.

The journey towards encoding these simplified forms in Unicode started almost exactly ten years ago with the inaugural proposal JTC1/SC2/WG2 N3695 by Andrew West, accessible at

http://std.dkuug.dk/jtc1/sc2/wg2/docs/n3695.pdf

and aimed at such uses as the correct representation of the texts published during the short implementation period. An additional upside of such an implementation would be the ability to render correctly some of the ancient manuscripts and printed works which originally, informally used the simplifications later incorporated in the system of the SSS.
The detailed situation with the Unicode encoding is summarized in the current author's document, *The Chart of Current Status for Second Stage Simplification in Unicode*, accessible in its up-to-date version at

https://drive.google.com/open?id=1e51AuBE7_G2bvkNUsMgmuEeddZdBCSIr

page. Currently, the characters from Part One, those actually appearing in print beside the table itself (but not the result of applying the guidelines for mass simplification, also given in the table, unless the table itself explicitly mentions them), are all in pipeline for the inclusion in CJK Extension

G. This proposal will refer to one controversial decision of unification made during this inclusion and argue for disunification and adding one more character to the extension.

2. Character 1.1.136 (simplification of U+96EA)

This is the excerpt from the Table of Part One simplifications, showing the unsimplified and simplified forms of the character U+96EA, 雪, side by side:

Fig. 1. U+96EA (雪) and its simplification, 彐

As the comparison between the two forms obviously shows, the intention of this simplification is to retain only the lowest part of the glyph 雪.

Together with the other characters from Part One of the table, this one was included with the UTC source identifier in the range 00953–01178, which encompasses the various proposals of Andrew West given in the UTC document L2/12-333, *Request to UTC to Propose 226 Characters for Inclusion in CJK Extension F*. Its number is UTC-01005.

However, during the following stages, the glyph was deleted from the proposals, with the USourceData.txt file giving the reasoning as “Encoded in the URO” (not even “Unifiable”, but “Encoded”). The apparently existing codepoint was given as U+5F50 彐. It is one of the 8 characters identified as already encoded in the URO from this request: the rest are non-controversial. 钖汪贲黉镉鲋铱 was encoded as U+9FCF, U+9FD1–U+9FD5, and U+9FED respectively due to other processes.

I argue that the decision was wrong and the character proposed as UTC-01005 彐 is not a duplicate of U+5F50 彐.

Consider the relevant unification rules (R1, Source Separation Rule, is decommissioned):

R2. Noncognate Rule. In general, if two ideographs are unrelated in historical derivation (noncognate characters), then they are not unified.

R3. By means of a two-level classification (described next), the abstract shape of each ideograph is determined. Any two ideographs that possess the same abstract shape are then unified provided that their unification is not disallowed by either the Source Separation Rule or the Noncognate Rule.

As of 2018-01-29, the following two rules were designated to "reduce the number of encoded variants": one unifies

1. **characters that have a different structure, but whose difference is not considered significant enough to encode them as separate unified ideographs, and for which strong evidence associating them as variants of encoded characters can be provided.**
2. **characters with the same structure, but with different components at the second (or subsequent) level that may not be generally unifiable, and for which strong evidence associating them as variants of encoded characters can be provided.**

The shared structure of the two representative glyphs under question is repeated here in close-up in BabelStone family of fonts:

Fig. 2. The glyphs of U+96EA (for comparison), UTC-01005, and U+5F50.

Indeed, they resemble a case of unification by structure (Rule R3) as depicted in Table 18-6 on p. 716 of The Unicode Standard Ver. 12.0, namely, unifiability by "[d]ifferences in protrusion at the folded corner of strokes", illustrated there with the glyphs

鉅　鉅

Fig. 3. The glyphs for U+9245 (Japano-Korean vs. Mainland), unified according to Rule R3.

for U+9245[1]. However, the first impression is tricky.

Consider the *semantics* of the given characters. The semantics of ⺕ is obviously given by its simplification relationship with its original form, 雪, as identical to it in all but the actual property of simplification. This is confirmed by the entry for the character in *Zhōnghuá Zìhǎi*:

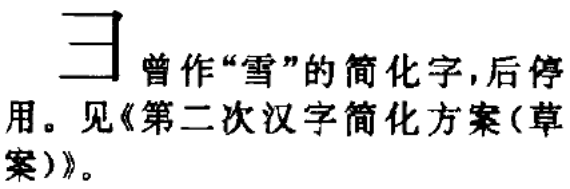

Fig. 4. Excerpt from *Zhōnghuá Zìhǎi* (p. 657) for ⺕: 『曾作“雪”的简化字，后停用。』

Note that this dictionary gives the character separately from ⺕, situated right above it. It inherits from its unsimplified version such properties as the general meaning “snow” and reading *xuě*.

Meanwhile, the semantics of ⺕ are defined on its own, independently of its connection to any other characters. The character ⺕ is a representation of Kangxi radical #58, U+2F39 KANGXI RADICAL SNOUT in the URO, with a conventional Pǔtōnghuà reading *jì* and Cantonese reading *gai*^{3}; it has reconstructible Middle Chinese (MC) reading and the ability to express the concept of “pig head” all of itself, though today it is rarely used for such a reference. By application of Rule R2, we find no reason toward unification. The changes introduced in 2018 are currently irrelevant, as they are to prevent the multitude of variants and put it upon the side proposing

[1] Even in such a case, the correct notation would be “V” for “Variant of an encoded character”, not “U” “Encoded in the URO”, apparently designed for the cases when the UTC proposal was *appended to the end* of the URO, not discovered in it by search. Nevertheless, the “V” notation is not used in the span under question even once.

unification to provide “strong evidence associating them as variants of encoded characters”, which is this case cannot be provided as the evidence is trivially opposite.

But what if it is still possible to unify the characters, at least graphically, just to stop the proliferation of characters with minor graphical differences? This, however, is also impossible, according to Dr. Ken Lunde’s database, *IICore2020*. If we consult it, we find out that while U+96EA has the kIICore2020 property of “GHJKMPT”, pointing at universal usage, U+5F50 is marked “GH”. This is rarer, but, vitally, still contains the letter “G” (Mainland use). The absence of the entry for the property kTGH under U+5F50 in Unihan_OtherMappings.txt file points at the source for U+5F50’s appearance: the basic coded standard set for Mainland China, GB 2312, contains the glyph now represented by U+5F50.

And this means: even it the most basic plaintext representation it is possible to encounter a Chinese text for Mainland use that contains both of the glyphs under question – at the same time. It is sufficient to mention “snow” somewhere and also discuss the Kangxi radicals; 彐部 *jìbù* is actually a dictionary word for the radical, according to wenlin.co. Even in a thorough enough text on the SSS itself, that lists the new forms of glyphs and gives the radical and stroke information for them, it is required to maintain the distinction between them. It seems unfeasible to maintain a Variation Selector for such a purpose, as turning “snow” into a “pig’s head” is outside the scope of VS mechanisms. Hence, the existence of the two glyphs under separate codepoints is the only available choice.

Additionally, it is possible to understand Rule R2 not synchronically, but diachronically: if the forms of 彐 and 彑 are the same *in historical derivation*, modern usage notwithstanding, the case can still be done for unification. The meaning of 彑 as “pig’s head” or “snout” has already been established and is even more obvious in its graphical variant, U+5F51 𢆶, given below in its Small Seal form.

Fig. 5. Small Seal (小篆) forms of “snout” and “snow” sinograms, courtesy wiktionary.org

Meanwhile, the glyph 雪 is composed phono-semantically, with the phonetic part being U+5F57 彗 “broomstick; comet”. Investigating further, we find the bottom part of 彗, which also is a phono-semantic compound, is a form of “hand” U+53C8 又, in Small Seal script  . The fact that, according to Unicode code charts, Hong Kong, Taiwan and Korea representative forms produce the extended central stem, 雪 instead of 雪, which is also the form from 康熙字典, not only supports this derivation but successfully demonstrates the absence of any historical connection between accidentally similar glyphs.

3. The Matter of Other Unifications

One should notice the discussion above does not make a case for the remaining two unifications that happened during the preparation of the Working Set 2015, precursor of Extension G.

- UTC-01024 毁 was unified at a much later stage of consideration: the IRG #48 Liaison Report of 2017-06-25 offers the unification with U+6BC0 毀 or U+6BC1 毁. “Justification: two SAT-submitted characters in Extension F were unified with U+22758 for the same reason, and a new UCV may be added.” As a non-compulsory comment, I would choose the more similarly-looking non-simplified form of U+6BC0 for the unification, as the attachment of the kSimplifiedVariant and similar properties is already enough of a mess (with the appearance of sequences traditional – simplified – Second Stage simplified), while using U+6BC1 would prevent using the same font to depict forms before and after SSS without resorting to an immediate registration of an IDS (which should still be registered, nonetheless).
- The simplification of U+8D5B 赛, 宷, was excluded from the L2/12-333 proposal due to its unifiability with the already encoded U+219F3 𢇁. This is a correct solution, as not only the graphical difference is down to variants of the same sub-component, the semantics are exactly the same: 𢇁 is a Singaporean simplification of the same character (used before moving to Mainland scheme) and was encoded as such. Note that the additional Sawndip usage of 𢇁 is now irrelevant due to different script. Still, an IVS should rather be registered.

4. Conclusion

This document now proposes:

1. Encode the character UTC-01005 with the representative glyph ⺕ in the Extension G to maintain the integrity of extension. Change the designation “U” to “G” in the file USourceData.txt.

Additionally, with less urgency,

2. When choosing among U+6BC0 毀 and U+6BC1 毀 for the unification of UTC-01024 毀, prefer U+6BC0; simultaneously register an IVS for the same glyph to distinguish graphical variants.
3. Add an IVS to the existing character U+219F3 𡈼, which chooses the form 宙.

Appendix 1: Font Data

This document has been typeset in EB Garamond. The Chinese characters were typeset in `Babelstone Han`. The SSS forms were typeset in `Babelstone Erjian 2`. Whenever a contrast was required between various local renderings of glyphs, Source Han Sans or Serif was used.

Appendix 2: References

1. 第二次　汉字简化方案一草案. Archived from archive.org.
2. West, Andrew. *Proposal to Encode Obsolete Simplified Chinese Characters.* Available at http://std.dkuug.dk/jtc1/sc2/wg2/docs/n3695.pdf.
3. Zapryagaev, Alexander. *The Chart of Current Status for Second Stage Simplification in Unicode.* Ver. 2.0. Available at https://drive.google.com/open?id=1e51AuBE7_G2bvkNUsMgmuEeddZdBCSIr.
4. The Unicode Standard. *U-Source Glyphs.* Available at http://www.unicode.org/Public/UCD/latest/ucd/USourceGlyphs.pdf.
5. West, Andrew. L2/12-333, *Request to UTC to Propose 226 Characters for Inclusion in CJK Extension F.* Available at https://unicode.org/L2/L2012/12333-cjk-f.pdf.
6. The Unicode Standard. *USourceData.txt.* Available at http://www.unicode.org/Public/UCD/latest/ucd/USourceData.txt.
7. The Unicode Consortium. *The Unicode Standard, Version 12.1.0*, (Mountain View, CA: The Unicode Consortium, 2019. ISBN 978-1-936213-25-2). Available at http://www.unicode.org/versions/Unicode12.1.0/.
8. Unicode Technical Standard #37: *Unicode Ideographic Variation Database.* Available at http://www.unicode.org/reports/tr37/tr37-12.html.
9. Leng, Yulong, and Yixin Wei, eds. *Zhonghua zihai.* Zhongguo youyi chuban gongsi, 1994.
10. *Wiktionary.* Available at https://en.wiktionary.org/.
11. Lunde, Ken (Adobe). *Proposal to define new Unihan Database property: kIICore2020.* Available at https://www.unicode.org/L2/L2018/18279-iicore-2020.pdf.
12. *Wenlin dictionary.* Available at https://wenlin.co/.
13. *IRG2015.xlsx.* Available at http://www.babelstone.co.uk/CJK/IRG2015/IRG2015.xlsx.