1. Background

1.1 a related document

ISO/IEC JTC 1/SC 2/WG 2 IRG N1859
Date: 2012-06-07
Source: SUZUKI Toshiya,
Title: Proposal for the Discussion How to Handle the Mistakenly Differentiated Glyphs in Huge Dictionaries

1.2 Relevant portion extracted from IRG N1859

... 

Taking an example "جا" (U+23A26), following evidences are found. Even for the publishing of Kangxi in different formats, these differences are not regarded as the shapes to be distinguished.

- The handwritten copy of Kangxi in Siqu Quanshu uses U+23A26 shape at the place corresponding to G_K0584.12
- The digitally typesetted version of Kangxi published by Shanghai Lexicographical Publishing House uses U+23A26 shape at the place corresponding to G_K0584.12
- The digitally typesetted index added to Kangxi published by Zhonghua Bookstore includes U+23A26 but not G_K0584.12 If these shape differences are regarded as different base characters, the impact may introduce the incompatibility among the existing and future versions of digitized dictionaries.
Proposed Item for the Discussion

Because the glyphic differences are not found in the precedent unifications, it is questionable if they should be dealt as generally unifiable difference. In the case of the difference ... difficult to take as generally unifiable. Therefore, “when two variants are found to be caused by the mistake in the dictionary compilation, and their difference is difficult to take as generally unifiable, how the variants should be handled?” is expected to be discussed by the experts. The possible options would be following:

(A) apply non-generic unification, and code at CJK Compatibility Ideograph.
(B) apply non-generic unification, and register the shape in IVD.
(C) classify the evidence taken from the dictionary as unreliable, and postpone the discussion until yet-another evidence to justify the separated encoding of the variants.
(D) code the variants separately, but make a record that these variants are cognate.

In summary, my proposed option is (C), and I want to receive the feedback about which option is the best for the submitters, with the concrete use cases.
2. ROK comments

- In principle, ROK agrees with Mr. Suzuki's proposed option C.

- ROK suggests that "yet-another evidence" be elaborated as follows:

  1) Glyphs in other dictionaries are not to be considered as yet-another evidence in general.

  - It seems desirable that actual usage in ordinary books or documents (other than dictionaries) is considered as yet-another evidence.

  2) However, suppose that none of two mistakenly differentiated glyphs in a Hanzi dictionary (for example, Kangxi) are encoded in UCS (this situation will occur rarely); furthermore, no actual usage other than dictionaries exists.

  In this situation, if we want to encode one of them in UCS, then we can probably accept usage in other dictionaries as yet-another evidence.

* * *