Comments on PRI 108: Combined registration of the Adobe-Japan1 collection and of sequences in that collection

Masahiro Sekiguchi
Fujitsu Limited
<seki@jp.fujitsu.com>

Please note that the opinion expressed here is the author's own, and does not represent that of his employer's and/or any organization that he or his employer participates.

The author of this comment finds the following two problems in the PRI 108. He believes the submission should be amended appropriately before accepted by the Unicode Consortium for registeration.

PROBLEM 1: The current submission is unclear regarding what is the coverage of each proposed glyphic subset.

As specified in UTS#37, the purpose of an IVS is to restrict the possible glyphs for a given CJK ideograph as a Unicode character. The submission needs to clearly define a glyphic subset associated by each IVS being registered.

The document pointed to by the submitted sequences.txt essentially shows just one representative glyph to specify a glyphic subset. We see no further definition, or even no hits sometimes, to explain coverage of a glyphic subset that corresponds to a representative glyph. As a result, the glyphic subsets are unclear.

To illustrate the problem, we take a close look at the three proposed glyphic subsets CID+3622, CID+14016, and CID+20229, just as an example. The following chart is an excerpt from the supplement document, 5078.Adobe-Japan1-6.pdf:

The three representative glyphs are shown to specify three glyphic subsets.

By a simple glance at these three representative glyphs, one can find there are combinations of two shapes in two components; two shapes for the radical part and two shapes for the phonetic part. Then, the following fourth glyph comes to mind:

返

The question is: Is this fourth glyph a member of any of the three glyphic subsets or not included in any of the three? Because the fourth glyph shares one of two features both with the representative glyphs for CID+3622 and with CID+20229, the fourth glyph may be included either in the glyphic subset CID+3622 or CID+20229, or it may be not included in

CID+3622 nor CID+20229. The current submission gives us no hints on the submitter's intention regarding this point. Moreover, we can think other possibilities. Since glyphic subsets specified by IVS need not be disjoint, the fourth glyph can be included both in CID+3622 and CID+20229. If so, why not all? I.e., the fourth glyph can be included in all the three glyphic subsets.

Of course we can break the features seen in the three representative glyphs and consider other combinations, e.g.,

or mix up other features allowed for unification to the Unicode character U+8FD4 but not seen in any of the representative glyphs to create more glyphs, e.g.,

There are hundreds of possible glyphs for U+8FD4. The author believes that the specification of a proposed glyphic subset needs to be clear enough to decide which of these glyphs are in the subset and which are not. However, the document designated as the web page for the *collection* (as in UTS#37) in this submission, i.e., Adobe Tech Note #5078, lacks the information.

This problem can be more serious for the cases on the glyphic subsets where only one IVS is proposed for a same base ideograph. Just showing one representative glyph gives us no useful information to determine the range of a glyphic subset.

The author imagines several solutions to this problem.

One possibility is to list as many glyphs as possible and specify which is in a subset and which is not. If the number of example glyphs for a subset is large enough and covers most of the possible features, it should illustrate specification of the subset.

Another possibility is, if the submitted *collection* is formed systematically, writing down the definition of the set of rules that groups possible glyphs into each glyphic subset as a part of the description of the collection may be a good way. The author assumes that the definition is something similar to S.1 of ISO/IEC 10646 or those seen on pages 417 to 420 in the Unicode Standard 5.0 book. Of course, this approach may not be possible if the collection is by non uniform way.

The last possibility the author can imagine is to explain definition of each glyphic subset in clear English sentences (and some supplementary figures where applicable.) The example for CID+3622 might be: "Any glyph whose *walk* radical part has just one dot on its top and straight stem (as opposed to a zig-zag stem). The shape of phonetic part is not significant and may be in any shape at all." (Note that this is the author's understanding of CID+3622, and may not match with the submitter's intention. The fact that one can misunderstand the subset is the problem the author is discussing here!)

Note that the example shown in B.2 of UTS#37 is a mixture of the first and the third method above, i.e., it gives some verbal description of the proposed glyphic subset as well as example glyphs that are and are not included in the subset.

PROBLEM 2: IVS that doesn't restrict allowed glyphs.

Some of the glyphic subsets in the proposed *collection* look no way to restrict allowed set of glyphs. For example, the proposal contains the following representative glyphs (taken from the supplement document, again):

The author can't imagine for what glyphic subsets CID+1200 or CID+8371 are intended, unless they correspond to unrestricted full sets for Unicode characters U+4E00 and U+4E28. Although the UTS#37 doesn't prohibit the registration of an IVS that corresponds to a set of all glyphs that is allowed for the base Unicode ideograph, registration of such an IVS doesn't make sense.

In the first example above, if CID+1200 is the set of all possible glyphs that is allowed for the Unicode character U+4E00, the sequence U+4E00 U+E0100 is exactly same as the sequence U+4E00. It will confuse users. The author believes such registration should be avoided.

Because of the problem 1 discussed above, the author may misunderstand the intention of the submitter, and the CID+1200 and CID+8371 are in fact some proper subsets of the glyphs allowed for U+4E00 and U+4E28 respectively. If it is the case, the author will withdraw this problem 2, assuming some clear definition for the glyphic subsets CID+1200 and CID+8371 are supplied in response to the problem 1.

As a related issue, the following sentences taken from the submission is questionable:

> Note that all Adobe-Japan1-6 kanji *(abridged)* are given IVS assignments, including those that have only one form assigned. *(abridged)* because kanji may be added in future Adobe-Japan1 Supplements that may be variants of such kanji.

The above sentences give an impression that, when a variant kanji is added in a future Adobe-Japan1 supplement, the set of allowed glyphs for an existing IVS for the existing glyphic subset changes in a way that the glyph for the newly added variant kanji is suppressed. The author believes that, once registered, the glyphic subset that corresponds to an IVS never changes, even if the Adobe-Japn1 is revised in a future. The author hopes that he simply misunderstood the quoted sentences, seeking for some reasonable clarification.

(END OF COMMENT)