

ISO/IEC JTC/1 SC/2 WG/2
Universal Multiple-Octet Coded Character Set (UCS)
Secretariat: ANSI

Title: Concerns on the VARIATION SELECTORS in ISO/IEC 10646-2, PDAM-1

Doc. Type: Experts Contribution

Source: The national body of Japan -tks

Project:

Status: to be reviewed at #42 WG2 meeting

Date: 2002-05-15

Distribution: JTC1 SC2 WG2

References:

Number of pages: 4

This paper expresses a concern on the VARIATION SELECTORS in ISO/IEC 10646-2 PDAM-1.

This paper is a concern of the NB of Japan. This paper can be treated as a new input for the FPDAM, or late additional comment from NB of Japan for PDAM-1 ballot.

This paper agrees with the sub-code assignment approach to express the variation(s) of the characters.

However, at the same time, this paper concerns an application of the VARIATION SELECTORS for CJK ideographs as well as non-CJK-ideographs, and also concerns differences of nature of the CJK ideographs and other scripts from view points of applying VARIATION SELECTOR. This paper expects more intensive discussion on the VARIATION SELECTORS before proceeding to the FPDAM and further, and just in case, the VARIATION SELECTORS may be delayed for Amendment-2 to keep other part of the AMD-1 on schedule.

This paper recommends to split the variation into several (at least two). And use the current for the "glyph variation" (will be expressed later) with minor additional text, and start the extensive discussion on the CJK application by opening ad-hoc group within WG2 for much clearer use of the selectors. Japan is willing to participate in the ad-hoc.

1. Collection and Block name and sub-setting.

As a data element, the nature of the VARIATION SELECTORS is explained reasonably well, but it is not enough as an element of the UCS standard. The concern is a definition of sub-set.

In PDAM-1, the block and collection number assignment is looked like as if those VARIATION SELECTORS are recognized as normal independent characters or script. This is creating a question for the specifying of subset.

This paper is using the word “mother character” as the character the VARIATION SELECTOR to be attached. “base character” might be better word for this, but the word is already used for other purpose.

1-1. Limited subset:

What if some of VARIATION SELECTORS are referred as a part of limited subset? Does this mean that any variations of mother character (referred as subset) with the variation to be referred as a part of subset? It is very difficult to understand this case. Real case might be that each subset elements are referred as a combination of mother character and the VARIATION SELECTOR. The combination should be referred as a single character in the subset, and not VARIATION SELECTOR alone can be referred as a part of subset.

This question does not request any test change of existing PDAM-1, but, may be some additional clarification text might be necessary as a part of subsets clause or as new annex.

1-2. Selected subset:

Selected subset is much more confusing than the limited subset if the VARIATION SELECTORS has its own block name and collection name.

1-2-1. What does no “VARIATION SELECTORS” referred selected as a part subset? This may be useful for CJK ideographs. It may mean no variation at all type sub-set. Besides, how about for non-CJK ideograph application? For Mongolian script, does no-VARIATION SELECTOR mean something? It means incomplete repertoire as Monglian. Does it mean anything useful? This sample hints that for some script, selecting one collection/block of mother character should automatically include all VARIATION SELECTORS for the mother characters (means no block/collection name necessary for the VARIATION SELECTOR). And for some script, selecting VARIATION SELECTOR as a block/collection might have some kind of meaning, but a sample case is not thought of as of now .

1-2-2. What does “VARIATION SELECTOR referred” selected subset?

If VARIATION SELECTORS collection is referred as an element of selected subset, what does that mean? Is that mean all (actually any) scripts referred as subset include all variation? Variation selection should be defined by script by script, not all or nothing. In most of non-CJK ideographs type scripts (math is somewhere in between), as described 1-2-1, specifying VARIATION SELECTORS means nothing anyway. And for CJK ideographs, the variation

requirement is not “all or nothing”, always “it depends”. Therefore, block/collection name assignment for the VARIATION SELECTOR is very questionable.

Above discussion indicates the fact that there are two (at least) different usage of the VARIATION SELECTOR. One is use as a “glyph” selector (i.e., Mongolian case) this fit well for WG2 model if WG2 write correct text. Another is (may be) a “shape definer” for the same character (CJK case), the nature of this case is not well nailed down yet.

However, it might be necessary to separate those two different cases clearly. If possible, there are two kinds of SELECTORS are necessary according to the difference of use.

And somewhere in between.

1-3. CJK-ideographs should use only limited subset: Is it practical?

One of the usages of the VARIATION SELECTOR is explained as an application for CJK-ideographs shape variation. However, a set of necessary variation (unification range in another word) is different from person to person, group to group and application to application. If someone should specify a set of necessary variation, the specification should be done mother character by mother character and VARIATION SELECTOR by VARIATION SELECTOR. This is limited subset approach, and if UCS introduces the VARIATION SELECTOR, users will be forced to define the limited subset because to let “unnecessary shape sneak in” is not acceptable practice. Easy use of the selected subset might cause serious problems. Besides, specifying a subset of CJK ideograph by limited subset is not a practical way to do because a set of too many characters should be specified one by one. There is a needs to introduce the new method of specifying a subset as a part of this standard (use of pre defined sub-repertoire as a extended selected subset might be one of the way), Anyway that means additional technical text and it should be recognized as a major technical enhancement of the UCS in formality of standard view.

2. Mother character issue with CJK ideographs.

Application of the VARIATION SELECTORS on CJK ideographs causes other questions. One of them is the question of “which mother character should be selected to define a necessary variation?”.

Remember that, in general, more than one code points are assigned for same CJK ideographs if the character shapes are different and recognized as “out side the unification rule”. Then, when a variation of shape to be defined to use the VARIATION SELECTOR, the question is which character should be the mother character for the variation.

For example, US style CJK is looked like a variation of simplified ideographs, but it is also a variation of the authentic shape of the character.

There are two ways of idea. One is to make tree structure such that the inter relation and history

of change can be easily tracked. Another way is that every thing are a variation of authentic shape such as Kang-XI shape. May be, someone propose random selection in free choice. There is very high possibility of multiple coding and user confusions if random or tree is selected. Besides, which is authentic discussion will happen if the single mother character method is selected.

Anyway, it is necessary to define on method for this question.

3. Relation with CJK unification:

Relation with CJK unification should be clearly defined.

Question-1: Does VARIATION SELECTOR define a specific shape within the unification rule?

Question-2: Does VARIATION SELECTOR define a character only outside unification range?

Question-3: Is that define both?

Question-4: Is that define only specific shape? Or actually independent character?

If the answer for question-1 is YES, then what will be the new unification range? This relates with the question-4.

If the answer for question-2 is yes, then does it include all shape variation within the unification rule centering the defined shape?

Many question and Answer are possible. What will be WG2 decision on this issue?

Conclusion:

Above discussion indicates two things. One is that there are two (or three) different kind of "variation" are going to be handled as one principle.

First of the two: One type of variation is "pure glyph" variation (Mongolian script). Second type is actually preferred shape of the same selector selection type application (CJK ideograph) and Math might be somewhere in between.

Second of the two: Issues are very different from type to type. The issue for "pure glyph" is simple and easy to resolve, but issues for CJK type is very complex and current PDAM is too un-mature. There is many philosophical decision (consensus) has to be made before WG2 proceed to further process.

It might be better to polish up current PDAM only for "pure glyph" variation type application. (may be, character name to be changed to GLYPH VARIATION SELECTOR).

And open the further discussion for CJK type application by creating new ad-hoc group within the WG2 . If necessary discussion on "open issue" with in the character/glyph model TR (CJK is out side of TR) should be opened. Japan likes to join the discussion there very positively.

---end---