L2/02-115

ISO/IEC JTC 1/SC 2/WG 2 N2423

DATE: 2002-03-25

ISO/IEC JTC 1/SC 2/WG 2

Universal Multiple-Octet Coded Character Set (UCS) - ISO/IEC 10646

Secretariat: ANSI

DOC TYPE:	National Body Contribution
TITLE:	**Proposal "Error correction on unified ideographs in UCS"**
SOURCE:	Japan: tks
PROJECT:	JTC 1.02.18 – ISO/IEC 10646
STATUS:	Input to WG2
ACTION ID:	For review by WG2
DUE DATE:	
DISTRIBUTION:	SC2/WG2 members and Liaison organizations IRG members and Liaison organizations
Reference	WG2 n2271(TCA), WG2 n2294(Japan), WG2 n2382(Japan) M41.11
NO. OF PAGES:	3

This paper proposes "Error correction on unified ideographs in UCS" for IRG and WG2

A principle on "error correction of CJK unified ideographs was proposed by Japan as WG2 n2382 at the #41 WG2 in Singapore. The proposal has been accepted in principle.

Note: the n2382 is a response for AI-39-9b from Japan.

However, a discussion after the acceptance, such as resolution M41.11 is not reflecting the idea expressed in Japanese proposal.

Japan, therefore, re-submit a proposal with a description of current practice that is in M41.11.

The practices that M41.11 describe is

a. Once a character is assigned a code position in the standard it cannot be reassigned in the interest of ensuring interoperability of standardized characters.
b. The arrangement of the characters in the standard is fixed; sorting and collation of the characters is outside the scope of the standard.
c. The character names chosen by WG 2 for the English version of the standard are unique, fixed and may be arbitrary; once a character name is assigned, it *cannot* be changed even if additional information is provided later. These name strings are used , for example to establish correspondences with characters in other standards.

Japan agrees with the reason behind the above principle, however, the principle is not applicable for some cases of error about the CJK unified ideographs. This is why Japan proposed a separated and unique principle for CJK unified ideographs.

There are two kids of errors in CJK unified ideographs.
Case-1: Ideographs that should be unified are assigned independent code point (TO BE UNIFIED error).
Case-2. Ideographs that should NOT be unified are unified and assigned only one code point (TO BE DISUNIFIED error). Example: request from TCA in n2271.

The M41.11 principle is applicable for case-1 error. The TO BE UNIFIED pair (this is just a multiple encoding problem) should not be unified (by deleting one code assignment, or by changing a code point).
Case-1 should follow the existing principle. If necessary, additional note, instead of changing a code point, should be a practice in this case.

However, case-2 is different situation. In the case of n2271, even though the cause of the mistakes are different, in both case, the character pair which should not be unified (because those are different each other) are unified simply by mistake.
There is no reason why UCS should assign one code point for two characters. Therefore, in this case, separation of the characters (DISUNIFICATION in some sense, and character name change in some sense also) into two code points should be done as soon as possible. This is what n2382 is requesting.

Once the proposal is accepted, there are some items to be agreed for practical application of the principle.

1. Which character should be relocated?

If the dis-unification is necessary, next question is that which one should be relocated and which one should use current code point?

Japan proposes "Glyph shape which is same as Unicode Book" should use the current code point. And "the new to Unicode Book glyph shape"" should get the new code point. There might be another discussion that the logically correct glyph (such as radical and stroke number of the code location to be) for the code point. However, the result of this method may, possibly, against the principle of M41.11. Thus, at least, the Unicode Book, fortunately it is in single column, is following the M41.11.

2. Error of unified ideographs in single column

If there is an error, the error might be discovered in Super CJK, the same correction principle should be applied. In this case, from the user of UCS, the result will be just an addition of new character and correction of the mapping table.

3. Real dis-unification request

If above practice is accepted, there is a possibility of "pure true dis-unification". This is almost like the new source code separation request. This kind of request shall not be accepted disregarding the reasoning behind.

4. Key difference between "TO BE DISUNIFIED" and "SHALL NOT BE DISUNIFIED" is:

   If. character pair is non-cognate (means different character), those pair are TO BE DISUNIFIED.

   If character pair is cognate (means the same but different shape), those pair are SHALL NOT BE UNIFIED.

5. Mis-application of the unification rule

   Dis-unification request with reason of mis-application (over-application usually) of unification rule should NOT be accepted due to the principle of M41.11.

---end----