ISO/IEC JTC/1 SC/2 WG/2
Information Technology, Multi-octet Coded Character Set

ISO/IEC JTC/1 SC/2 **WG/2 N** **IRG N453**

DATE: 1997-04-08

REPLACES:
none

DOC TYPE:
Personal contribution

TITLE:
Questions on the "Han structure method" described in WG2 N1490(IRG N436)

SOURCE: Takayuki K. Sato -Japan

PROJECT: ISO/IEC 10646 Han enhancements (IRG project)

STATUS: To be reviewed by NANJING IRG and Crete WG2, 1997

ACTION ID: ACT

DUE DATE:

DISTRIBUTION: ISO/IEC JTC/1 SC/2 WG/2 and SC/2 IRG

REFERENCE: WG2 N1490

MEDIUM: P

DISKETTE NO:
NO. OF PAGES:

Secretariat,

Questions on the "Han structure method" described in WG2 N1490(IRG N436)

To raise a question on the N1490 by 15th of April 1997 is one of the actions from the Singapore WG2 meeting. This is an individual response for the action request by Takayuki K. Sato. This contribution is sent to both the WG2 convenor and the IRG raporter for circulation.

The N1490 describes well about the technological details of the IRG proposal.
Even though there are some more technical questions, the N1490 answers enough for the evaluation of the proposal from its objectives. At this moment of time, therefore, the review (and questioning) should be done from view point of:

- Is this meeting the "objective" of the addition of this method?
- Does the technology harmonise with ISO/IEC 10646?
- Are there any draw backs (or changes) of ISO/IEC 10646 if this method is added?
- Are there any modification of the method needed for adapting into ISO/IEC 10646?
- Does the modified method still meeting the objective?
- Is this optimum and the best solution to meet the objective?
- Are there any better solutions to meet the objective than the proposed method?
- Are there any differences in reasoning to stick on the proposal?

1. Confirmation of the "Objective."
In Copenhagen WG2 meeting, it was said that the purpose of this proposal "structured characters" is "To support the CJK ideograph that is not on the ISO/IEC 10646." And also it was confirmed at same time that any other possible usage of the method is just a "side benefit" of the purpose, the side benefits should not justify the adaptation of the method by the side benefits themselves.

Note that if, for any reason, the project objective should be revised, it is necessary to have an approval on the new objective by upper level committee.

2. Further confirmation of the "objective."
There are two kinds of the CJK ideographs that are not on ISO/IEC 10646. One is "new unified ideographs per current unification rule" but not on the UCS yet, this means that the addition of this kind is still under unification principle. Another is "CJK ideographs that are to be (or has been) unified for a single code point of UCS (now or future)". This means that the CJK ideograph of this kind is “out side of control of existing unification rules."

WG2 N1490 (N1432-3 and N1433-3) indicates that "not existing CJK ideographs" to be added by this method are still under the unification rule.
This is in line with approved scope of IRG, if it is out side of unification rule, it is out side scope of IRG, and if this is a purpose, IRG needs to request new approval for upper level.

3. Ideographic Variation mark.

In N1490 (IRGN436) answers that this is used for ideograph or structured sequence only (answer 3). Besides that, WG2 defined it as a "general mark" at its Singapore meeting.
Any problem? Any specific reason for the answer?

4. Canonical Ideographic Structure Sequence (CISS).

4-1, What is the purpose of the CISS?

4-2, Does it define "unified CJK ideograph" uniquely?

If yes, Does the unified CJK ideograph including the unified CJK ideograph already in ISO/IEC 10646 (including ext-A)? It should not be so, because of the objective of this method.

4-3. Is equivalence table between CISS and the "unified CJK ideograph" provided?

If yes, how many of "unified CJK ideographs" to be selected for the equivalence table?

If No, why? How we can say which sequence is which?

4-4. What does it mean CISS defines SIS uniquely?

SIS is not a canonical (answer 7), so it can not be uniquely defined. Is this true?

4-5, Assume all of the rest of "unified CJK ideograph" are in plane-2,

Is CISS still needed?

What is the advantage of CISS beyond the "simple addition"?

4-6. Some may say, because the character to be represented by using only BMP.

Does it really have a benefit compare with use of UTF-16?

Additional complexity of the any Structure Sequence is too much if it is unique and complex for CJK ideograph only. Besides the UTF-16 is a common solution for any scripts (even though it add something beyond UCS2).

UTF-16 is far simpler than supporting the sequence.

4-7, What if someone create same string as CISS as just someone's SIS?

It would be highly possible. Should it be handled as CISS?

5. Radical supplements.

The answers in N1490(IRGN436) is based on "before Singapore WG2 meeting" status. In Singapore WG2, some (most of) radicals are uniquely defined as "radical" and not a part of the unified ideographs. This means that the situation has changed.

5-1. Are there any change needed for the answers in N1490? It must be.

5-2. Specially, The proposed supplement radicals are defined as "radical" also.

Are there any change of the principle of total proposal?

5-3. Are there any needs of additional radicals due to the change?

Note that the "radicals" should be a part of SIS now, there are two similar shapes in ISO/IEC 10646.

6. Component Supplements.

6-1. Are there any possibility of growing the number of the supplemental components?

The Structure Character is defined as to combine 2 or 3 components which is minimum number of components. Usually, to represent the large characters, the less number of components in the string requires the more number of components to be used.

7. Ideographic Structure Characters.

7-1. Level-3 issue.

Clause 15.3 says that level-3 means "every thing possible". Sounds good. And also it is requested to write a description of the Structure Characters in a style of ISO/IEC 6429 or clause 23 of ISO/IEC 10646-1. Once behaviour of the Structure Characters defined, if nothing mentioned, it should be applied to any ISO/IEC 10646 characters such as Structure Character followed by alphabet (LTR followed by A and E may mean character AE!!).
Is this OK? If NO, specific characters for valid operands to be listed, then it would be almost like different implementation level. Note that, implementation levels by themselves as of now defines a set of characters only, and the behaviour of the characters are limited (or attached) to the character itself only, it does not have any restriction on what to be used with the characters. (Therefore, the Ideographic Structure Characters are exceptional case of current UCS, This may mean additional implementation level)

7-2. Nesting
If nesting is allowed (yes, it is necessary, because number of combined symbols are limited within 2 or 3), how deeply can we allow the nesting level? Or, any concern about complexity of multi-level nesting?

7-2. Valid code.
Assume LTR, 4EBB, 4E8C is almost 4EC1, is RUE 4EBB, 4E8C also like 4EC1?
How about UTE 4EBB, 4E8C?

If those are almost same shape, are they same characters, or independent character?
If some of above are not valid, how to check them?

7-3. Dual encoding.
Is LTR+4EBB+4E8C representing 4EC1?
If yes, why we allow it? If no, Is that mean the string should be handled as independent CC-data-element from equivalent? then why?

7-4 Searching and comparing
Can SIS searched, compared and ordered as independently form equivalent?

8. Accuracy of representation.
Some of SIS may have a possibility of more than two characters such as following pairs:
5183 and 5184, 571F and 58EB, 65E5 and 66F0, 6737 and 6738, 6739 and 673A.
How can we differentiate? Some may say, use different SIS to represent the characters.
However, that idea invites the valid and invalid sequence issue. Is this OK?
Is it in-valid to use LMR for 6737 or 6738?

Comments:
a, Additional complexity and ambiguity can not pay off the merit of adding this method on to existing ISO/IEC 10646.
b, The method should be evaluated after inclusion of all necessary "unified CJK ideographs" into plane 2.
c, Still there is a question of "is this an efficient and best way to represent "unified ideograph which is not coded in ISO/IEC 10646?". This does not mean that this method can not represent the CJK, this means that it may not make sense after having over 85K of unified CJK ideograph.

9. Possible contribution
If similar to this method is applied to express real variation within the unified ideograph, there might be a possibility of the contribution.

<operator2><CJK code point><operand> or
<operator3><CJK code point><operand1><operand2>

the operands defines exactly which shape is needed. and still CJK code point represents the attribute of the characters. (If variation is not necessary, ignore operator and operands)

But, this is an out of scope of either 10646 project or IRG for now, and also, the operands have to be non-unified components (means another animal)..

------end----