L2/00-099

ISO/IEC JTC/1 SC/2 WG/2 **N2196**

2000-03-15

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

Title:	**CJK COMPATIBILITY IDEOGRAPH**
Doc. Type:	national body contribution
Source:	**Japan** -tks
Project:	02.18
Status:	To be discussed at the WG2 meeting in Beijing
Date:	2000-03-15
Distribution:	ISO/IEC JTC1 SC2 WG2
Reference:	WG2 N2095, N2142, N2143, N2159, WG2 AI-37-11-d, Updated CJK COMPATIBILITY IDEOGRAPH request from Japan,

This is a part of response for a WG2 action item AI-37-11-d from Copenhagen.

At Singapore IRG, there has been consensus among the participants about the necessity of the coding of CJK COMPATIBILITY IDEOGRAPH. (to guarantee a round trip conversion with other code standards)

Japan believes that the CJK COMPATIBILITY IDEOGRAPH is different ideograph in nature from CJK UNIFIED IDEOGRAPH. Thus, before jumping into the actions of the addition of the CJK CPMPATIBILITY IDEOGRAPH, Japan proposes to clarify the key characteristics of the CJK COMPATIBILITY IDEOGRAPH and document it. Japan believes that this is very important to avoid future confusions and struggle between different understandings.

The existing CJK COMPATIBILITY IDEOGRAPH is a good **bench mark** to test the characteristics of them. Taking F900-FA0D and FA10-FA2D as a sample case, the characteristics can be tested and defined. (FA0E and FA0F do have the same characteristics as well some of FA10-FA2D).

The CJK COMPATIBILITY IDEOGRAPH FA10-FA2D are included for ISO/IEC 10646-1:1993 for a compatibility purpose with IBM-Kanji set for Japan (and MS-Windows cp932). Even though there was an unusual discussion about Canadian minority, those are IBM-Kanji set compatibility ideographs. It has

been necessary for IBM to guarantee a round trip conversion between UCS and the IBM (and MS) code.

Bench Mark-1:

At the time of CJK UNIFIED IDEOGRAPH EXTENSION-A introduced, FA1F and FA23 were categorized as a part of the extension-A. Because there were requirement of CJK ideographs with the same shapes as the two CJK COMPATIBILITY IDEOGRAPHs, and test against the unification rule showed that there are no CJK UNIFIED IDEOGRAPHS to be unified with those two characters.

This fact leads us to the first rule for the CJK COMPATIBILITY IDEOGRAPH that is:

Rule-1: *A CJK ideographs "which can not be unified with any of existing CJK UNIFIED IDEOGRAPHs under the unification rule" should not be a CJK COMPATIBILITY IDEOGRAPH. It should be coded as CJK UNIFIED IDEOGRAPH (whatever reason for the proposal of inclusion to ISO/IEC 10646).*

WG2 decision at the time of development of 1993 version was: .. Those CJK ideographs should not be assigned for CJK COMPATIBILITY IDEOGRAPHS even though those are "compatibility with IBM-Kanji" purpose. In this view, at least, the later decision of the WG2 is correct. The decision is "whatever collection name is, once it is coded, there are no differences between CJK UNIFIED IDEOGRAPHS and CJK COMPATIBILITY IDEOGRAPH". (but bench mark-2 proofs that this decision is mistake).

Under this principle, at the CJK UNIFIED IDEOGRAPH EXTENSION-B development process, some more CJK COMPATIBILITY IDEOGRAPHS are re-categorized as CJK UNIFIED IDEOGRAPHS such as FA11, FA13, etc.…..(see final sheet of original CD 10646-2 ballot copy for more of those ideographs).

Bench mark-2:

The bench mark-1 is an exceptional case, comparison between FA25 and 9038 may give another view. FA25 does have exact same shape as G-column of 9038. FA25 should be unified with 9038 in principle. Because the IBM-Kanji needs both J-column shape of 9038 and FA25 with independent to each other code point, the FA25 has been included into the UCS to guarantee a round trip conversion between UCS and IBM-Kanji-set (MS Windows cp932 as well). It is not possible to handle FA25 as if it is new independent CJK UNIFIED IDEOGRAPH as WG2 decided. Because G-column shape of the 9038. to make FA25 as CJK UNIFIED IDEOGRAPH, it is necessary to remove the G-column of 9038 to FA25, or remove T, J and K column to somewhere else and put FA25 into J-column of 9038. This is not practical approach, this means an inpompatible change of existing code table. And this is typical case for the CJK COMPATIBILITY IDEOGRAPH.

Definition-1: CJK COMPATIBILITY IDEOGRAPH *CJK ideograph that has to be unified with existing CJK UNIFIED IDEOGRAPH in principle. However, different code position is assigned for the CJK COMPATIBILITY IDEOGRAPH to guarantee "round trip conversion" between UCS and existing national standard(s) or industrial practices.*

Definition-2: CJK UNIFIED IDEOGRAPH *CJK ideograph that is independent ideograph from any other CJK UNIFIED IDEOGRAPH per the unification rule of the UCS. The unification is done by the unification rule defined in the ISO/IEC 10646.*

Therefore, the new definition is necessary. The definition may require the revision of previous WG2 decision. The new decision should be " CJK COMPATIBILITY IDEOGRAPH and CJK UNIFIED IDEOGRAPH are different in nature each other. CJK COMPATIBILITY IDEOGRAPH should be used only for the purpose for round trip conversion guarantee".

Bench mark-3:

CJK COMPATIBILITY IDEOGRAPH for KS C 5601(F900-FA0D) has different nature. For example, F900 is the same as 8C48 in its shape. The reason for compatibility is different from the case FA25 described above. F900-FA0D are necessary not because of the shape, but because of the "pronunciation". This fact indicates that the reason for compatibility is not only for the shape but sometimes it would be different (such as pronunciation) attributes of the character.

Rule-2: *CJK COMPATIBILITY IDEOGRAPH should be proposed and coded with the information of its corresponding CJK UNIFIED IDEOGRAPH and the source standard (or industrial practice) that the round trip conversion should be guaranteed. If possible, reason why the source standard is separating the ideograph to be added.*

If special attention is paid for the bench-mark 3, then one more rule to be provided.

Rule-3: *CJK COMPATIBILITY IDEOGRAPH should not be unified with other CJK COMPATIBILITY IDEOGRAPH by their shapes. They should be unified by the reason why it is coded separately in the source standard. Since the reason for separation in source standard(s) may change from time to time, in principle, CJK COMPATIBILITY IDEOGRAPH should not be unified unless having good reasoning to unify.*

Additional rule:

In addition to above three rules, one more guidance is needed.

Since the needs for the CJK UNIFIED IDEOGRAPH is for compatibility with other computer implementation. Not for coding of anything else.

Rule-4: *The CJK COMPATIBILITY IDEOGRAPH should only be available for round trip conversion purpose with coded character set which are implemented (or be able to implement) for computer system. No other reason for compatibility shoule be allowed (such as book and dictionary).*

Bench Mark-4 (Round trip conversion):

KS C 5601 compatibility ideograph (F900-FA0D) guarantees the round trip conversion between the UCS and KS C 5601. However, (since there are no such an ideograph in other countries), the KS C 5601 compatibility ideograph can not be mapped to other than KS C 5601. Therefore, the round trip conversion is not possible from KS C 5601 to UCS then (for example) to JIS (and back to KS C 5601 via UCS). Let's name this conversion as "Around the world conversion".

Rule-5: *CJK COMPATIBILITY IDEOGRAPH guarantees a round trip conversion between one specified coded character set and UCS. It does not guarantees the around the world conversion. The around the world conversion may be done by user's own risk.*

Bench Mark-5 (Source Code Separation) :

There is another type of compatibility ideographs are definded in annex S of UCS. It is called as "source code separation". This is very similar with CJK COMPATIBILITY IDEOGRAPH, but it is different animal. For example, Annex S of ISO/IEC 10646 says that 4E1F and 4E22 are the source code separation. Per unification theory, they should be unified, but since those are separated before UCS is designed, for round trip conversion purpose, both ideographs are coded as CJK UNIFIED IDEOGRAPH.

What's difference? 4E1F originally has open G-column and 4E22 has J and K columns opened. Taiwan needs both. But others need one of those. If G-column of 4E1F is filed by the shape at G-column of 4E22, (and G-column of 4E22 is open) the T-column of 4E22 can be a CJK COMPATIBILITY IDEOGRAPH. But since this is not a case, it is easy to handle those relations as independent each other. (means CJK UNIFIED IDEOGRAPH) Unless, there would be many confusions by the user.

Rule-6: *The source code separated ideograph is a CJK UNIFIED IDEOGRAPH and it does have it's own shape variation range which does not overlap with any other CJK UNIFIED IDEOGRAPH. (even though generic unification rule works differently)*

Rule-7: *No more source code separation with in UCS. If compatibility is needed, then use CJK COMPATIBILITY IDEOGRAPH.*

Caution: There will be two kinds of thinkings within the WG2 in near future. Pick one as the WG2 decision, unless, there will be another battle.

Alternative-1: Over unification of ideograph may cause this problem again and again. Therefore, if there is a possibility of future separation, unification of such ideograph should not be done

Alternative-2: Since we have safety valve, namely CJK COMPATIBILITY IDEOGRAPH, now UCS does not necessary to consider an useful variation of the same ideograph, it may be supported by CJK COMPATIBILITY IDEOGRAPH. Thus much more tight unification might be possible.

Japan proposes to have this rules accepted by the SC2 and included some where at ISO/IEC 10646.

----end-----