Universal Multiple-Octet Coded Character Set UCS

ISO/IEC JTC1/SC2/WG2 N4663
ISO/IEC JTC1/SC2/WG2/IRG N2067
Date: 2015-04-01

Source:	Japan
Title:	Further discussion on the disposition of PDAM2.2 ballot comments
Meeting:	
Status :	
Actions required	SC2/WG2 and IRG for review
Distribution:	Electronic
Medium :	
Page:	
Appendix:	
References:	WG2 N4656

Preface

Comments of PDAM2.2 ballot are disposed by project editor as WG2 N4656 and ballot of revised draft for PDAM has been started. However Japan NB is not satisfied with this revision because CJK F is removed without making consensus, for example. Most decisions regarding CJK are made following UK comments but most of them are already discussed in IRG in the past or ignoring rule defined in ISO/IEC 10646.

Japan sees that IRG's submission was mature enough and it is not appropriate to remove because of very small number of faults. Japan also think that IRG should examine dispositions in WG2N4656 carefully and discuss what IRG should do.

This paper describes Japan's view on the disposition regarding CJK so that IRG can review easily. Japan also expects that CJK F will be moved back to amendment 2 draft if the issues became clear.

Summary

Japan expects only three characters should be discussed at IRG meeting. Japan also requests IRG to confirm that other issues are not necessary to discuss if any other reasonable rationale is submitted.

T16. Clause 31 (32) – CJK Unified Ideographs Extension F

With respect to the request to remove USAT source references from 50 characters in CJK Unified Ideographs Extension F (See IRG N2041), and consequently remove from CJK Unified Ideographs Extension F those 49 characters that only had a USAT source reference:

A. We agree to the removal of 48 characters with USAT source references;

B. We request that 2D30C (USAT-00856) not be removed as it is a character attested in use in Bernard Karlgren's highly influential *Grammata Serica* (1940) and *Grammata Serica Recensa* (1957). This character and other unencoded characters in Karlgren's *Grammata Serica Recensa* were proposed as part of the UTC Extension F submission (see IRG N1888), but the entire submission was sumarily rejected by IRG on procedural grounds, which we consider to have been extremely unfortunate. As this character has been proposed by the UTC we request that it be kept in CJK Unified Ideographs Extension F with the source reference changed to UTC-01155.

C. We note that four of the fifty USAT characters that are requested to be removed are identical to existing encoded characters:
U+2D6AC (USAT-01869) = 2266C 𦚬
U+2D9AE (USAT-02066) = 2AC87 𪮇
U+2D9DE (USAT-03431) = 234C3 𣓃
U+2DBD2 (USAT-01739) = 6FD3 濂
These four characters have simple IDS sequences, and we would have expected that automated IDS checking of CJK-F would have identified these characters as duplicates. That these duplicates were not detected prior to submission to WG2 indicates a failure in the IRG Quality Assurance process, which we consider to be very worrying.
Proposed change by UK:
Remove the following 48 characters from CJK Unified Ideographs Extension F:
2CED4 (USAT-04335)
2CEE7 (USAT-04345)
2CF42 (USAT-02160)
2CF6D (USAT-60012)
2CF75 (USAT-03966)
2D06B (USAT-00332)
2D16F (USAT-05701)
2D18A (USAT-60046)
2D1BF (USAT-05796)
2D20A (USAT-01366)
2D29A (USAT-01778)
2D402 (USAT-05302)
2D442 (USAT-03388)
2D4BD (USAT-04810)
2D4E0 (USAT-90141)
2D531 (USAT-02486)
2D588 (USAT-03840)
2D59C (USAT-03564)
2D5D0 (USAT-03076)
2D5D4 (USAT-02802)
2D623 (USAT-60095)
2D6AC (USAT-01869)
2D732 (USAT-60123)
2D744 (USAT-02396)
2D74E (USAT-03777)

2D78E (USAT-03807)
2D810 (USAT-02889)
2D82D (USAT-00313)
2D82E (USAT-01556)
2D83F (USAT-00055)
2D86C (USAT-04803)
2D912 (USAT-03855)
2D987 (USAT-00998)
2D98C (USAT-02114)
2D9AE (USAT-02066)
2D9DE (USAT-03431)
2D9E8 (USAT-04809)
2DA24 (USAT-03899)
2DA3E (USAT-01373)
2DAD2 (USAT-02590)
2DB04 (USAT-60177)
2DB41 (USAT-03269)
2DBD2 (USAT-01739)
2DC09 (USAT-00268)
2DC6D (USAT-00966)
2DCEB (USAT-04194)
2DCF9 (USAT-03483)
2DDBB (USAT-00307

Accepted in principle

See also comments T1 from China.

See also comments T18 and G19 from UK which are resulting in the removal of CJK Ext F from this amendment.

T16 of UK comments suggests dropping 48 USAT single source characters with changing one source for keep. Changing USAT-00856 to UTC-01155 because UTC once proposed a character with the same shape to CJK F. However UTC submission to CJKF are rejected at IRG#39 because of lack of information and missing the target date. Thus, that source is never reviewed in IRG and it is unfair to keep that source in CJK F (i.e. delete 49 characters from CJK F as IRG concluded.)

T17. Clause 31 (32) – CJK Unified Ideographs Extension F

IRG N2042 identifies 11 further characters in CJK Unified Ideographs Extension F that should be removed as duplicates or unifiable with existing characters:

2D127 (USAT-01722) : unifiable with 2057D 汞
2D3AD (GCY-0697.00) : unifiable with 2144F 壝
2D5A5 (USAT-03456) : unifiable with 536E 卮
2D666 (USAT-04922) : unifiable with 224BF 犯
2D6B9 (USAT-01338) : unifiable with 22758 惺
2D754 (KC-01326) : unifiable with 2F8B1 □
2D834 (USAT-04653) : identical to 2ABBE 攘
2D9A4 (USAT-01096) : unifiable with 6752 朷
2DAE2 (KC-01963) : identical to 27BF8 彀
2DD0B (JMJ-059937) : unifiable with 20924 眞
2DD82 (JMJ-058841) : unifiable with 488B 辺

That these unifiable and duplicate characters were only identified after CJK-F was submitted to WG2 further indicates a failure in the IRG Quality Assurance process, and suggests that CJK-F may have been prematurely submitted to WG2.

Proposed change by UK:
Remove the following 11 characters from CJK Unified Ideographs Extension F:
2D127 (USAT-01722)
2D3AD (GCY-0697.00)
2D5A5 (USAT-03456)
2D666 (USAT-04922)
2D6B9 (USAT-01338)
2D754 (KC-01326)
2D834 (USAT-04653)
2D9A4 (USAT-01096)
2DAE2 (KC-01963)
2DD0B (JMJ-059937)
2DD82 (JMJ-058841).
Accepted in principle
See also comments T1 from China.
See also comments T18 and G19 from UK which are resulting in the removal of CJK Ext F from this amendment.

T17 of UK comments is based on IRG#43 Editorial Group Report (IRG N2042). All comments here was found and examined by IRG themselves and concluded. This is a success of IRG's Quality Assurance Process that IRG is continue working after submit to WG2.

T18 Clause 31 (32) – CJK Unified Ideographs Extension F
We have carried out a partial review of CJK-F, focusing primarily on the USAT source characters, and note the following issues.
2CEF3 (JMJ-056849) has a round dot above which is not a stroke used in the Han script. Is this really a distinct character? Or is it simply 20000 𠂉 with an editorial dot, in which case it can be represented as 20000 𠂉 plus 0307 combining dot above.
2D13F (USAT-00061) is unifiable with 20991 𡈼.
2D260 (JMJ-059428) should be 30.10 strokes not 30.11.
2D459 (USAT-60078) 𠘨啵女 is actually 5619 嘍. This is evident from the 口 element which is large and not aligned with 波 (cf. the size and position of口 in 2D446 USAT-00947).
2DB74 (USAT-05567) may be unifiable with 6EDB 𤶼.
2DD0F (JMJ-058197) should be radical 86.13 not 112.12.
Proposed change by UK:
Remove 2CEF3 (JMJ-056849) for further study.
Remove 2D13F (USAT-00061).
Reorder 2D260 (JMJ-059428) as appropriate.
Remove 2D459 (USAT-60078).
Remove 2DB74 (USAT-05567) for further study.
Reorder 2DD0F (JMJ-058197) as appropriate.
Accepted in principle
Based on this, IRG needs to review these comments and issue a new repertoire. As a result, CJK Ext F is removed from this amendment and postponed to the Committee Draft of the 5th edition.

T18 of UK comments seemed by original work of UK because they are not in IRG document. Japan feels so happy because other members outside IRG are working to brush up IRG's work.

Japan sees these three issues are clear because some of them are already discussed at

IRG or clear from the rule.

2CEF3 (JMJ-056849) has a round dot above which is not a stroke used in the Han script. Is this really a distinct character? Or is it simply 20000 𠂉 with an editorial dot, in which case it can be represented as 20000 𠂉 plus 0307 combining dot above.
→ *This is a Han Character although there Japan NB already submitted copy of DaiKanwa Dictionary as evidence. Is there any rationale to deny the description of authentic dictionary?*

2D260 (JMJ-059428) should be 30.10 strokes not 30.11.
→ *This issue was already discussed at IRG#40 and #41 as recorded in IRGN2044 and concluded to "30.10".*

2D459 (USAT-60078) 𠗈破女 is actually 5619 嚏. This is evident from the 口 element which is large and not aligned with 波 (cf. the size and position of口 in 2D446 USAT-00947).
→ *The proposed character shape has different relative position of component therefore they should be separated because of S.1.4.2. What is the reason to request changing radical?*

On the other hand, rest three are not ever discussed at IRG. They might be worth discussed at IRG.

2D13F (USAT-00061) is unifiable with 20991 𣥂.
2DB74 (USAT-05567) may be unifiable with 6EDB 𤾺.
2DD0F (JMJ-058197) should be radical 86.13 not 112.12.

It is truly difficult to expect perfect regarding repertoire of CJK Unified Ideographs because there are so huge characters that is 80k characters with over 200k sources. So it is important to continue working even after the draft is released from IRG.

G19 Clause 31 (32) – CJK Unified Ideographs Extension F
We note that a very large number of the JMJ source characters do not appear to be suitable candidates for encoding.
A very large proportion of the JMJ source characters appear to be idiosyncratic, calligraphic or semi-cursive variants of the same character (e.g. 2D004 through 2D007 and 2D009). Encoding these variants seems to us to go against the spirit of Annex S, and we believe that such variants would be best dealt with as IVS sequences.
Some of the JMJ source characters are weird squiggles, which look nothing like CJK characters (e.g. 2CEF8 through 2CEFD and 2CEFF), and we wonder whether they really are distinct CJK characters.
Proposed change by UK:
Consider removing all JMJ source characters from CJK-F for further consideration by IRG, in order to determine which of these characters are appropriate for encoding according to the Character-Glyph model, and which would be better dealt with as IVS sequences.
Accepted in principle
Based on this, IRG needs to review these comments and issue a new repertoire. As a result, CJK Ext F is removed from this amendment and postponed to the Committee Draft of the 5th edition.

IRG already discussed many times about similar issue that G19 of UK comments raised. Some conclusion are recorded in IRG documents clearly. (e.g. IRG N2045, IRGM39.5, etc.) It is true that there are many variants for all sources (not only for JMJ) in draft

CJK F and already encoded CJK Ideographs. All of they are carefully reviewed under unification rule (i.e. Annex S) to determine if they are separately encoded and finally submitted to WG2 because IRG concluded that they are appropriate to be coded separately. That is the spirit of Annex S.
Addition to this, UK teases u+2cefc(USAT-04910) and u+2ceff(GZ-1382301) as "weird squiggles", but they are not JMJ source.

Japan feels really inconvenient with this situation that such strange comments without reasonable rationale is adopted at ballot stage.

Postscript

It was usual manner in the past that disposition of ballot comments are discussed at WG2 meeting in F2F manner to reach consensus, however, it is difficult because frequency of the meeting was changed to 12 month interval. This may be the first case that project editor makes disposition without expert's consensus. Japan expects WG2 to discuss about the procedure of how the work progress under JTC1 directives, and IRG should be more careful how the work is discussed after submitting to WG2.

(End of Document)