JTC1/SC2/WG2/IRG N922R
Date: 2002.05.08

ISO-IEC JTC1/SC2/WG2/IRG
Ideographic Rapporteur Group

Report from the ExtC Editorial Group

Source : IRG
Meeting: Macau Cultural Centre, Macao SAR
Title : Report from the Ext C1 Editorial Group

1. Summary:

The editorial group has carried out two days of checking on N907 (Ext C1) and concluded:

a. We cannot finish the checking with very high degree of accuracy within a short period of time. Rushing will compromise quality and we will suffer later.
b. We have experimented with different checking methods and decided on two best approaches (will be presented later)
c. We have summarized some common errors.
d. We have concluded a working plan and a schedule (to be presented)

2. Sources of errors:

The error types reported in the following does not include any errors due to submission source errors from members. The errors reported are mainly duplications or characters have already been included in Super CJK. After some exercises, we have found the common sources of errors are:

a. Wrong choice of radical for reference 帎, use 巾 not 日.
b. Ambiguity in stroke count

免 免	鬼 鬼	差 差	者 者	韋 韋	嬁嬥
7 or 8	9 or 10	9 or 10	8 or 9	9 or 10	

c. Problem with first stroke code

 1 or 3

d. Some characters should be unified
e.g. Moon 月and meat肉; sun 日and the fatter version曰, grass ⺾ ⺿ etc.


The view of the editorial groups is we should follow Annex S, the unification rules

3. Checking Methods:

We have carried out checking using three different methods.

a. Manual check: members reviewed each entry by checking against Super CJK and dictionary and counting strokes manually. They also rely on memory to identify possible duplications.
b. With the help of an Input method: All characters will be typed in one by one using an approximate or very intelligent input method. If a similar shape character is found, we shall investigate the source of error. In some cases, the characters are different but it some cases these are duplications due to the problem mentioned in 2. Even if no entry is found, it is unsafe to assume there is no duplication.
c. With the help of a CJK seeker Database: We can use an intelligent composite search method to show all CJK characters with the radical and strokes specified in the Ext C1 entry. We can visual inspections on possible errors. To double check, we also need to go through alternate radicals and different stroke counts.

Our conclusion is we need to carry out both methods b & c independently together with a to ensure quality.

4. Our proposal

To carry out quality checking, we propose to adopt the following checking methodology.

a. The document N907 will be divided into 4 parts
b. Members will be divided into 8 teams:
   i. China1
   ii. Japan
   iii. TCA
   iv. Two Koreas
   v. Hong Kong + Singapore
   vi. Vietnam + Macao
   vii. Unicode + USA
   viii. China2

Two teams will carry out checks on one of the 4 parts (Project Editor will decide) in parallel. One of the two teams will use Input Method approach and the other will use DB method.

5. Report format

The reports will be in Excel Spread Sheet format for easy editing in the future. For easy consolidation, the excel file should use ASCII code only (no Hanzi Characters). The format of the report is as followed:

Flag	Ext C1 ref. No. 1	Ext C1 Ref. No 2	CJK Ref (U + **)	KX Ref.	Radical (KX)	1st Stroke No.	Stroke Count	Note

Flag: M: safe for the editor to merge or to delete
D: more review and discussion required

Ext C1 ref. No 1: The Ext C1 reference number for the character in question.

Ext C1 ref. No 2: Optional. The Ext. C1 reference number for the character which may merge with Ext. C1 ref. No. 1.

CJK Ref. No: Optional. The CJK's Unicode reference number for the character which may merge with Ext C1 ref. No. 1.

KX ref.: The correct KX reference number and it is optional. It is needed only if there is an error.

KX Radical No: The correct KX radical number and it is optional. It is needed only if there is an error.

1st Stroke No.: The correct 1st stroke reference number and it is optional. It is needed only if the DB is wrong.

1	2	3	4	5
一				乙

Stroke Count: The correct stroke count and it is optional. It is needed only if the DB stroke count is wrong.

Examples:

a. Possible duplicated entries in Ext. C1 (N907)

Flag	Ext C1 ref. No. 1	Ext C1 Ref. No 2	CJK Ref (U + **)	KX Ref.	Radical (KX)	1st Stroke No.	Stroke Count	Note
M	1234	ABCD						

b. More than two characters look the same (more than 1 entry is needed)

Flag	Ext C1 ref. No. 1	Ext C1 Ref. No 2	CJK Ref (U + **)	KX Ref.	Radical (KX)	1st Stroke No.	Stroke Count	Note
D	1234	5678						
D	1234	ABCD						
D	1234	WXYZ						

These entries imply 1234, 5678, ABCD and WXYZ are candidates for merging. **Do not make a separate entry for ABCD with WXYZ because the multi row entry is sufficient.**

c. Error in Ext. C1 DB (N907):

Flag	Ext C1 ref. No. 1	Ext C1 Ref. No 2	CJK Ref (U + **)	KX Ref.	Radical (KX)	1st Stroke No.	Stroke Count	Note
M	1234	-----	-----	Correct data as required				

d. Error in Super CJK:

Flag	Ext C1 ref. No. 1	Ext C1 Ref. No 2	CJK Ref (U + **)	KX Ref.	Radical (KX)	1st Stroke No.	Stroke Count	Note
M	-----	-----	U+123	Correct info as required				

If two characters can be merged, we make an entry like example a. If one of the two has incorrect information, we also need to make another entry like example d.

If we find two characters are possible variants, we should make an entry and marked it as D for discussion.

For characters with D tags, if the characters were from the same member, the member should make the initial decision if these characters can be unified. If these characters are submitted by more than one country, we should resolve it in IRG#20.

If two or more characters can be merged (as reported with M tags), editor will remove the redundant characters from the later versions but the source information of the deleted characters should be added to the source information section of the retained character.

6. Schedule

By	Milestone	Actions
		Members will carry out checking* and produce reports, based on Ext C1 V1 (N907). (Members are encouraged to provide or update their searching tools by June 1.)
2002.07.05	Error Reports to Project Editor	
		Project Editor will consolidate and produce review report
2002.07.26	New N907 (Ext C1 V2) and a database changes report and a discussion report	
		Members will carry out checking and produce reports, based on Ext C1 V2. Member will carry out checkings using a and b, or a and c, on a different part of the Ext C1. (Members are encouraged to update their searching tools by Aug. 9.)
2002.09.27	Another round of error reports to Project Editor	
		Project editor will consolidate and produce review report
2002.10.25	New N907 (Ext C1 V3) and a database changes report and a discussion report	
		Circulate reports for review and discussions. (Members are encouraged to update their searching tools by Nov. 8.)
2002.11.17	IRG#20	Editorial meeting.

* The Chief Project Editor will assign to each editor team which part of the Ext C1 they will be responsible by email.