JTC1/SC2/WG2/IRG N1018
Date: 2003.11.19

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

IRG #21 Summary report of IICore Ad Hoc Group

Source : IRG
Meeting: Guilin, China
Title : IRG #21 Summary report of International Ideographs Core Ad Hoc Group

The IRG#21 IICORE Ad Hoc Group meeting was attended by China, HKSAR, Japan, Macao SAR, R.O.Korea, TCA, USA, and Unicode.

The following is a summary of the discussions. Please note that the numbers are indicatives. (Changes to these numbers are expected with minor removals and additions due to version change or editorial/compilation mistakes).

1. Review of submissions

The group has reviewed the submissions from G, T, H, J, DPRK, ROK, M with total number of submissions of 10,663, respectively.

Members agreed to further investigate the characters in the set and further clarify and justify the submissions. The investigation has produced some useful statistics listed below.

A set of characters, named the Level 1 Common Set(L1CS), is produced based on the union of all the characters from Level 1 of GB2312, Level 1 of Big5/CNS11643, Level 1 of JISX 0208, Level 1 of KS X 1001 with a total of 7,772 characters.

Members have then produced two separate sets based on individual member submissions for the characters in L1CS and characters not in L1CS(NL1CS). The summary of these data are given as follows:

Member body	In L1CS	NL1CS
G:	4,888	2,054
H:	4,897	327
T:	5,762	799
J:	4,334	259
K(S):	4,639	110
K(N):	4,634	19
M:	4,788	172

A common set for characters in all the NL1CSs is further analyzed as follows:

Appearance in No. of NL1CS Set	Total No. of characters
4:	16
3:	78
2:	403
1:	2,509

Thus there are around 497 number of characters which appear in more than one member body submission, which we call the Edited Common Set(ECS)

For those characters that are unique submissions(NL1CS_MU) or single source submissions, further information are provided(the classification information is based on IRG N947):

Member	NL1CS_MU	Members Level2	Other information
G:	1,702	1,473	229
T:	413	382	31(B: 17, E:1, G:12, T3:1)
J:	175	167	8(D)
K(S):	77	68	4:K3(D) 5:K4(D)
H:	159(98+61)	0	B:3 D:3 E:15 F138
M:	29		
KN:	15		

2. Agreed principles and further actions

2.1 Naming of character sets in IICORE:

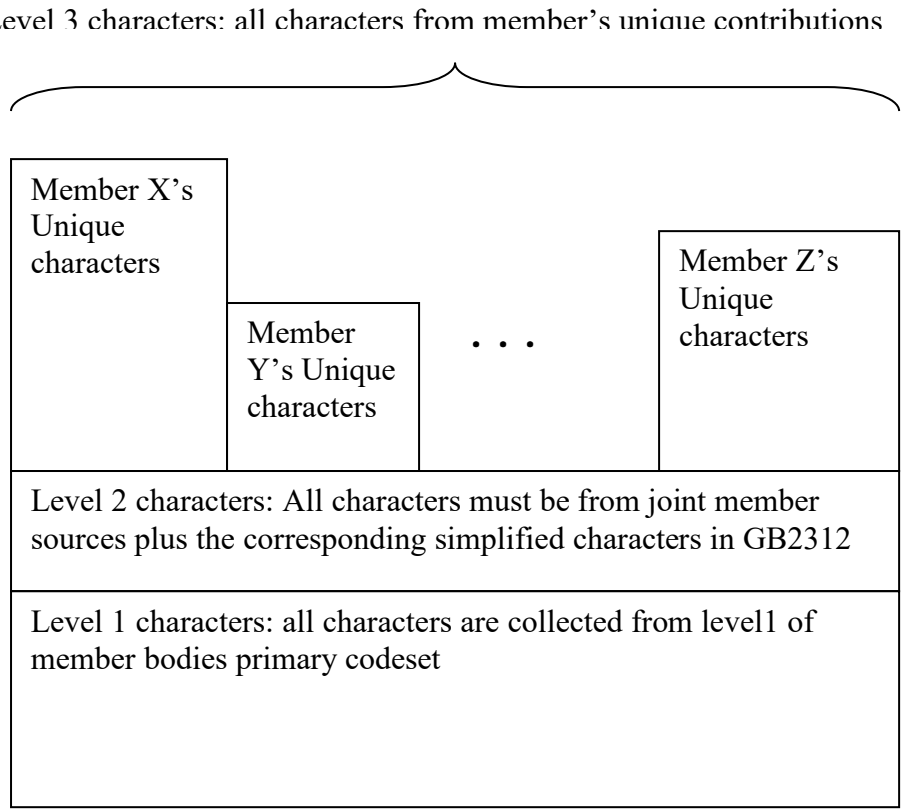
Category A characters: Characters in L1CS are referred to as Category A characters.

Category B characters: Characters in ECS are Category B characters. For those characters already defined in Category A and Category B, their corresponding simplified characters (defined in the level of GB2312 or the *General List of Simplified Hanzi* and the *General Purpose Hanzi List for Modern Chinese Language* [1]) are also considered Category B characters.

Category C characters: Single source character submissions (NL1CS_MU) are called Category C characters. Only characters in Unified CJK ideograph blocks can be considered in Category C inclusion. The characters for inclusion should be qualified within the scope of IICore and the total number of characters in each NL1CS_MU should be of reasonable size.

The following diagram gives a graphic indication of the relationships among the different categories of characters in IICore

[1] 《简化字总表》and《现代汉语通用字表》



2.2.2. Recommendation to WG2

The group has agreed unanimously that IICore should be accepted as the normative part of ISO 10646 to be included in Annex A.

2.3 Additional explanations

1. IICore as a whole set is considered a single indivisible named subset. The classifications into different categories are used for internal editing purpose only. Category C characters are not considered optional characters. They are specified to recognize the difference in day-to-day need in different member body's represented countries or regions.

2. IICore will be produced as a named subset under ISO 10646 for convenience of implementations for applications with limited memory, input/output capabilities, and/or applications where the use of complete ISO 10646 ideograph repertoire would be cumbersome.

3. All characters in IICore must be coded ideographs in the Unified CJK ideograph blocks. No character in the Ideograph compatibility Zone, the radical blocks, or other symbols should qualify for IICore inclusion.

2.4 Working schedule

2.4.1 Member Body(2:00 pm Wed. Nov.19, 2003): confirm the data or submit corrected data to technical editor(TE), Ms. Wang Xiao Ming

2.4.2 Technical Editor, Ms. Wang Xiao Ming, (Thursday Nov. 20, 2003): Produce and distribute Draft1(D1) of IICore where Category A and Category B should be merged in one file, but Category C data should be in separate files for different member bodies.

2.4.3 Member Body(Friday Jan. 16, 2004): Review of D1 including checking Member's Level 3 characters

2.4.4 IRG Raperteur(Feb. 2, 2004): Report to WG2 the work of IICore with D1 and explanatory notes (IRG N1018)

2.4.5 Mr. John Jenkins(at the next UTC meeting) : Inform UTC on IICore work with supporting documents(D1 and IRG N 1018) and also get their feed back

2.4.6 Technical Editor(Monday March 15, 2004): Produce and distribute D2

2.4.7 Member body(Friday April 30, 2004): Review of D2 and feed back to TE

2.4.8 Technical Editor(one week before IRG#22 meeting): Produce D3 for preview before IRG #22 meeting

2.4.9 Technical Editor and member body(during or after IRGN22 meeting): Produce D4

2.4.10 IRG Reperteur(before WG2#45 meeting): Submit D4 to WG2 #45 meeting

2.5 Data format

The submission should be a text file with an entry for one character per line. Each line consists of 8 or 9 bytes:

Bytes 1 through 5: the ISO/IEC 10646 code point.(0xxxx for BMP; 2xxxx for Plane 2)

Byte 6: the source ID (G, T, J, K, V, D, H, M, S, U)

Byte 7: indicates the source subID, such as "0" for G-0, "7" for T-7, and so on. Sources which are not subdivided should use "1".

Byte 8: one of the following, depending on where the character comes from:

- A: for level 1 of the source encoded character set
- B: for education
- C for level 2 of the source encoded character set
- D for personal names
- E for place names
- F for colloquial characters
- G for anything else
- P .. Z: member's own classification. Explanatory notes should be provided for other members.

Byte 9: an optional byte. If members would like to provide additional information for level 1 submission, they can indicate the character nature using categories B to Z as listed in Byte 8.