

Date: 1994-09-12
Supersedes: IRG N128

ISO/IEC JTC 1/SC2/WG2/IRG

Ideographic Rapporteur Group

Secretariat: IPSJ/ITSCJ (Japan)

Source / Contribution Identifier: **IRG**

Meeting: @ / for **Beijing '94**

Issue Number: Project / Task Identifier: **Annex-1, Annex-2**

**TITLE: Final draft of Annex: Rules of Unification and Arrangement
for CJK Ideographs**

Status: **Approved draft**

Keywords: **Rules of unification, identification for ideograph**

Short Description:

This document is the final draft for PDAM of annex to ISO/IEC 10646-1

Proposed Conclusion / Requested Action:

To be forwarded to ISO/IEC JTC 1/SC2/WG2

Argument / Text of Contribution:

See attachment.

Contact:

Mr. Akio Kido, Rapporteur of ISO/IEC JTC 1/SC2/WG2/IRG

DBCS Technical Coordination Office, IBM Japan, Ltd.

1623-14, Shimotsuruma, Yamato-shi, Kanagawa-ken 242, Japan

Tel: +81 462 73 5436

Fax: +81 462 73 7415

E-mail: kido@vnet.ibm.com

Annex

(informative)

Rules of unification and arrangement for CJK Ideographs

The CJK Unified Ideographs subset contains 20,902 ideographs that are produced based on over 54,000 source ideographs. They are unified and arranged under some rules. Rules are shown as follows.

Note

- 1 These rules are applied only for unification of CJK UNIFIED IDEOGRAPHS repertoire.
- 2 These rules may change in future extensions.

1. Source code separation rule

Ideographs that are distinct in one of the sources listed below must not be unified.

Sources are:

G-source: GB2312-80, GB12345-90, GB7589-87, GB7590-87, GB8565-88,

General Purpose Hanzi List for Modern Chinese Language

T-source: TCA-CNS 11643/1st plane, TCA-CNS 11643/2nd plane,

TCA-CNS 11643/14th plane

J-source: JIS X 0208-1990, JIS X 0212-1990

K-source: KS C 5601-1989, KS C 5657-1991

Source code separation rule is applied prior to other unification rules.

2. Unification rules

2.1 Scope of unification

In general, ideographs that are unrelated in historical derivation (non-cognate characters) will not be unified.

Example: 士, 士

2.2 Two level classification

The two level classification consists of the difference of abstract shapes and the difference of actual shapes with actual typeface designs. Variant forms are identified based on the difference of abstract shapes.

2.3 Rules

The two level classification can be performed with the following algorithm, which is applied to each ideograph to be compared.

2.3.1 Analysis of component structure

The component structure of each ideograph is examined. The component is a geometrical combination of primitive elements, and is able to configure an alternative ideograph in conjunction with other components. Some components can be combined to be a component with more complicated structure. As for an ideograph, therefore, there can be defined a component tree with the top node of the ideograph and the bottom nodes of primitive elements as shown in Fig.1.

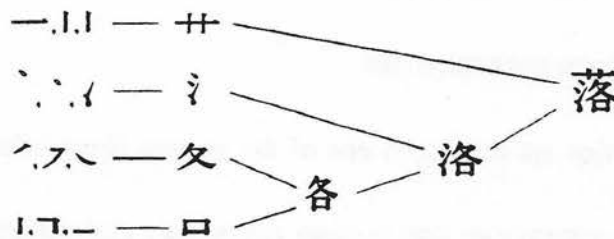


Figure 1 - Component structure

2.3.2 Analysis of component features

In this procedure, such components located at the most superior node are employed as shown in Fig.2.

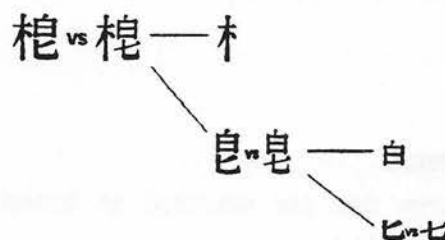


Figure 2 - The most superior node of component

- The following features of each ideograph to be compared are examined:
 - a* : number of components
 - b* : relative position of components in each complete ideograph
 - c* : structure of corresponding component
 - d* : treatment in a source ideograph set

e : radical(s) contained in a component

- If one or more features (*a* to *e* above) are different between the ideographs to be compared, the ideographs are considered to have different abstract shapes and are therefore not to be unified.
- If all the features (*a* to *e* above) are identical between the ideographs, the ideographs are considered to have the same abstract shape and are therefore to be unified.

2.4 Examples of differences of abstract shapes (not to be unified)

To illustrate the above rules (*a* to *e* in 2.3.2), some typical examples of differences of abstract shapes are shown:

2.4.1 Examples of different number of components (*a*)

崖·厓, 肱·肱, 降·冢

2.4.2 Examples with the same number of components, but with different relative positions of components (*b*)

峰·峯, 荊·荊

2.4.3 Examples with the same number and the same relative position of components, but with different structures of corresponding components (*c*)

穉·秋, 扌·擴, 策·箒

2.4.4 Examples treated differently in a source ideograph set (*d*)

𠂇·𠂇, 𠂇·𠂇, 𠂇·𠂇, 𠂇·𠂇,
夾·夾, 單·單, 隹·隹, 𠂇·𠂇,
贊·贊, 襄·襄, 隹·隹, 間·間,
朶·朶, 隹·隹, 恒·恆, 人·人

2.4.5 Examples with different radical(s) in a component (e)

禾·示, 文·又

2.5 Examples of differences of actual shapes (to be unified)

The examples shown here are considered to have the same abstract shape, but to have different actual shapes.

讠·讠·讠, 示·示·木, 艮·艮·艮, 食·食·食,
 黄·黄, 盥·盥, 曷·曷, 包·包,
 青·青, 每·每, 册·册, 爭·爭,
 畺·畺·畺, 畵·畵, 步·步, 者·者,
 臭·臭, 并·并, 开·开·开, 骨·骨,
 呂·呂, 直·直, 鼎·鼎, 吳·吳·吳,
 眞·眞·眞, 爲·為, 單·单, 曾·曾·曾,
 成·成, 專·專, 內·內, 卅·卅

The differences are further classified according to the following examples.

a) Differences in writing sequence

周·周, 告·告

b) Differences in rotated strokes/dots

半·半, 勺·勺, 羽·羽·羽,
 酉·酉, 兼·兼, 益·益

c) Differences in overshoot at the stroke initiation and/or termination

身·身，雪·雪，拐·拐，
不·不，非·非

d) Differences in contact of strokes

奥·奥，酉·酉，兕·兕，
查·查，奔·奔

e) Differences in protrusion at the folded corner of strokes

巨·巨

f) Differences in bent strokes

西·西

g) Differences in folding back at the stroke termination

朱·朱

h) Differences in accent at the stroke initiation

父·父，丈·丈，互·互

j) Differences in "rooftop" modification

八·八，宀·宀

k) Combinations of the above differences

刃·刃·刃，

The differences in actual shapes are illustrated in clause 26 in the four columns representation.

3. Source code separation

The following list shows exceptions to the unification rules caused by the source code separation.

丟 丟	T	偷 偷	T	劒 劒	J
4E1F 4E22		5077 5078		5292 5294	
么 么	GT	偽 偽	TJ	勻 勻	T
4E48 5E7A		507D 50DE		52FB 5300	
争 爭	GTJ	冊 冊	TJ	单 单	T
4E89 722D		518A 518C		5355 5358	
仞 仞	J	刃 刃	TJ	即 即	TK
4EDE 4EED		5203 5204		5373 537D	
併 併	T	刊 刊	T	卷 卷	J
4F75 5002		520A 520B		5377 5DFB	
侶 侶	T	刪 刪	T	叁 叁	T
4FA3 4FB6		5220 522A		53C1 53C2	
俁 俁	T	別 別	T	參 參	T
4FC1 4FE3		5225 522B		53C3 53C4	
俞 俞	T	剎 剎	T	呂 呂	T
4FDE 516A		5239 524E		5415 5442	
俱 俱	T	𠂔 𠂔	T	吞 吞	T
4FF1 5036		524F 5259		541E 5451	
值 值	T	剝 剝	T	吳 吳 吳	TJ
5024 503C		525D 5265		5433 5434 5449	

呐 呐
5436 5450

T

告 告
543F 544A

T

唧 唧
5527 559E

T

喻 喻
55A9 55BB

T

嘘 嘘
5618 5653

T

国 国
56EF 56FD

T

圈 圈
5708 570F

TJ

垚 垚
5759 5DE0

T

埒 埒
57D2 57D3

J

埒 埒
57F3 586A

T

塹 塹
5848 588D

T

填 填
5861 586B

TJ

增 增
5897 589E

T

壮 壮
58EE 58EF

TJ

簋 簋
5910 657B

T

本 本
5932 672C

J

奥 奥
5965 5967

J

獎 獎 獎
5968 596C 734E

TJ

妆 妝
5986 599D

T

妍 妍
598D 59F8

T

姍 姍
59CD 59D7

T

娛 娛 娛
5A1B 5A2F 5A31

T

婕 婕
5A55 5AAB

T

嫵 嫵
5A7E 5AAE

T

媪 媪
5AAA 5ABC

T

嫫 嫫
5AAF 5B00

T

嫫 嫫
5B0E 5B14

T

孳 孳
5B73 5B76

T

宫 宮
5BAB 5BAE

T

寬 寬
5BDB 5BEC

T

寧 寧
5BDC 5BE7

T

寢 寢
5BDD 5BE2

TJ

專 專
5C02 5C08

J

将 將
5C06 5C07

TJ

尔 尔
5C13 5C14

T

尙 尙
5C19 5C1A

T

尫 尫
5C2A 5C2B

T

檻 檻
5C36 5C37

T

屏 屏
5C4F 5C5B

T

崢 崢
5CE5 5D22

GT

帡 帡
5E21 5E32

T

并 并
5E76 5E77

T

廐 廐
5EC4 5ECF

T

弑 弑
5F11 5F12

T

弹 彈
5F39 5F3E

T

𠂇 𠂇 TJ
5F50 5F51

𠂇 𠂇 T
5F54 5F55

彙 彙 T
5F59 5F5A

彙 彙 J
5F5B 5F5C

彙 彙 T
5F65 5F66

徵 徵 T
5FB4 5FB5

惠 惠 TJ
6075 60E0

悅 悅 T
6085 60A6

惠 惠 T
60B3 60EA

愠 愠 T
6120 614D

愠 愠 TJ
613C 614E

戲 戲 T
622F 6231

戶 戶 戶 T
6236 6237 6238

掄 掄 T
6329 635D

插 插 插 TJ
633F 63D2 63F7

捏 捏 T
634F 63D1

搜 搜 TJ
635C 641C

揭 揭 T
63B2 63ED

搖 搖 搖 TJ
63FA 6416 6447

搵 搵 T
63FE 6435

敫 敫 T
6553 655A

既 既 T
65E2 65E3

晚 晚 T
665A 6669

暨 暨 T
66A8 66C1

曾 曾 J
66FD 66FE

查 查 T
67E5 67FB

柵 柵 T
67F5 6805

稅 稅 T
68B2 68C1

榆 榆 T
6961 6986

榼 榼 T
6985 69B2

檇 檇 T
699D 6A27

楨 楨 J
69C7 69D9

橫 橫 T
6A2A 6A6B

步 步 T
6B65 6B69

殼 殼 TJ
6BBB 6BBC

毀 毀 T
6BC0 6BC1

每 每 T
6BCE 6BCF

氫 氫 T
6C32 6C33

淨 淨 T
6D44 6DE8

涉 涉 T
6D89 6E09

浼 浼 T
6D97 6D9A

淚 淚 T
6D99 6DDA

淥 淥 T
6DE5 6E0C

清 清 T
6DF8 6E05

渴 渴 T
6E07 6E34

溫 溫
6E29 6EAB

T

滙 滙
6E88 6F59

T

漑 漑
6E89 6F11

T

為 爲
70BA 7232

TJ

煢 煢
712D 7162

GT

煨 煨
7174 7185

T

狀 狀
72B6 72C0

T

瑤 瑤
7464 7476

TJ

瓶 瓶
74F6 7501

T

產 產
7522 7523

T

瘦 瘦
75E9 7626

J

眞 眞
771E 771F

TJ

研 研
7814 784F

T

祿 祿
797F 7984

TJ

褚 褚
7987 891A

T

禿 禿
79BF 79C3

T

稅 稅
7A05 7A0E

T

穗 穗
7A42 7A57

TJ

箏 箏
7B5D 7B8F

GJ

箴 箴
7BB3 7C08

T

粵 粵
7CA4 7CB5

T

絕 絕
7D55 7D76

T

綠 綠
7DA0 7DD1

T

緒 緒
7DD2 7DD6

T

緣 緣
7DE3 7E01

T

緼 緼
7DFC 7E15

T

胼 胼
80FC 8141

T

脫 脫
812B 8131

T

膾 膾
817D 8183

T

舍 舍
820D 820E

TJ

鋪 鋪
8216 8217

J

莊 莊
8358 838A

TJ

菑 菑
83D1 8458

TJ

藍 藍
8480 8495

T

蔣 蔣
848B 8523

J

薦 薦
848D 853F

T

蒞 蒞
8570 8580

T

薰 薰
85AB 85B0

T

蘊 蘊
85F4 860A

T

虛 虛
865A 865B

T

蛻 蛻
86FB 8715

T

裝 裝
88C5 88DD

JK

訐 訐
8A2E 8A7D

T

說 說
8AAA 8AAC

T

諫 諫
8ACC 8AEB

J

謠 謠	J	鎮 鎮	TJ	鳳 鳳	T
8B20 8B21		93AD 93AE		9CEF 9CF3	
𪔐 𪔐	T	閱 閱	T	鶻 鶻	J
8C5C 8C63		95B1 95B2		9D87 9DAB	
𪔐 𪔐	T	青 青	T	鷓 鷓	J
8EFF 8F27		9751 9752		9DC6 9DCF	
輜 輜	J	靜 靜	GTJ	麪 麪	T
8F1C 8F3A		9759 975C		9EAA 9EAB	
輜 輜	T	靱 靱	J	麼 麼	T
8F3C 8F40		976D 9771		9EBC 9EBD	
迸 迸	T	頰 頰	T	黃 黃	T
8FF8 902C		9839 983D		9EC3 9EC4	
遙 遙	J	顏 顏	TJ	黑 黑	T
9059 9065		984F 9854		9ED1 9ED2	
邢 邢	T	顛 顛	J		
90A2 90C9		985A 985B			
郎 郎	T	飲 飲	J		
90CE 90DE		98EE 98F2			
醞 醞	T	餅 餅	TJ		
9196 919E		9905 9920			
醬 醬	J	駢 駢	T		
91A4 91AC		99E2 9A08			
鉞 鉞	T	飢 飢	T		
9203 9292		9AA9 9AAB			
銳 銳	T	高 高	T		
92B3 92ED		9AD8 9AD9			
錄 錄	T	鬪 鬪	T		
9304 9332		9B2C 9B2D			
鍊 鍊	T	鰓 鰓	TJ		
932C 934A		9C1B 9C2E			

4. Arrangement rules

4.1 Scope of arrangement

The CJK unified ideographs are arranged following the order based on the dictionaries listed below.

Priority	Dictionary	Version
1	Kangxi Dictionary 康熙字典	Beijing, 7th edition
2	Daikanwa Jiten 大漢和辭典	9th edition
3	Hanyu Dazidian 汉语大字典	1st edition
4	Daejaweon 大字源	1st edition

4.2 Rules

4.2.1 Ideographs found in the dictionaries

- a) If an ideograph is found in the Kangxi Dictionary, it follows the Kangxi Dictionary order.
- b) If an ideograph is not found in the Kangxi Dictionary but found in the Daikanwa Jiten, it is given a position at the end of the radical-stroke group under which is indexed the nearest preceding Daikanwa Jiten character that also appears in the Kangxi dictionary.
- c) If an ideograph is found in neither the Kangxi nor the Daikanwa, the Hanyu Dazidian and the Daejaweon dictionaries are referred to with a similar procedure.

4.2.2 Ideographs not found in any of the four dictionaries

If an ideograph is not found in any of the four dictionaries, it is given a position at the end of the characters present in the dictionaries indexed under the same radical-stroke count.

5. Details of Source Ideograph sets

Details of source ideograph sets are shown as follows.

Source		Primary Source			Code of BMP
ID	Characters of BMP	Source Name	Characters of Primary Source.	Characters of BMP	
G0	6,763	GB2312-80	6,763	6,763	-
G1	2,352	GB12345-90	6,866	2,192	-
		Hong Kong Characters	-	58	See Note 1
		Idu Characters	-	92	See Note 2
G3	4,835	GB7589-87	7,237	4,835	-
G5	2,842	GB7590-87	7,039	2,842	-
G7	42	General Purpose Hanzi List for Modern Chinese Language	7,000	42	-
G8	290	GB8565-88	7,053	290	-
T1	5,410	TCA-CNS 11643/1st plane	5,401	5,401	-
		TCA-CNS 11643/1st plane Symbols	-	9	See Note 3
T2	7,650	TCA-CNS 11643/2nd plane	7,650	7,650	-
TE	4,198	TCA-CNS 11643/14th plane	6,319	3,951	-
		CCCII Characters	-	237	See Note 4
		Cantonese Characters	-	10	See Note 5
J0	6,356	JIS X 0208-1990	6,355	6,355	-
		JIS X 0208-1990 Symbol	-	1	4EDD
J1	5,801	JIS X 0212-1990	5,801	5,801	-
K0	4,620	KS C 5601-1987	4,888	4,620	See Note 6
K1	2,856	KS C 5657-1991	2,856	2,856	-

Note

1 BMP codes of Hong Kong Characters on G1 source are shown as follows.

4E78, 519A, 51A7, 528F, 53FE, 5497, 54CB, 54E3, 551E, 5525, 5569, 556B, 5571, 5572, 5579, 5590, 55BC, 55F0, 55F1, 5605, 5622, 5625, 56A1, 56A4, 56B9, 56BF, 56D6, 5AF2, 5B6D, 5C83, 63B9, 63FC, 64DD, 6530, 66F1, 713E, 7140, 71F6, 7534, 7666, 7793, 77CB, 7F49, 810C, 8137, 81A5, 81B6, 8226, 8254, 8804, 88C7, 8E0E, 8E2D, 8E80, 8EDA, 922A, 9385, 9BED

2 BMP codes of Idu Characters on G1 source are shown as follows.

4E06, 4E37, 4E41, 4E4A, 4E5B, 4E64, 4E65, 4E6A, 4E6B, 4E6C, 4E6D, 4E6E, 4E6F, 4E72,
4E76, 4E7A, 4E7B, 4E7C, 4E7D, 4E87, 4EAA, 4EBD, 4ED2, 4FA4, 516F, 517A, 536A, 5381,
5391, 53BC, 54DB, 551F, 55B8, 55ED, 5655, 56CD, 56D5, 5788, 58ED, 591E, 593B, 5AA4,
5CBE, 5DEA, 5DEC, 5DED, 5DFC, 5EE4, 603E, 65C0, 65D5, 6729, 6730, 680D, 6927, 693A,
698B, 698C, 6A74, 6A75, 6A7B, 6AF7, 6B0C, 6B15, 6D4C, 70BA, 7233, 729C, 7320, 7364,
7553, 785B, 7873, 78D7, 7A24, 7A52, 7B7D, 7E07, 7F56, 7F98, 824D, 83BB, 848A, 8644,
87A6, 8968, 8FF2, 9007, 9064, 95AA, 95CF, 97B0

3 BMP codes of TCA-CNS 11643/1 Symbols on T1 source are shown as follows.

5159, 515B, 515D, 515E, 5161, 5163, 55E7, 74E9, 7CCE

4 BMP codes of CCCII Characters on TE source are shown as follows.

4F66, 4F68, 4FE7, 503F, 50A6, 510F, 517F, 523E, 5324, 5365, 539B, 54CB, 556B, 5571,
5573, 55F4, 5620, 5622, 5691, 5692, 56B0, 56BA, 5759, 578A, 580F, 5812, 5813, 5847,
589B, 5900, 5911, 594D, 5AD1, 5AD3, 5B67, 5C21, 5C57, 5C77, 5CD5, 5D75, 5D8E, 5DA5,
5DB6, 5DBF, 5E65, 5E71, 5E85, 5ECD, 5EED, 5F94, 5F9A, 5FBA, 6125, 6150, 61D3, 62A3,
6360, 6364, 63B6, 6403, 6449, 64B6, 651A, 66E2, 6702, 6711, 6725, 67A4, 67AC, 6806,
6810, 685A, 685E, 6929, 692C, 6A2D, 6A77, 6A7A, 6ACA, 6AE6, 6AF5, 6B0D, 6B0E, 6BDC,
6BDD, 6BF6, 6C1E, 6C63, 6DA5, 6E0F, 6E7C, 6E84, 6E8A, 6E8B, 6F48, 6F49, 6F4C, 6F99,
6F9D, 6FF8, 702D, 702E, 705C, 70BF, 70E5, 70EA, 7111, 7112, 7139, 713B, 713D, 713F,
7171, 7175, 7176, 7177, 7193, 7196, 71B4, 71DD, 71DE, 720E, 7218, 7347, 7348, 73EF,
7412, 743B, 748D, 74A4, 74B4, 7673, 7677, 76BC, 7819, 781B, 781E, 783D, 7853, 7854,
7858, 78B7, 78D8, 78EE, 7922, 794D, 7986, 7999, 79A3, 79BC, 79CC, 7A25, 7AA7, 7B37,
7B59, 7BD0, 7C2F, 7C32, 7C42, 7C4E, 7C68, 7CA9, 7CED, 7DD0, 7DD3, 7E07, 7E64, 7F40,
8002, 8041, 8063, 80BB, 8248, 8310, 8312, 8362, 841E, 8421, 84DE, 84E1, 84E2, 8573,
85D4, 85F5, 8637, 8645, 8672, 874A, 87A5, 87A9, 87F5, 8834, 8850, 8887, 8954, 8984,
8B03, 8C52, 8CD8, 8DOC, 8D18, 8DB0, 8EBC, 8ED5, 8FAA, 909C, 915C, 9221, 922B, 9273,
92F4, 92F5, 933F, 9342, 9386, 93BC, 93BD, 93BE, 93EF, 93F1, 93F2, 9422, 9423, 9424,
9466, 9467, 9597, 95CE, 95E7, 973B, 974D, 98E4, 990E, 9942, 9B1D, 9B98, 9D49

5 BMP codes of Cantonese Characters on TE source are shown as follows.

5528, 5572, 55BA, 55EE, 55F0, 56B8, 56B9, 56C4, 8053, 92B0

6 The following 268 ideographs from KS C 5601-1981 are located in the CJK

COMPATIBILITY IDEOGRAPHS, because they duplicated existing ideographs in order
to denote alternative pronunciation.

F900 - FA0B