

Date: 1994.07.07

Supersedes: N071

ISO/IEC JTC 1/SC2/WG2/IRG**Ideograph Rapporteur Group****Secretariat: IPSJ/ITSCJ (Japan)**

Source / Contribution Identifier: T. Koike/TK-19

Meeting: @ / for Beijing '94

Issue Number: Project / Task Identifier:

TITLE: Draft text of ANNEX: Rules of CJK Unified Ideographs

Status:

Keywords:

Short Description:

This draft text is revised version of N072. Source code separation list and details of source ideographs sets are not completed.

Proposed Conclusion / Requested Action:

- 1. To be discussed and confirmed on the 3rd IRG Beijing Meeting.**
- 2. Roll call the data or opinion from China and Korea as to non-completed item.**

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 CJK Unified 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 the future extension

1. Source code separation rule

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

G-source: GB 2312-80, GB 12345-90, GB 7589-87, GB 7590-87, GB 8565-89,

General purpose Hanzi list for modern Chinese language

T-source: TCA-CNS11643/1st plane, TCA-CNS11643/2nd plane,
TCA-CNS11643/14th 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 prior to other unification rules.

2. Unification/separation rule

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 2-level classification

The 2-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 2-level classification can be performed with the following algorithm, which is applied to each ideograph to be compared:

a) 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 top node of the ideograph and bottom nodes of primitive elements as shown in Fig.1.

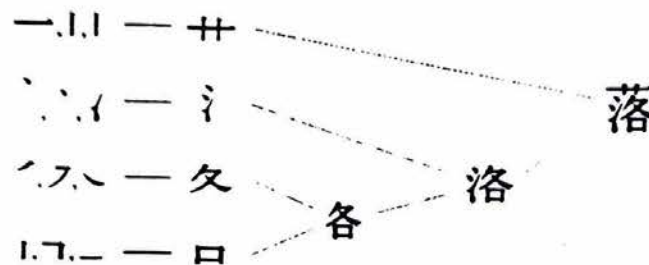


Figure 1 Component structure

In the next procedure, such components located at the most superior node are employed (Fig.2).

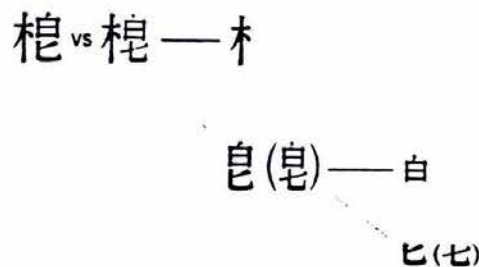


Figure 2-The most superior node of component

- b) 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
- c) If one or more features (*a* to *e*) are different between the ideographs to be compared, the ideographs are defined to have difference in abstract shapes and are therefore to be

separated.

d) If all the features (*a* to *e*) are identical between the ideographs, the ideographs are considered to have identical abstract shape and therefore to be unified.

2.4 Example of differences of abstract shapes (to be separated)

To make the above discussion clearer, some typical examples of difference of abstract shapes are shown:

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

崖·厓, 肱·肱, 降·𡿨

2.4.2 Examples with the same number of components, but with the difference in relative position of components (*b*)

峰·峯, 荊·荊

2.4.3 Examples with the same number and the same relative position of components but with the different structure of a corresponding component (*c*)

種·秋, 扌·扌, 策·箒

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

𠂔·𠂔, 𠂔·𠂔, 𠂔·𠂔, 𠂔·𠂔,
 夾·夾, 單·單, 隹·隹, 𠂔·𠂔,
 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·禾, 示·示, 文·文

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

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

迂·迂·迂, 示·示·示, 艮·艮·艮, 食·食·食,
 黄·黄, 盥·盥, 曷·曷, 包·包,
 青·青, 每·每, 册·册, 爭·爭,
 𡗗·𡗗·𡗗, 𡗗·𡗗, 步·步, 者·者,
 臭·臭, 并·并, 开·开·开, 骨·骨,
 呂·呂, 直·直, 𡗗·𡗗, 吳·吳·吳,
 眞·眞·眞, 爲·為, 單·单, 曾·曾·曾,
 成·成, 專·專, 內·內, 𡗗·𡗗

The differences are further classified as follows:

2.5.1 Examples of the difference in writing sequences

周·周, 告·告

2.5.2 Examples of the difference in rotated strokes/dots

半·半, 勺·勺, 羽·羽·羽,
 酉·酉, 兼·兼, 益·益

2.5.3 Examples of the difference in overshoot at the stroke termination

身·身, 雪·雪, 拐·拐, 刃·刃·刃,
不·不, 非·非

2.5.4 Examples of the difference in contact of strokes

奥·奥, 酉·酉, 兕·兕, 奔·奔,
查·查

2.5.5 Examples of the difference in protruding at the folded corner of strokes

巨·巨

2.5.6 Examples of the difference in bent strokes

西·西

2.5.7 Examples of the difference in folding back at the stroke termination

朱·朱

2.5.8 Examples of the difference in accent at the stroke initiation

父·父, 丈·丈, 互·互

2.5.9 Examples of the difference in "rooftop" modification

八·八, 穴·穴

2.5.10 Four-columns representation

Four-columns representation shows concrete actual shape difference.

3. Source code separation

not completed

The following listed exceptions to the unification rules caused by the source code separation rule.

丟丟	T	冊冊	TJ	卷卷	J
4E1F 4E22		518A 518C		5377 5DFB	
么么	T	刃刃	TJ	叁叁	T
4E48 5E7A		5203 5204		53C1 53C2	
争争	TJ	刊刊	T	叁叁	T
4E89 722D		520A 520B		53C3 53C4	
仞仞	J	刪刪	T	呂呂	T
4EDE 4EED		5220 522A		5415 5442	
併併	T	別別	T	吳吳吳	TJ
4F75 5002		5225 522B		5433 5434 5449	
侶侶	T	剎剎	T	訥訥	T
4FA3 4FB6		5239 524E		5436 5450	
俣俣	T	勑勑	T	告告	T
4FC1 4FE3		524F 5259		543F 544A	
俞俞	T	剝剝	T	唧唧	T
4FDE 516A		525D 5265		5527 559E	
俱俱	T	劒劒	J	喻喻	T
4FF1 5036		5292 5294		55A9 55BB	
值值	T	勻勻	T	噓噓	T
5024 503C		52FB 5300		5618 5653	
偷偷	T	单单	T	国国	T
5077 5078		5355 5358		56EF 56FD	
偽偽	TJ	即即	TK	圈圈	TJ
507D 50DE		5373 537D		5708 570F	

垚 垚	T	婕 婕	T	檻 檻	T
5759 5DE0		5A55 5AAB		5C36 5C37	
埈 埈	J	嫃 嫃	T	屏 屏	T
57D2 57D3		5A7E 5AAE		5C4F 5C5B	
埊 埊	T	媪 媪	T	崢 崢	T
57F3 586A		5AAA 5ABC		5CE5 5D22	
堅 堅	T	媯 媯	T	𪛗 𪛗	T
5848 588D		5AAF 5B00		5E21 5E32	
埧 埧	TJ	嬖 嬖	T	并 并	T
5861 586B		5B0E 5B14		5E76 5E77	
增 增	T	孳 孳	T	廨 廨	T
5897 589E		5B73 5B76		5EC4 5ECF	
壯 壯	TJ	宮 宮	T	弑 弑	T
58EE 58EF		5BAB 5BAE		5F11 5F12	
夤 夤	T	寬 寬	T	彈 彈	T
5910 657B		5BDB 5BEC		5F39 5F3E	
本 本	J	寧 寧	T	𠂇 𠂇	TJ
5932 672C		5BDC 5BE7		5F50 5F51	
奧 奧	J	寢 寢	TJ	𠂇 𠂇	T
5965 5967		5BDD 5BE2		5F54 5F55	
獎 獎 獎	TJ	專 專	J	彙 彙	T
5968 596C 734E		5C02 5C08		5F59 5F5A	
妝 妝	T	將 將	TJ	彝 彝	J
5986 599D		5C06 5C07		5F5B 5F5C	
妍 妍	T	尔 尔	T	彦 彦	T
598D 59F8		5C13 5C14		5F65 5F66	
姍 姍	T	尙 尙	T	徵 徵	T
59CD 59D7		5C19 5C1A		5FB4 5FB5	
娛 娛 娛	T	尙 尙	T	惠 惠	TJ
5A1B 5A2F 5A31		5C2A 5C2B		6075 60E0	

悅 悅	T	晚 晚	T	氫 氫	T
6085 60A6		665A 6669		6C32 6C33	
惠 惠	T	暨 暨	T	淨 淨	T
60B3 60EA		66A8 66C1		6D44 6DE8	
愠 愠	T	曾 曾	J	涉 涉	T
6120 614D		66FD 66FE		6D89 6E09	
愼 愼	TJ	查 查	T	浼 浼	T
613C 614E		67E5 67FB		6D97 6D9A	
戲 戲	T	柵 柵	T	淚 淚	T
622F 6231		67F5 6805		6D99 6DDA	
戶 戶 戶	T	稅 稅	T	淥 淥	T
6236 6237 6238		68B2 68C1		6DE5 6E0C	
掄 掄	T	榆 榆	T	清 清	T
6329 635D		6961 6986		6DF8 6E05	
插 插 插	TJ	榼 榼	T	渴 渴	T
633F 63D2 63F7		6985 69B2		6E07 6E34	
捏 捏	T	櫟 櫟	T	溫 溫	T
634F 63D1		699D 6A27		6E29 6EAB	
搜 搜	TJ	楨 楨	J	漚 漚	T
635C 641C		69C7 69D9		6E88 6F59	
揭 揭	T	橫 橫	T	漑 漑	T
63B2 63ED		6A2A 6A6B		6E89 6F11	
搖 搖 搖	TJ	步 步	T	為 為	TJ
63FA 6416 6447		6B65 6B69		70BA 7232	
搵 搵	T	殼 殼	TJ	堯 堯	T
63FE 6435		6BBB 6BBC		712D 7162	
敝 敝	T	毀 毀	T	煨 煨	T
6553 655A		6BC0 6BC1		7174 7185	
既 既	T	每 每	T	狀 狀	T
65E2 65E3		6BCE 6BCF		72B6 72C0	

瑤 瑤	TJ	綠 綠	T	薰 薰	T
7464 7476		7DA0 7DD1		85AB 85B0	
瓶 瓶	T	緒 緒	T	蘊 蘊	T
74F6 7501		7DD2 7DD6		85F4 860A	
產 產	T	緣 緣	T	虛 虛	T
7522 7523		7DE3 7E01		865A 865B	
瘦 瘦	J	縕 縕	T	蛻 蛻	T
75E9 7626		7DFC 7E15		86FB 8715	
眞 眞	TJ	胼 胼	T	衛 衛	JK
771E 771F		80FC 8141		885B 885E	
研 研	T	脫 脫	T	裝 裝	JK
7814 784F		812B 8131		88C5 88DD	
祿 祿	TJ	膾 膾	T	訐 訐	T
797F 7984		817D 8183		8A2E 8A7D	
褚 褚	T	舍 舍	TJ	說 說	T
7987 891A		820D 820E		8AAA 8AAC	
禿 禿	T	舖 舖	J	諫 諫	J
79BF 79C3		8216 8217		8ACC 8AEB	
稅 稅	T	莊 莊	TJ	謠 謠	J
7A05 7A0E		8358 838A		8B20 8B21	
穗 穗	TJ	菑 菑	TJ	𪔐 𪔐	T
7A42 7A57		83D1 8458		8C5C 8C63	
箏 箏	J	藍 藍	T	𪔐 𪔐	T
7B5D 7B8F		8480 8495		8EFF 8F27	
箴 箴	T	蔣 蔣	J	輜 輜	J
7BB3 7C08		848B 8523		8F1C 8F3A	
粵 粵	T	蔦 蔦	T	輶 輶	T
7CA4 7CB5		848D 853F		8F3C 8F40	
絕 絕	T	蒞 蒞	T	迸 迸	T
7D55 7D76		8570 8580		8FF8 902C	

遙 遙
9059 9065

J

邢 邢
90A2 90C9

T

郎 郎
90CE 90DE

T

醞 醞
9196 919E

T

醬 醬
91A4 91AC

J

鉞 鉞
9203 9292

T

銳 銳
92B3 92ED

T

錄 錄
9304 9332

T

鍊 鍊
932C 934A

T

鎮 鎮
93AD 93AE

TJ

閱 閱
95B1 95B2

T

青 青
9751 9752

T

靜 靜
9759 975C

TJ

靱 靱
976D 9771

J

頹 頹
9839 983D

T

顏 顏
984F 9854

TJ

顛 顛
985A 985B

J

飲 飲
98EE 98F2

J

餅 餅
9905 9920

TJ

駢 駢
99E2 9A08

T

飢 飢
9AA9 9AAB

T

高 高
9AD8 9AD9

T

鬪 鬪
9B2C 9B2D

T

鯁 鯁
9C1B 9C2E

TJ

鳳 鳳
9CEF 9CF3

T

鶉 鶉
9D87 9DAB

J

鷓 鷓
9DC6 9DCF

J

麪 麪
9EAA 9EAB

T

麼 麼
9EBC 9EBD

T

黃 黃
9EC3 9EC4

T

黑 黑
9ED1 9ED2

T

4. Arrangement rules

4.1 Scope of arrangement

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

Priority	Dictionary	Version
1	Kangxi Dictionary	Beijing, 7th edition
2	Daikanwa Jiten	9th edition
3	Hanyu Dazidian	1st edition
4	Daejaweon	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 end of 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 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 Ideographs sets

Details of source Ideographs sets are shown as follows.

not completed

Source		Primary Source			Code of BMP
ID	Characters of BMP	Source Name	Characters of P.S.	Characters of BMP	
G0	6,763	GB 2312-1989	6,763	6,763	-
G1	2,352	GB 12345-1990		2,192	-
		Hong Kong Character	-	58	See Note 1
		Idu Character	-	92	See Note 2
G3	4,835	GB 7589-1987	7,237	4,835	-
G5	2,842	GB 7590-1987	7,039	2,842	-
G7	42	General purpose Hanzi list for modern Chinese language	-	42	-
G8	290	GB 8565-1989		290	-
T1	5,410	TCA-CNS 11643 1st plane	5,401	5,401	-
		TCA-CNS 11643 1st plane Symbol	-	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 Character	-	237	see Note 4
		Cantonese Character		10	see Note 5
J0	6,356	JIS X0208-1990	6,355	6,355	-
		JIS X0208-1990 Symbol	-	1	4EDD
J1	5,801	JIS X0212-1990	5,801	5,801	-
K0	4,620	KS C 5601-1987	4,888	4,620	-
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.

4E6A, 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, 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,
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, 4F6S, 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, 6FFS, 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, 74SD, 74A4, 74B4, 7673, 7677, 76BC, 7819, 781B, 781E, 783D, 7853, 7854,
7858, 78B7, 78D6, 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, 89S4,
8B03, 8C52, 8CD8, 8D0C, 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