

Date: 2008-03-21

**ISO/IEC JTC1/SC2/WG2
Coded Character Set
Secretariat: Japan (JISC)**

Doc. Type: Input to ISO/IEC 10646

Title: Representation of CJK Unified Ideographs in multi-column

Source: Michel Suignard, project editor

Project: JTC1 02.10646

Status: For review by WG2

Date: 2008-03-21

Distribution: WG2

Reference:

Medium:

The need to publish a new edition of ISO/IEC 10646 creates the requirement to republish all code charts containing CJK Unified Ideographs. However, sources have grown from eight to nine with the addition of the M source (Macao SAR) since the last edition: ISO/IEC 10646:2003. In addition, even in the last edition, three sources: KP (Democratic People Republic of Korea), H (Hong-Kong SAR), and U (UTC) were not adequately represented in the multi-column charts. However, it is not practical to show all nine sources for all CJK characters because no character has all nine sources, and very few characters go beyond 6 sources. Furthermore, the CJK extensions (A, B, and C) have even less sources, calling for a more optimal format based on average number of sources for a given collection.

In fact, if one looks at the various CJK Unified collections, the number optimal of sources to be shown varies from 6 (CJK Unified Ideograph: 4E00-9FC5), 4 (CJK Unified Ideographs Extension A 3400-4DB5), and 2 (CJK Unified Ideographs Extension B: 20000-2A6D6, and proposed CJK Unified Ideographs extension C: 2A700-2B734).

The following pages show some examples of what could be done as a compromise between size and legibility. The main focus is to provide a more consistent amount of information per page, thus allowing a variable amount of columns per characters depending on the collections. Because the standard is typically not printed in its entirety anymore, paper preservation is not as critical; this in return allows some flexibility. In addition, some additional information is provided such as the radical image and the 'radical, stroke' count for each ideograph.

There are still some issues worth considering:

- For the main CJK Unified Ideographs collection (4E00-9FC5), should we preserve spatial positioning for the sources, therefore requiring frequently 2 rows when a source is not part of the first six sources, or should we show these additional sources highlighted in the same row? (The author prefers the second solution)
- Can we get away with the spatial positioning for all other collections, allowing a compression to 4 or even 2 sources per characters (with exceptions dealt through additional rows)? (The author is in favor)
- Some G sources (such as KangXi: G_KX) have very long labels. This can be accommodated by shortening to just the prefix (like 'G_KX') or allowing the first source (always G when existing) to bleed over the character

code position and pushing the radical.count info upward. Note that the source representation of the Japanese ARIB Kanji characters may have to be shortened to 9 characters instead of the current 11 to fit in the sub-column box. (The author prefers the second solution, that is preserving the information by tweaking the layout)

41D5	立	端	端	端	端
立 117.14					
	G_KX087321	T3-5A68	K3-2E29	KP1-63A4	

- There are multiple ‘radical,stroke’ data sets. One has to be selected. The Unicode Unihan database contains such a data set. Ideally, the information should be added in the CJKU_SR.txt file.

The following pages show some samples:

- Format using large glyph: for reference only. Although the glyphs are large the amount of information per page is limited.
- Format using 2 columns per page with 6 sub-columns per characters allowing 6 sources per line. The first six sources are always represented in the same order if present (G, T, J, K, V, and KP) in the first line. If other sources exist (H, U, and M), they are represented in that order in a second line. The main issue with that format is that it is very wasteful in space when only 2 sources exists but on different lines (such as 4E04, 4E1A, and 4E1C in the example). This format allows about 32 characters per page (versus 48 for the current multi-column format for the main CJK Unified collection).
- Variant of the format above but representing the second line sources (H, U, and M) in holes in the first line starting from the right. These sources are highlighted to make clear that they are not related to the sub-column standard source. In the example, the H source is shown in the KP column for 4E04, 4E1A, 4E1C and the V column for 4E21. This format allows about 36 characters per page in the main CJK Unified collection.
- Format using 3 columns per page with 4 sub-columns per characters allowing 4 sources per line. The sources are represented in the order of appearance in the CJKU_SR.txt (G, T, J, K, V, H, KP, U, and M). This format allows about 50-54 characters per page in the CJK Extension A collection (versus 48 for the current multi-column format for the same collection).
- Format using 5 columns per page with 2 sub-columns per characters allowing 4 sources per line. The sources are represented in the order of appearance in the CJKU_SR.txt (G, T, J, K, V, H, KP, U, and M). This format allows about 80-90 characters per page in the CJK Extension B and Extension C collections (versus 128 for the current single-column format for Ext B and 32 for the multi-column format for Ext C).

Final note: the following examples use commercial fonts available to the author to represent source characters. Many sources were unavailable and are therefore not correctly represented. No claim is made about accuracy of these glyphs. It is assumed that appropriate fonts will be delivered to the contributing editors in charge of font production. The source info has been copied manually from CJKU_SR.txt, so the examples may contain mistakes. The real process will collect data directly from the file.

Format using large glyphs (2 Cols)

4E00	一	一	一	一		丈	丈		
—									
1.0	G0-523B	T1-4421	J0-306C	K0-6C69		V1-4A25	KP0-E6DD		
	一	一				4E09	三	三	三
						—	三	三	三
	V1—4A21	KP0-FCD6				1.2	G0-487D	T1-4435	J0-3B30
4E01	丁	丁	丁	丁			三	三	
—						G0-487D	T1-4435		
1.1	G0-3621	T1-4421	J0-437A	K0-6F4B		4E0A	上	上	上
	丁	丁				—	上	上	上
	V1-4A22	KP0-E8B9				1.2	G0-494F	T1-4438	J0-3E65
4E02	𠂇	𠂇	𠂇				上	上	
—						V1-4A27	KP0-E1C2		
1.1	G5-3021	T4-2126	J1-3021			4E0B	下	下	下
4E03	七	七	七	七		—	下	下	下
—						1.2	G0-4F42	T1-4436	J0-323C
1.1	G0-465F	T1-4424	J0-3C37	K0-7652			下	下	
	七	七				V1-4A28	KP0-F2BA		
	V1—4A23	KP0-EFA6				4E0C	𠂇	𠂇	𠂇
4E04	𠂇	𠂇	𠂇	𠂇		—	𠂇	𠂇	𠂇
—						1.2	G0-5822	T2-2127	J1-3024
1.1	GE-2121	T3-2126	J1-3022	H-9EB3		4E0D	不	不	不
4E05	𠂇	𠂇	𠂇			—	不	不	不
—						1.3	G0-323B	T1-4462	J0-4954
1.1	GE-2122	T3-2125	J1-3023				不	不	
4E06	𠂇	𠂇				V1-4A29	KP0-DFBE		
—						4E0E	与	与	与
1.1	G1-7D3D	K2-2121				—	与	与	与
4E07	万	万	万	万		1.3	G0-536B	T2-212F	J0-4D3F
—							与	与	
1.2	G0-4D72	T2-2126	J0-4B7C	K0-5832		V1-4A21	KP0-FCD6		
	万	万				4E0F	𠂇	𠂇	与
	V1—4A24	KP0-DAB9				—	𠂇	𠂇	与
4E08	丈	丈	丈	丈		1.3	G3-3021	T2-212D	K2-2124
—									
1.2	G0-5549	T1-4437	J0-3E66	K0-6D5B					

Ext A format (3 Cols)

41C0	竊 竊 竊	穴 116.15 G5-5E61 T4-637C K3-2E22	41D1	啤 啤 啤 啤	立 117.8 G5-5E2A T4-3F5A K3-2E27 KP1-638A	41E3	笱 笱 笱 笱	竹 118.5 G5-6324 T4-337B K3-2E31 KP1-63E3
41C1	籛 籛 籛	穴 116.17 G3-5F71 T4-6922 KP1-6366	41D2	暨 暨 暨	立 117.11 G5-5E2D T4-4563 JA-254B	41E4	笱 笱	竹 118.5 G_KX0881 T3-3775
41C2	辛 辛	立 117.1 G_KX0870 T3-2434	41D3	𪔐 𪔐 𪔐	立 117.12 G3-5F52 T4-5752 KP1-639F	41E5	筮 筮	竹 118.5 G7-2368 T6-4276
41C3	訇 訇	立 117.3 G_KX0870 T3-2A46	41D4	羸 羸 羸	立 117.13 G3-5F54 T4-5C3B K3-2E28	41E6	筮 筮 筮	竹 118.5 G3-6324 T4-3378 JA-7347
41C4	竝 竝 竝	立 117.3 GS-237B V0-3F51 H-994B	41D5	孺 孺 孺 孺	立 117.14 G_KX08732 T3-5A68 K3-2E29 KP1-63A4	41E7	筮 筮	竹 118.6 G3-6363 T4-396D
41C5	竝 竝 竝	立 117.4 G3-5F4A T4-2A76 K3-2E22	41D6	笱 笱 笱	竹 118.3 G_KX08780 T5-2B21 K3-2E2A	41E8	笱 笱 笱	竹 118.6 G3-6371 T4-3974 KP1-63F9
41C6	竝 竝	立 117.4 G-HZ T3-2E4A	41D7	笱 笱 笱	竹 118.4 G3-634E T4-2E75 KP1-63C3	41E9	笱 笱 笱	竹 118.6 G3-6366 T4-3971 KP1-641B
41C7	竝 竝	立 117.5 GS-237A T3-3325	41D8	笱 笱 笱 笱	竹 118.4 G5-6276 T4-2E70 K3-2E2B KP1-63C8	41EA	笱 笱 笱	竹 118.6 G_KX0882 T3-3D74 V3-3649
41C8	竝 竝	立 117.5 G3-5F4E T4-2E6D	41D9	笱 笱	竹 118.4 G_KX08782 T3-3328	41EB	笱 笱 笱	竹 118.6 G3-637D T4-3F58 KP1-6414
41C9	竝 竝	立 117.5 G_HZ T3-3322	41DA	笱 笱 笱	竹 118.4 G5-6278 T4-2E77 K3-2E2C	41EC	笱 笱 笱 笱	竹 118.6 G3-636A T43972 K3-2E32 KP1-6409
41CA	竝 竝	立 117.5 JA-2549 H-8E55	41DB	笱 笱 笱 笱	竹 118.4 G_KX08791 T3-3329 V2-7F4B H-8EFE	41ED	笱 笱 笱 笱	竹 118.6 G5-632D T3-3D7A H-8D5F KP1-6402
41CB	竝 竝 竝	立 117.6 G-KX08710 T5-3446 JA-254A	41DC	笱 笱 笱 笱	竹 118.4 G3-634F T4-2E73 K3-2E2D KP1-63B6	41EE	笱 笱 笱 笱	竹 118.6 G5-6334 T4-3975 JA-254D V2-7F50
41CC	竝 竝 竝	立 117.7 G_KX08711 T3-3D6F KP1-6386	41DD	笱 笱	竹 118.4 G3-634A T4-2E72		笱	KP1-641A
41CD	竝 竝 竝 竝	立 117.7 G3-5F4F T4-396A K3-2E23 KP1-6386	41DE	笱 笱 笱 笱	竹 118.5 G3-6355 T4-3376 K3-2E2E KP1-63D0	41EF	笱 笱 笱 笱	竹 118.6 G5-632E T3-3D73 H-8E59 KP1-641A
41CE	竝 竝 竝 竝	立 117.8 G5-5E28 T3-4348 K3-2E24 KP1-638D	41DF	笱 笱 笱	竹 118.5 G5-6327 T4-337D K3-2E2F	41F0	笱 笱 笱	竹 118.6 G3-6379 T4-396F KP1-640A
41CF	竝 竝 竝 竝	立 117.8 G3-5F50 T4-3F54 K3-2E26 H-994E	41E0	笱 笱 笱 笱	竹 118.5 G_KX08792 T3-3774 V0-3F71 KP1-63EB	41F1	笱 笱 笱	竹 118.6 G_KX0884 T6-4C5B JA-254E
	竝	KP1-6390	41E1	笱 笱 笱 笱	竹 118.5 G_KX08800 T5-3448 JA-254C KP1-63E7	41F2	笱 笱	竹 118.6 GS-233A T6-4C56
41D0	竝 竝 竝	立 117.8 G3-5F51 T4-3F55 KP1-6389	41E2	笱 笱 笱 笱	竹 118.5 G3-6356 T4-3377 K3-2E30 KP1-63D1	41F3	笱	竹 118.6 JA-254F

Ext B and Ext C format (5 Cols)

20000 巳 巳 — 1.1 3_KX00750 T5-2515	20011 丙 — 1.4 G_HZ	20022 此 此 — 1.6 G_KX0078 T5-232F	20034 𧈧 — 1.7 V0-354A	20045 虜 — 1.10 G_HZ
20001 𠂇 — 1.1 G_HZ	20012 𠂇 𠂇 — 1.4 G_HZ T6-222B	20023 𧈧 — 1.6 TF-2274	20035 𧈧 — 1.8 TF-2922	20046 𧈧 𧈧 — 1.11 TF-3932 H-9376
20002 乙 — 1.1 TF-2121	20013 引 — 1.4 G_HZ	20024 𧈧 — 1.6 G_HZ	20036 𧈧 — 1.8 TF-2923	20047 𧈧 — 1.11 TF-3933
20003 𠂇 𠂇 — 1.2 G_KX0076 T6-212F	20014 亥 亥 — 1.4 G_HZ T5-214D	20025 𧈧 𧈧 — 7.5 G_HZ T6-2567	20037 𧈧 □ 30.6 G_HZ	20048 𧈧 𧈧 — 1.11 G_HZ T6-472D
20004 工 — 1.2 T6-212D	20015 互 — 1.4 G_HZ	20026 𧈧 — 1.6 G_HZ	20038 𧈧 — 1.8 G_HZ	20049 𧈧 几 16.10 G_HZ
20005 𠂇 𠂇 — 1.2 G_HZ T6-212F	20016 共 — 1.4 V0-3F5F	20027 𧈧 — 1.6 V0-354F	20039 𧈧 □ 30.6 G_HZ	2004A 𧈧 — 1.11 G_HZ
20006 丰 — 1.2 K4-0002	20017 共 — 1.4 V0-3F60	20028 𧈧 — 1.6 V2-6E21	2003A 𧈧 — 1.8 G_HZ	2004B 𧈧 — 1.11 TF-3B73
20007 且 且 — 1.3 3_KX00770 T6-2142	20018 兂 兂 — 1.5 G_KX007 T6-2340	20029 𧈧 𧈧 J 6.5 G_HZ T6-2563	2003B 𧈧 𧈧 — 1.8 G_HZ T6-2E66	2004C 𧈧 — 1.12 TF-4035
20008 丘 丘 — 1.3 3_KX00770 T6-2143	20019 𧈧 𧈧 — 1.5 G_KX007 T5-233E	2002A 其 — 1.6 V0-456C	2003C 夏 文 35.7 G_HZ	2004D 𧈧 — 1.12 TF-4075
20009 𧈧 𧈧 — 1.3 T5-2133 KP1-3408	2001A 𧈧 𧈧 — 1.5 G_KX007 T6-233F	2002B 其 — 1.6 V0-456D	2003D 𧈧 — 1.9 G_HZ	2004E 𧈧 — 1.13 H-9548
2000A 𧈧 — 1.3 G_HZ	2001B 𧈧 — 1.5 G_HZ	2002C 𧈧 𧈧 — 1.7 3_KX00781 T6-2937	2003E 𧈧 — 1.9 H-9375	2004F 𧈧 — 1.13 G_HZ
2000B 𧈧 𧈧 — 1.3 G_HZ T3-2144	2001C 𧈧 — 1.5 G_HZ	2002D 𧈧 𧈧 — 1.7 G_KX0078 T6-293A	2003F 𧈧 — 1.9 V2-6E27	20050 𧈧 — 1.13 G_HZ
𧈧 𧈧 J3-2E22 KP1-340C	2001D 𧈧 — 1.5 G_HZ	2002E 𧈧 𧈧 — 1.7 G_KX0078 T6-2938	20040 𧈧 — 1.9 V0-3F68	20051 𧈧 — 1.13 V0-354C
2000C 𧈧 — 1.3 G_HZ	2001E 𧈧 — 1.5 G_HZ	2002F 𧈧 □ 31.5 G_HZ	20041 𧈧 𧈧 — 1.10 G_KX0078 T5-3072	20052 𧈧 — 1.13 TF-4742
2000D 𧈧 — 1.4 3_KX00771	2001F 𧈧 — 1.5 G_HZ	20030 𧈧 — 1.7 G_HZ	𧈧 KP1-341E	20053 𧈧 — 1.14 TF-4D56
2000E 𧈧 — 1.4 G_4K	20020 𧈧 𧈧 — 1.5 G_HZ T6-2467	20031 𧈧 — 1.7 G_HZ	20042 𧈧 — 1.10 V2-6E26	20054 𧈧 — 1.14 V2-6F21
2000F 𧈧 — 1.4 TF-213E	20021 𧈧 𧈧 — 1.6 G_KX007 T6-255F	20032 𧈧 — 1.7 V0-305F	20043 𧈧 𧈧 — 1.10 G_HZ T5-3323	20055 𧈧 𧈧 — 1.14 G_KX0078 T7-2121
20010 𧈧 — 1.4 TF-213F	𧈧 H-9C71	20033 𧈧 — 1.7 V2-6E25	20044 𧈧 — 1.10 V0-354B	𧈧 KP1-341F