INCITS/L2/04- 161R
WG2 N2807
Date: June 17, 2004

Title:	HKSCS and GB 18030 PUA characters, background document
Source:	UTC/US
Authors:	Michel Suignard, Eric Muller, John Jenkins
Action:	For consideration by UTC and IRG

Summary
This documents describes characters still encoded in the Private Use Area of ISO/IEC 10646/Unicode as commonly found in the mapping information for Chinese coded characters such as HKSCS and GB-18030. It describes new encoding proposal to eliminate these Private Use Area allocation, so that the PUA can really be used for its true purpose. Doing so would tremendously improve interoperability between the East Asian market platforms because support for Government related encoded repertoire would not interfere with local comprehensive usage of the PUA area.

Hong Kong Supplementary Character Set (HKSCS)

According to http://www.info.gov.hk/digital21/eng/hkscs/download/big5-iso.txt there are a large number of HKSCS-2001 characters still encoded in the Private Use Area (PUA). A large majority of these characters looks like CJK Basic stroke that could be used to describe the appearance of CJK characters. Although there are already collections of various CJK fragments (such as CJK Radicals Supplement, Kangxi Radical) and methods to describe their arrangement using the Ideographic Description Characters, these 'stroke' elements stands on their own merit as an interesting mechanism to describe CJK characters and corresponding glyphs.

Most of these characters have been proposed for encoding on the CJK Extension C. However that extension is not yet mature, but at the same time removing characters from the PUA is urgent. A better solution seems to create a new block containing these Basic CJK unified Stroke characters and to fill the current CJK Unified Ideographs blocks with the other characters that cannot be considered as BASIC stroke elements but should just be encoded as regular CJK Unified Ideographs.

This covers the large majority of the HKSCS PUA characters. The remaining cases are as following:

- Ê̄ (HKSCS code 8862), Ê̌ (HKSCS code 8864), ê̄ (HKSCS code 88A3) and ê̌ (HKSCS code 88A5) can already be encoded as sequence of base characters followed by combining characters. Therefore they don't need a new encoding. Furthermore, should they be encoded, they could never be used in any process involving character normalization. For example they would never be usable as part of an International Domain Name (IDN). Instead, the four composite sequences identified by the UCS Sequence identifiers should be formally identified in ISO 10646 as such.
- ⺶ (HKSCS code 88A9: ground symbol) and ⺘ (HKSCS code 88AA: fuse symbol) should be encoded as addition to the Miscellaneous Technical block.
- ⺾ (HKSCS code C87E) looks very similar to the CJK Unified Ideograph U+4491 ⺾ which is a simplified view of the grass radical encoded itself as U+2F8B 艸. There are many grass radicals already encoded in the standard, either as part of the CJK Unified set or in other parts:

    1) U+2F8B 艸 Kangxi Radical Grass
    2) U+2EBE ⺾ CJK Radical Grass One
    3) U+2EBF ⺿ CJK Radical Grass Two

4) U+2EC0 ⻀ CJK Radical Grass Three
5) U+8278 艸 CJK Unified Ideograph
6) U+8279 艹 CJK Unified Ideograph
7) U+4491 ⺾ CJK Unified Ideograph
8) U+FA5D ⻀ CJK Compatibility Ideograph
9) U+FA5E 艹CJK Compatibility Ideograph

Fonts used for HKSCS such as “MingLiu_HKSCS” do not seem to contain a character for U+4491. Therefore it seems appropriate to encode that private use character as U+4491, unless the intended usage is a CJK radical which would have different character properties. However, encoding a fourth variant of the CJK Radical Grass seems unwise unless strong evidence is provided.

The following table shows the following information:

- HKSCS Code
- UCS PUA Code
- Representative glyphs as in MingLiu_HKSCS
- Proposed Code (xx represents the block position that would be used for the proposed Unified CJK Basic Stroke, all proposed code values are new characters except for U+4491 and composite sequences)
- Comment as appropriate

(The shaded area in the glyph cells shows the character bounding box, underlined proposed code shows addition)

HKSCS Code	PUA Code	Glyph	Proposed Code	Comment
8840	F303		31C0	Proposed for CJK Unified Ext C-0001
8841	F304		31C1	Proposed for CJK Unified Ext C-0518
8842	F305		31C2	Proposed for CJK Unified Ext C-0519
8843	F306		9FB2	Proposed for CJK Unified Ext C-0520, not encoded as a stroke character because it is a variant of HKSCS 8843
8844	F307		31C3	
8846	F309		31C4	Proposed for CJK Unified Ext C-0400
8849	F30C		31C5	HKSCS C87A already maps to similarly shaped U+200CC ⺄
884A	F30D		31C6	Proposed for CJK Unified Ext C-0402
884D	F310		31C7	Proposed for CJK Unified Ext C-0399
884F	F312		31C8	Proposed for CJK Unified Ext C-0403
8850	F313		31C9	Proposed for CJK Unified Ext C-0404
8851	F314		31CA	Proposed for CJK Unified Ext C-0405
8852	F315		31CB	Proposed for CJK Unified Ext C-0406
8854	F317		31CC	Proposed for CJK Unified Ext C-0407
8855	F318		31CD	Proposed for CJK Unified Ext C-0408
8862	F325		<00CA,0304>	Already encoded as composite sequence

8864	F327	Ế	<00CA,030C>	Already encoded as composite sequence
88A3	F344	ế	<00EA,0304>	Already encoded as composite sequence
88A5	F346	ế	<00EA,030C>	Already encoded as composite sequence
88A9	F34A	𠀋	23DA	Miscellaneous Technical Block
88AA	F34B	𠀋	23DB	Miscellaneous Technical Block
8C43	F57A	熚	9FA6	Proposed for CJK Unified Ext C-13160
8C6D	F5A4	曙	9FA7	Proposed for CJK Unified Ext C-10543
8C74	F5AB	匪	9FA8	Proposed for CJK Unified Ext C-3169
8CB7	F5CC	薪	9FA9	Proposed for CJK Unified Ext C-18342
8CB9	F5CE	赵	9FAA	Proposed for CJK Unified Ext C-19996
8CBB	F5D0	幹	9FAB	Proposed for CJK Unified Ext C-20303
8CC0	F5D5	鋼	9FAC	Proposed for CJK Unified Ext C-21171
8CD7	F5EC	騈	9FAD	Proposed for CJK Unified Ext C-22927
8CD8	F5ED	驤	9FAE	Proposed for CJK Unified Ext C-22952
8CDA	F5EF	鈸	9FAF	Proposed for CJK Unified Ext C-21016
C879	F7E5	𠂉	31CE	Proposed for CJK Unified Ext C-0211
C87E	F7EA	𦍌	4491	Already encoded as CJK Unified Ideograph
C8A1	F7EB	𠂤	9FB0	Proposed for CJK Unified Ext C-3309
C8A3	F7ED	冈	9FB1	Proposed for CJK Unified Ext C-5373

Hong Kong HKSCS extension

In addition to the current version of the HKSCS, the Hong Kong Government has a process of character 'approval' by which new characters get added to a list for future inclusion in HKSCS. If the characters do no exist yet in ISO/IEC 10646, they are supposed to be submitted to the appropriate channel, for more details see http://www.info.gov.hk/digital21/eng/hkscs/applicn.html. Unfortunately in the meantime, such characters are allocated to the PUA. The current list at http://www.info.gov.hk/digital21/eng/hkscs/download/newchar.pdf contains three such characters.

Of those three characters, one seems to be already encoded in CJK Unified Extension B block as U+23551 𣕑. The following table shows the following information:

- HKSCS Code
- UCS PUA Code
- Representative glyphs as in the HK web site.
- Proposed Code
- Comment as appropriate

(Underlined proposed code shows addition)

HKSCS Code	PUA Code	Glyph	Proposed Code	Comment
8CEB	F600		23551	Serial Number 16, already encoded as U+23551 𣲗
8CED	F602		9FB3	Serial Number 17
8D48	F61C		9FB4	Serial Number 44

It should be noted that the character coded in 23551 𣲗 is slightly different. It could be argued that they should be dis-unified.

GB-18030 Coded Character Set

GB-18030 also contains some Private Use Area characters and many of these characters can be categorized along the same criteria as the HKSCS set: CJK Basic Stroke, CJK Unified Ideographs, Latin letter and various symbols. Symbols and Latin letter can be categorized as follows:

- Vertical Variants (GB Codes A6D9-A6DF, A6EC-A6ED, A6F3). Should they be encoded, they should become part of a supplement block to the CJK Compatibility Block U+FE30-4F which is fully populated. The new block could be called 'Vertical Forms' and use the range U+FE10-U+FE1F.
- ḿ (GB Code A8BC) is a Latin character already encoded in U+1E3F.

The following table shows the following information:

- GB-18030 Code
- UCS PUA Code
- Representative glyphs as in SimSun-18030
- Proposed Code (xx represents the block position that would be used for the proposed Unified CJK Basic Stroke, yyy represents the block position that would be used for a possible CJK Compatibility Forms Supplement)
- Comment as appropriate

(The shaded area in the glyph cells shows the character bounding box, underlined proposed code shows addition)

GB18030 Code	PUA Code	Glyph	Proposed Code	Comment
A6D9	E78D	﹐	FE10	
A6DA	E78E	﹒	FE12	Vertical variant of U+3002 。
A6DB	E78F	﹑	FE11	Vertical variant of U+3001 、
A6DC	E790	﹕	FE13	
A6DD	E791	﹔	FE14	
A6DE	E792	﹖	FE15	
A6DF	E793	﹗	FE16	
A6EC	E794	﹇	FE17	Vertical variant of U+3016 〖

A6ED	E795	﹀	FE18	Vertical variant of U+3017 〗
A6F3	E796	⋮	FE19	Vertical ellipsis, but can't use U+22EE as it does not decompose in <vertical>U+2026
A8BC	E7C7	ḿ	1E3F	
FE51	E816	𠂇	20087	
FE52	E817	𠂉	20089	
FE53	E818	𠃌	200CC	
FE59	E81E	𠂢	9FB5	
FE61	E826	𢆶	9FB6	
FE66	E82B	𢌞	9FB7	
FE67	E82C	𢒼	9FB8	Proposed for CJK Unified Ext C-0537
FE6C	E831	𣲘	215D7	
FE6D	E832	𠔃	9FB9	
FE76	E83B	𢦏	2298F	
FE7E	E843	𠚕	9FBA	Similar to U+20509 尖 but as high half component
FE90	E854	𠂋	9FBB	Similar to U+2099D 卓 but as left half component
FE91	E855	𤇾	241FE	
FEA0	E864	𤔔	9FBC	Similar to U+470C 鱻 but as high half component

Conclusion

Addressing the PUA issues as suggested for these two important coded character set (HKSCS and GB-18030) would go a long way in facilitating data interchange across East Asia. It would also be extremely beneficial to bring the requirements for any new characters similar to these to the attention of the Unicode Technical Committee and ISO SC2/WG2 before taking the decision to encode them in the Private Use Area.