The Hong Kong Supplementary Character Set(HKSCS) and Mingration to ISO / IEC 10646

Qin Lu

The Hong Kong Polytechnic University

Outline

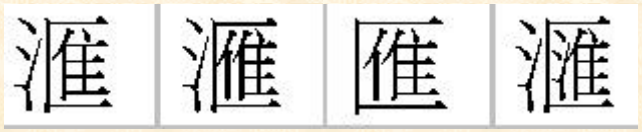
- Introduction
- Collection & Coding Allocations
- Mappings into ISO/IEC 10646
- Extension of HKSCS

- HK is a bilingual society
- Majority use Big-5 based systems with 13,000 Chinese characters in traditional form
- Lack of support for some Cantonese/HK unique characters
- Examples: (From GCCS)
  - Personal names: 滙 (FAC0), 娚(FBFB),
  - simplified Chinese: 沟 (9076), 笔 (9FE5)
  - Cantonese characters: 嗰 (9DF5), 喺 (9DF6)
  - Variants: 靜 (90DC), 羣 (8EC4)
  - Foreign characters: 鮀 (9DCD)

Government Common Character Set (GCCS)

- First appeared in Govern. Tender doc. late 1995
- 3,049 characters defined in User-Defined Areas (UDCs)
- Intended for Govern. internal use
- Sources: Various Government Departments
- Made available to public in 1997 for download with font and the Changjie input method
- Marked the first attempt by HK Govern. for “standardization”

GCCS continued

- Problems with GCCS
  - Not truly exchangeable
  - Lack of criteria for inclusion
  - Inclusion of “incorrect” characters:
  - Example: 滙 
- Digital 21(Nov. 1998): HKSARG IT strategy:
  - Open and Common Chinese Language Interface
  - Adoption of ISO/IEC 10646
    - Superset of Big-5
    - Evolving standard and possible to include GCCS and future extensions

1st Extension of GCCS

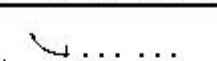
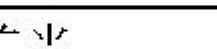
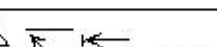
- Additional 3,000 some candidate characters by May 99 collected by the Official Language Agency(OLA)
- Limited code space in Big-5
- Need for inclusion criteria and the removal of “incorrect” characters(characters without clear source)
- Establishment of the Chinese Language Interface Advisory Committee(May, 99)
- Published in September 28, 1999
- Renamed:
  - **Hong Kong Supplementary Character Set**(HKSCS)

Hong Kong Supplementary Character Set (HKSCS)

- 4,702 character:
  - 2,943 from GCCS(106 from GCCS removed)
  - 1,759 newly included
- Chinese characters: 4,261

Character Class	Total	Samples
Cantonese characters	109	啞𠺢峨瘺……
Characters found in major dictionaries	3,114	癐擗䶞䬳……
Scientific names	12	檘鉎鈹聕……
Personal names	890	覍巽㡘弢……
Unified Extension B characters	99	傍荎僂偮……
Radicals and variants	37	⺮ 月 阝 ⻗……
Total	4,261	

- Special Symbols

Character Class	Total	Samples
Pinyin Symbols	49	Ā Ǎ Ă À … ···
International phonetic symbols	10	∫ ɐ ɛ ɔ ɵ œ ø ŋ ʊ I
Stroke symbols	22	
Hiragana	83	
Katakana	86	
Components	33	
Numerals	30	①②. . . . . vi vii. . . . .
Misc symbols	26	
Graphic symbols	34	
Electrical symbols	2	
RUSSIAN	66	А Б В Г · · · · ·
Total	441	

VDA 1																	
Row C6	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
A0		①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩	(1)	(2)	(3)	(4)	(5)
B0	(6)	(7)	(8)	(9)	(10)	i	ii	iii	iv	v	vi	vii	viii	ix	x	丶
C0	丿	亅	一	冂	一	冫	勹	匚	冂	厶	夂	亠	巛	幺	广	
D0	ヨ	彡	攴		疒		辶		¨	ˆ	丶	゛	ゝ	ゞ		
E0	々	〆	〇	ー	[	]	*	ぁ	あ	ぃ	い	ぅ	う	ぇ	え	お
F0	お	か	が	き	ぎ	く	ぐ	け	げ	こ	ご	さ	ざ	し	じ	

Row C7	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
40	す	ず	せ	ぜ	そ	ぞ	た	だ	ち	ぢ	っ	つ	づ	て	で	と
50	ど	な	に	ぬ	ね	の	は	ば	ぱ	ひ	び	ぴ	ふ	ぶ	ぷ	へ
60	べ	ぺ	ほ	ぼ	ぽ	ま	み	む	め	も	ゃ	や	ゅ	ゆ	ょ	よ
70	ら	り	る	れ	ろ	ゎ	わ	ゐ	ゑ	を	ん	ァ	ア	ィ	イ	
A0		ゥ	ウ	ェ	エ	ォ	オ	カ	ガ	キ	ギ	ク	グ	ケ	ゲ	コ
B0	ゴ	サ	ザ	シ	ジ	ス	ズ	セ	ゼ	ソ	ゾ	タ	ダ	チ	ヂ	ッ
C0	ツ	ヅ	テ	デ	ト	ド	ナ	ニ	ヌ	ネ	ノ	ハ	バ	パ	ヒ	ビ
D0	ピ	フ	ブ	プ	ヘ	ベ	ペ	ホ	ボ	ポ	マ	ミ	ム	メ	モ	ャ
E0	ヤ	ュ	ユ	ョ	ヨ	ラ	リ	ル	レ	ロ	ヮ	ワ	ヰ	ヱ	ヲ	ン
F0	ヴ	ヵ	ヶ	А	Б	В	Г	Д	Е	Ё	Ж	З	И	Й	К	

Row C8	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
40	Л	М	Н	О	П	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ
50	Ы	Ь	Э	Ю	Я	а	б	в	г	д	е	ё	ж	з	и	й
60	к	л	м	н	о	п	р	с	т	у	ф	х	ц	ч	ш	щ
70	ъ	ы	ь	э	ю	я	⇧	↖	⇆	㇏	㇆	㇗	㇇	⺉	[illegible]	
A0		[illegible]	冈	図	[illegible]											
B0																
C0														㇕	¦	'
D0	”	㈱	№	℡	゛	゜	⺀	⺄	⺆	⺇	⺈	⺊	⺌	⺍	⺕	⺲
E0	⺝	⺥	⺧	⺪	⺭	⺮	⺶	⺼	⺾	⻆	⻊	⻌	⻍	⻏	⻖	⻗
F0	⻞	⻣				ʃ	ɐ	ɛ	ɔ	ɵ	œ	ø	ŋ	ʊ	ɪ	

VDA 2																
Row F9	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
D0							碁	銹	裏	墻	恒	粧	嫺	╔	╦	╗
E0	╠	╬	╣	╚	╩	╝	╒	╤	╕	╞	╪	╡	╘	╧	╛	╓
F0	╥	╖	╟	╫	╢	╙	╨	╜	║	═	╭	╮	╰	╯	■	

- UDA3

Row 88	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
40	㇀	㇁	㇂	㇃	㇄	㇅	㇆	㇇	㇈	㇉	㇊	㇋	㇌	㇍	㇎	㇏
50	㇐	㇑	㇒	㇓	㇔	㇕	Ā	Á	Ǎ	À	Ē	É	Ě	È	Ō	Ó
60	Ǒ	Ò	Ē̄	Ḗ	Ē̌	Ḕ	Ê	ā	á	ǎ	à	a	ē	é	ě	è
70	ī	í	ǐ	ì	ō	ó	ǒ	ò	ū	ú	ǔ	ù	ǖ	ű	ǔ	
A0		ǜ	ü	ē̄	ḗ	ě	ề	ê	g	㇟	㇠					

Repertoire Selection Principles

- Exclusion Principles:
  - Characters already defined in Big-5
  - Variants of character(s) defined in Big-5 that can be unified(using the ISO/IEC 10646 unification rules):84

removed	弌	蠏	瑨	牐	煢	包	嘅	婷	开	覶	涅
remained	弌	蠏	瑨	牐	煢	包	嘅	婷	幵	覶	涅

  - Characters whose source information and usage cannot be verified : 22

9EAC	獾	9FAD	嗯	9FDA	膺	A054	䘥	A072	蚟	A0D3	亂
9EC4	髒	9FB1	[illegible]	9FE6	瀋	A057	熵	A0A5	嗽	A0E1	憔
9EF4	跭	9FC0	薛	9FEA	罸	A05A	悉	A0AD	磬		
9F4E	纒	9FC8	丞	9FEF	[illegible]	A062	癥	A0AF	糃		

Big-5 Coding Ranges

Range	Total	Name of Block (Total code points)
8140 – 8DFE	2,041	User-Defined Area 3 (UDA3)
8E40 – A0FE	2,983	User-Defined Area 2 (UDA2)
A140 – A3FE	471	Big-5 Symbols and Control Codes
A440 – C67E	4,501	Big-5 Primary Character Set
C6A1 – C8FE	408	Vendor-Defined Area (VDA1)
C940 – F9D5	7,652	Big-5 Secondary Character Set
F9D6 – F9FE	41	Vendor-Defined Area (VDA2)
FA40 – FEFE	785	User-Defined Area 1 (UDA1)

HKSCS Code Allocation in Big-5

- UDA 1 (FA40 – FEFE) : 763 Characters
- UDA 2 (8E40 – A0FE) : 2,898 Characters
- UDA 3 (8140 – 8DFE) : 641 Characters
- VDA 1 (C6A1 – C8FE) : 359 Characters
- VDA 2 (F9D6 – F9FE) : 41 Characters

- Future extension in UDA 3

Range (Total code points)	Sub-blocks (Total code points)	Purpose
User-Defined Area 3 (UDA3) 8140 – 8DFE (2,041 code points)	8140 – 84FE (628 code points)	Will not be used by HKSCS nor for future extensions of HKSCS.
	8540 – 8DFE (1,413 code points)	Reserved for HKSCS. Currently, 641 characters are defined.

Compatibility points

- Introduced to provide full backward compatibility to GCCS
- Principles:
  - Code points for removed characters are reserved
  - No new assignment of these compatibility points
  - Flexible implementation :
    - Font can be provided
    - Input methods can be disabled

Row FA		0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
40		冇	錍	凂	漌	蕌	紈	珯	况	𡋾	琊	錞	鍄	袕	菈	洸	砼
50		杆	拟	玔	鉚	僗	莶	墬	倰	後	倈	蒀	諳	槮	璆	倴	倩
60		偃	徤	傑	傢	傜	僚	僞	僬	儁	儀	顲	孫	莧	現	兒	兆

HKSCS in Unicode Scheme

- Mappings to both Unicode 2.0 and Unicode 3.0

FA	0	1	2	3	4	5	6	7	8
40	𠕇	𠕋	𠘨	𠸘	𡃤	𡉏	𡋾	𡌛	𡏾
U-2	E000	92DB	E002	E003	854C	E005	73EF	51B5	E008
U-3	E000	92DB	E002	E003	854C	E005	73EF	51B5	3649

- Only some characters are mapped into Private Use Area of Unicode
- Use of compatibility points in PUA
- Converting functions in existing systems

Extension of HKSCS

- Will be handled by CLIAC
- Public consultation paper out Friday 24 March, 2000
- 3 parts: Exclusion rules, Inclusion rules, Procedures for submission and review
- Exclusion rules:
  - Check against Big-5 repertoire
  - Follow ISO/IEC 10646 unification rules
  - No simplified Chinese in principle

    Exceptions: 葉 vs. 叶

- Inclusion Rules:

Characters used “commonly” in HK

  - Characters in use (in printed materials) already a place, etc:

    埗 (96F5), 糱 (8E78) vs 粽 ,

  - Cantonese characters(may be newly created)

    嗰(9DF5), 喺(9DF6)

  - Characters used in personal names, building names, etc, which can be verified in major dictionary:

    塬 (9254), 邨 (9068) vs 村

  - Non-regional names, new materials, names, etc
  - Special symbols

- Procedures:
  - Separate submissions:
    - Govern agencies: requires timely reply(in a matter of days)
    - individuals: scholarly, news papers,
  - Around 3 months for review, and available in internet
  - Publish at most once a year and stop after Extension B of ISO/IEC 10646 is published.

Conclusion

- HKSCS is the first standard in HK
- Government is playing more roles in standardization
- More efforts/resources will be allocated to Unicode migration related issues
- Encourage vendors to make systems that are Unicode enabled
- http://www.digital21.gov.hk/chi/hkscs/download