Hong Kong Character Set - 2015

Version ~~1~~2.0

Sept.~~Aug.~~ 2015

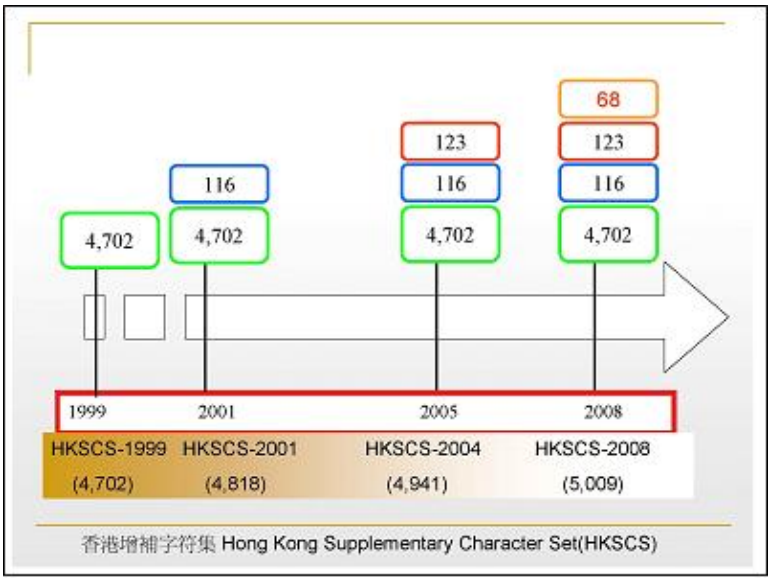
Table of Contents

Preface

Information stored in a computer or transmitted in electronic communication is coded according to a pre-defined coding scheme. For information in Chinese, there are different coding schemes including Big-5, GB (Guo Biao) and the ISO/IEC 10646 international coding standard. When coding schemes do not cover all the Chinese characters used in the Hong Kong Special Administrative Region (HKSAR), users may need to create unencoded characters on their computers and assign internal codes for them in the user-defined area. Among these characters, many are commonly used in the HKSAR. Some of them are used in the names of persons and places while some in the Cantonese dialect. This works well in stand-alone computers, but when computers are connected to each other, these user-defined characters may give rise to problems in communication and data exchange.

One of the initiatives under the Government's "Digital 21" Strategy for IT development is to develop an open and common Chinese language interface in the HKSAR for users who choose to communicate electronically in Chinese. The pivotal elements of this initiative include: (1) building the Hong Kong Supplementary Character Set (HKSCS) as a standard for the Chinese characters used in Hong Kong so as to set up a common platform; (2) migrating to use of the ISO/IEC 10646 international standard gradually to promote the realization of the common Chinese system. ISO/IEC 10646 is an international coding standard being developed under the aegis of the ISO to embrace characters used in all major languages in the world, including traditional and simplified Chinese characters.

To facilitate electronic communication within the Government, the Hong Kong Government developed the Government Common Character Set (GCCS) in 1995. The release of the GCCS marked the first step in coordinating the adoption of



user-defined Chinese characters and it was well received by the public as a supplement to the standard character set of Big-5. This supplementary character set was later enhanced by the Government in collaboration with the Chinese Language Interface Advisory Committee (CLIAC), which comprises representatives from academia, language and linguistics associations, the information technology industry and the publishing industry. The enhanced character set included characters collected from various sectors in the HKSAR and represented a common set for the community. It was named the Hong Kong Supplementary Character Set (HKSCS) and was published in September 1999. This version has 4,702 characters and is also known as HKSCS-1999 for aligning with the versions published afterwards.

There were two code allocation schemes for HKSCS when it was initially published, one for the Big-5 system used by Hong Kong at that time and the other for the ISO/IEC 10646 international coding standard.
(HKSCS-2008 is the last version of HKSCS published with Big-5 code points.)

Recognising the need of the public and government departments to add new characters into HKSCS from time to time, the Government worked together with the CLIAC and published the procedures and principles for inclusion of new characters in HKSCS in April 2000. The CLIAC meets regularly to consider applications for inclusion of characters in HKSCS. Once approved, the new HKSCS characters will be submitted to the Ideographic Rapporteur Group (IRG), a working group under the International Organisation for Standardisation (ISO), for inclusion in the new releases of the ISO/IEC 10646 standard as vertical extension.

In December 2001, HKSCS-2001 was released with 116 newly approved characters. In May 2005, HKSCS-2004 was published with a further addition of 123 new characters. In April 2008, HKSCS-2008 was published with 68 newly approved characters, making a total of 5,009 characters. HKSCS-2004 technically aligns with the ISO/IEC 10646:2003 and its Amendment 1 published in April 2004 and November 2005 respectively. In October 2009, five previously unencoded characters in HKSCS-2008 were accepted in ISO/IEC 10646:2003 Amendment 6. As such, all the characters in HKSCS-2008 have been assigned code points in the ISO/IEC 10646 standard, marking an important step of Hong Kong's migration from the Big-5 platform to ISO 10646-enabled platforms. In April 2008, the Government promulgated the revised principles for the inclusion of characters in HKSCS. For newly included HKSCS

characters, only ISO/IEC 10646 code points will be assigned and Big-5 code points will cease to be assigned. All HKSCS characters already assigned with Big-5 code points before will not be affected. This decision is in line with the adoption of the common Chinese language interface which comprises the ISO/IEC 10646 standard and HKSCS as recommended by the "Digital 21" Strategy for IT development.

With the acceptance of five previously unencoded HKSCS-2008 characters by the ISO, HKSCS-2008 aligns technically with the ISO/IEC 10646:2003 and its Amendments 1 to 6. In view of this, the CLIAC has decided to further revise the principles for the inclusion of characters in HKSCS such that newly approved characters will be submitted directly to the ISO for approval before formal inclusion in HKSCS. In addition, characters will not be assigned code points in the Private Use Area (PUA). Instead, they will be given ISO/IEC code points directly once approved by the ISO.

However, HKSCS has its limitation. At present, the principle for extending it only supports the addition of new characters to ISO/IEC 10646. In other words, HKSCS only supports Vertical Extension to ISO/IEC 10646. There is no complete character set to define all the characters commonly used in Hong Kong, which makes it confusing to users. Without official definitions of Hong Kong character set, it is also difficult for the industry to support the use of Chinese in electronic communication in Hong Kong.

To support the coding platform of the computer system, ISO/IEC 10646 provides a method named Horizontal Extension to define the Chinese characters used in a certain country or region as a **named character set** [1]. Through this method, we can define and use those accepted Chinese characters and symbols proposed by other countries and regions. This enables vendors to follow the named character set definition to support the common Chinese language interface. In previous Horizontal Extension to ISO/IEC 10646 for the Hong Kong characters, referred to as the H-Column, only those characters in HKSCS are defined. This does not reflect the actual use and cannot completely satisfy the needs of Hong Kong.

To enable vendors to support the Chinese characters used in Hong Kong and the related localization technology, and to help HKSAR to migrate completely from a Big-5-centric

[1] ISO/IEC 10646 allows the defining of named character set to support localization.

coding system to the use of the ISO/IEC 10646 international standard, it is necessary to form a complete collection as a named character set under the ISO/IEC 10646 international coding standard as the **Hong Kong Character Set** (**HKCS**). In this way, the H-Column in ISO/IEC 10646 will be completely defined.

HKCS defined in this file will include (1) all characters and symbols in HKSCS, (2) all characters and symbols in the Big-5 character set (Etian Big-5), and (3) Hong Kong's Horizontal Extension to ISO/IEC 10646. HKCS-2015 consists of 17,654 ideographic characters and 83~~8~~9 symbols, including (1) 4,579 ideographic characters and 430 symbols in HKSCS-2008, (2) 13,053 ideographic characters and 408 symbols in the Big-5 character set (excluding 33 transmission control codes in Big-5), and (3) 22 Hong Kong's Horizontal Extension characters~~ confirmed by CLIAC~~, and 1 newly added symbol, Euro Sign (HD-20AC)[2].

HKCS will be extended if there are future needs in the community. As HKCS has enlarged its scope, the principles for inclusion of new characters in HKSCS need to be revised accordingly. That is, the principles should include not only Vertical Extension but also Horizontal Extension to ISO/IEC 10646. The revision of the principles and the review of new submissions are already accepted as part of CLIAC's work items.

HKCS-2015 can be downloaded from xxxx.

[2] For details, see Section 3.6.

Acknowledgement

To be added

Section 1 Overview

1.1 This document specifies the overall coding architecture and code points of HKCS-2015 in ISO/IEC 10646. The characters and symbols in Big-5 are also coded accordingly. HKCS-2015 is fully compatible with the previous versions of HKSCS and the GCCS.

1.2 HKCS-2015 is a complete named character set under the ISO/IEC 10646 international standard, including (1) all characters and symbols in HKSCS-2008, that is, 5,009 in total, (2) all characters and symbols in Big-5, namely, 13,461 in total, and (3) 22 Hong Kong's Horizontal Extension characters ~~confirmed by CLIAC~~, and 1 newly added symbol, Euro Sign (HD-20AC). A total of 18,49~~2~~3 characters and symbols in HKCS-2015 will be listed in two tables. The first table lists the CJK Unified Ideographs, and the second one lists other symbols, including some radicals and components of Chinese characters.

1.3 HKCS-2015 is a coded character set rather than a glyph standard. For glyph guidelines, please refer to "Glyph Specification for the Chinese Characters in HKCS".

(Note: "Glyph Specification for the Chinese Characters in HKCS" can be found at xxxx.)

1.4 For the purpose of this document, the following definitions will apply:

Term	Definition
Basic Multilingual Plane (BMP, Plane 0)	The first code plane in the ISO/IEC 10646 coding framework (i.e. "Plane 0" or basic plane). Code points are from U+0000 to U+FFFF.
Block	A continuous collection of code points. The assigned characters to the block share some common characteristics.
Character	A member of a set of elements used for the organisation, control or representation of data.
Character Glyph	In ISO/IEC 10646, it refers to a character in its abstract form as an image. It is independent of a specific image. The

Term	Definition
	basic elements to form an ideograph character are strokes, radicals, components and their relative positions.
Character Set	A defined set of characters.
CJK Ideographs Main Block	The first block assigned to the unified ideographs including Chinese, Japanese and Korean. Code points are from U+4E00 to U+9FFF.
CJK Ideograph Source	The CJK ideographs in the ISO/IEC 10646 international standard are defined based on the original computer character standards of China, Japan, Korea and other countries and regions. The original computer character standard or specification is called CJK Ideograph Source. The countries and regions are represented by letters as follows: Mainland China (G), Hong Kong (H), Japan (J), South Korea (K), Singapore (S), Taiwan (T), and Vietnam (V).
CJK Unified Ideographs	A set of ideographs defined in the ISO/IEC 10646 international standard after the process of unification, which is derived from the original character standards of China, Japan, Korea, and other countries and regions. As the first version of the standard includes only characters from China (including Taiwan), Japan and Korea, the name “CJK” has been used ever since.
Code Point	An assigned binary code value to represent a character. To distinguish from other coding standards, the notation of “U+” is used to indicate an ISO/IEC 10646 code point.
Coded Character Set	A character set established using a set of unambiguous rules

Term	Definition
	to indicate the relationship between the characters of the set and their coded representation.
Extension Blocks	All blocks extended after the CJK ideographs main block. Extension A block is placed on the BMP and the subsequent extension blocks are on the Supplementary Ideographic Plane (SIP), Plane 2.
Government Common Character Set (GCCS)	A coded character set developed by the Hong Kong Government in 1995 for Chinese character interchange and processing within the Government.
Horizontal Extension	A method provided by ISO/IEC 10646 to list the characters used by a particular country or region X in the X-Column of the CJK ideographs main block and other extension blocks so as to support the coding platforms of the computer systems. The characters used by Hong Kong are listed in the H-Column.
Ideograph	Refers to a writing system in which the scripts are not primarily used to represent sound, but to represent meaning. Chinese characters are often referred to as ideographs.
ISO/IEC 10646	An ISO standard on a coded character set. It aims at providing one single character set to encompass the characters of all major languages.
ISO/IEC 10646:2003	Published in 2004, this version of the ISO/IEC 10646 standard is a single publication following the merger of the previous two releases of the ISO/IEC 10646 standard: ISO/IEC 10646-1:2000 and its supplement ISO/IEC 10646-2:2001.

Term	Definition
Named Character Set	A character set defined for a particular country, region, or language under ISO/IEC 10646 to support localization. Chinese characters are listed in the X-Column in CJK ideographs blocks.
Private Use Area (PUA)	A block of code points in the ISO/IEC 10646 standard reserved for users and vendors to give their own definitions. In principle, it is intended for use by individuals or vendors privately and not for data interchange.
Source Reference	A reference established by associating a CJK Ideograph code point with one or several values in the source standards from which the CJK Unified Ideographs in ISO/IEC 10646 are derived.
Supplementary Ideographic Plane (SIP, Plane 2)	The ISO/IEC 10646 coding framework assigned Plane 2 for CJK ideograph extensions. Code points are from U+20000 to U+2FFFF.
CJK Unification	The process of assigning one code point to two or more CJK character glyphs which, though images seemingly different, are actually variants representing the same element in data representation. Consequently, only one of the variants is selected as the representative.
Vertical Extension	A method for adding new ideographs to the CJK ideographs main block and other extension blocks. Source references are required when new ideographs are added.

Section 2 Coding Scheme of HKCS

2.1 HKCS, as a complete named character set under the ISO/IEC 10646 international coding standard, covers all approved characters that are commonly used characters in Hong Kong computer systems. HKCS includes (1) all characters and symbols in HKSCS, (2) all characters and symbols in the Big-5 character set (Etian Big-5), (3) Hong Kong's Vertical Extension characters and symbols, and (4) Hong Kong's Horizontal Extension characters and symbols.

2.2 The ISO/IEC 10646 international coding standard collects and encodes ideographs from different sources. The code charts for CJK ideographs list not only the codes in ISO/IEC 10646 but also the source references and glyphs of the ideographs. The sources of HKCS include: (1) HKSCS, (2) the Big-5 character set (Etian Big-5), (3) Hong Kong's Vertical Extension to ISO/IEC 10646 in the future, and (4) Hong Kong's Horizontal Extension to ISO/IEC 10646.

2.3 Under the ISO/IEC 10646 international coding standard, the four subsets of HKCS as source references are coded as follows:

2.3.1 H-XXXX: used to refer to all characters already coded in HKSCS-2008, in which "XXXX" is the corresponding Big-5 code in HKSCS-2008.

2.3.2 HB(Y)[3]-XXXX: used for all characters in the Big-5 character set, in which "Y" (being optional) can take values of "1" or "2" to indicate the character that belongs to either the basic block or the secondary block in Big-5; "XXXX" is the Big-5 code.

2.3.3 HC-XXXX: used for characters to be vertically extended to ISO/IEC 10646 in the future, in which "XXXX" is an HKCS assigned source reference code between 0001 and 9999 in sequence.

2.3.4 HD-XXXX(X): used for all characters as part of Hong Kong's Horizontal Extension to ISO/IEC 10646, in which "XXXX(X)" is the code of the character in the ISO/IEC 10646 international standard.

2.4 When reviewing the glyphs of the characters in HKCS, CLIAC found that some Hong Kong preferred glyphs were different from character glyphs in Big-5. For example,

[3] "( )" is to indicate an optional item. In actual code point expression, "( )" does not appear.

the character glyph in Big-5 is "兌" (Big-5 0xA749; U+514C), while the Hong Kong preferred glyph is "兑". If we change the glyph of the character "兌" (Big-5 0xA749; U+514C) in Big-5 to "兑", it will have the same glyph with another character "兑" (U+5151). As a result, a character glyph will have two U+ codes[4], which is a violation of the ISO/IEC 10646 coding rules. On the other hand, the character "兌" cannot be displayed on the Chinese platforms conforming to Hong Kong's standard. Therefore, CLIAC decided to keep the original glyph of the character "兌" (Big-5 0xA749; U+514C) in Big-5 and propose Horizontal Extension to ISO/IEC 10646 for the character "兑" (U+5151). In other words, such characters in ISO/IEC 10646 which were not proposed by Hong Kong will be included in HKCS. In this way, we are able to specify the Hong Kong preferred glyphs without violating the mapping between Big-5 and ISO/IEC 10646. Table 1 lists the 22 characters which will be horizontally extended to ISO/IEC 10646.

2.5 HKCS-2015 contains 18,49~~2~~3 characters and symbols, including: (1) 17,654 characters, which are listed in Table 2 using their code points in ISO/IEC 10646 and source references; (2) 83~~8~~9 symbols, which are listed in Table 3 using their code points in ISO/IEC 10646 and source references. It should be noted that all characters shown in the tables use the Song style (print style) glyphs and the glyphs follow the guidelines given in "Glyph Specification for the Chinese Characters in HKCS-2015".

2.6 According to ISO/IEC JTC1/SC2 working group, the code points of two HKSCS characters in ISO/IEC 10646 are changed as follows:

Source Reference	Original Code Point	New Code Point
H-9D73	U+4CA4	U+9FD0
H-91B5	U+3D1D	U+2A3ED

[4] The code in ISO/IEC 10646 is referred to as U+ code.

Section 3 Problematic Code Points in Big-5 and Solutions in HKCS

3.1 Two characters with duplicate Big-5 code points

In Big-5, there are two characters with duplicate Big-5 code points, which are mapped to two different ISO/IEC 10646 code points. The two mappings between Big-5 and ISO/IEC 10646 code points are kept in HKCS:

Seq.	HKCS (Big-5)	Character	U+ Code
1	HB-A461	兀	U+5140
	HB-C94A	兀	U+FA0C
2	HB-DCD1	殼	U+55C0
	HB-DDFC	殼	U+FA0D

3.2 Three problematic numerals

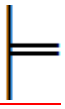
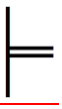
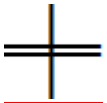
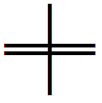
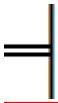
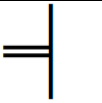
"卄" (Big-5 0xA2CD) is currently mapped to "卄" (U+5344) in CJK Ideographs Main Block. However, in Big-5, it is grouped with other Suzhou Numerals[5]. Another two Big-5 symbols, namely, "  " (Big-5 0xA2CC) and "  " (Big-5 0xA2CE), are unmapped to ISO/IEC 10646 code points.[6] This problem is probably because the three symbols were added to the standard at a later point in time (ISO/IEC 10646-1:2000). In HKCS, the mappings of the three numerals are corrected as follows:

Seq.	HKCS (Big-5)	Symbol	U+ Code
1	HB-A2CC		U+3038 (HANGZHOU NUMERAL TEN)
2	HB-A2CD		U+3039 (HANGZHOU NUMERAL TWENTY)
3	HB-A2CE		U+303A (HANGZHOU NUMERAL THIRTY)

3.3 Eight symbols with 2-1 mappings

[5] They are encoded in the CJK Symbols and Punctuation block in the ranges U+3021..U+3029 and U+3038..U+303A. "HANGZHOU NUMERAL" is used in the names of the code points, which is a misnomer. (See page 369 of ISO/IEC 10646:2014) . Since the names of the code points cannot be changes, "HANGZHOU NUMERAL" is used to refer to particular code points.

There are eight symbols with both Big-5 code point and HKSCS code point but only one ISO/IEC 10646 code point. Therefore, both Big-5 code point and HKSCS code point are mapped to one ISO/IEC 10646 code point. For the conversion from U+ code to Big-5 code, the system should choose one from the Big-5 code point and HKSCS code point. Details are as follows:

U+ Code	Big-5[7]	HKSCS
 U+256D	 HB-A27E	 H-F9FA
 U+256E	 HB-A2A1	 H-F9FB
 U+2570	 HB-A2A2	 H-F9FC
 U+256F	 HB-A2A3	 H-F9FD
U+2550	 HB-A2A4	H-F9F9
 U+255E	 HB-A2A5	 H-F9E9
 U+256A	 HB-A2A6	 H-F9EA
 U+2561	 HB-A2A7	 H-F9EB

The solution of HKCS is as follows:

[6] For details about the mapping between Big-5 and ISO/IEC 10646 code points, see Unihan_IRGSources.txt: http://www.unicode.org/Public/UCD/latest/ucd/Unihan.zip

(1) For conversion from Big-5 code to U+ code, both original Big-5 code point (A2XX) and HKSCS code point (F9XX) are mapped to the same U+ code.

(2) For conversion from U+ code to Big-5 code, four symbols are mapped to the original Big-5 code point (A2XX) and the other four are mapped to HKSCS code points (F9XX). Details are as follows:

U+ Code	HKCS
U+2550	H-F9F9（HKSCS）
U+255E	H-F9E9（HKSCS）
U+2561	H-F9EB（HKSCS）
U+256A	H-F9EA（HKSCS）
U+256D	HB-A27E
U+256E	HB-A2A1
U+256F	HB-A2A3
U+2570	HB-A2A2

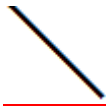
3.4 Five symbols without mapping

In the original Mapping Table[8] of Unicode, five Big-5 symbols are mapped to U+FFFD, which means no mapping. In HKCS, the mappings are as follows:

Big-5	U+ Code
 HB-A15A	U+2574
HB-A1C3	U+FFE3
 HB-A1C5	U+02CD

[7] The images of the symbols in the Big-5 column are from: Computer Chinese Glyph and Character Code Mapping Table, the Industrial Standard of Big-5 in Taiwan, Technical Report C-26, Institute for Information Industry of Taiwan, 1984

[8] See: http://www.unicode.org/Public/MAPPINGS/OBSOLETE/EASTASIA/OTHER/BIG5.TXT

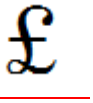
Big-5	U+ Code
 HB-A1FE	 U+FF0F
 HB-A240	 U+FF3C

3.5 12 symbols with problematic mappings

There are 12 symbols with problematic mappings: the mappings in the original Mapping Table[9] of Unicode are different from those of Microsoft or Taiwan CNS11643. In HKCS, they are as follows:

Big-5	U+ Code
 HB-A145	 U+2027
 HB-A14E	 U+FE51
 HB-A156	 U+2013
 HB-A1E3	 U+FF5E
 HB-A1F2	 U+2295
 HB-A1F3	 U+2299

[9] See: http://www.unicode.org/Public/MAPPINGS/OBSOLETE/EASTASIA/OTHER/BIG5.TXT

Big-5	U+ Code
 HB-A241	 U+2215
 HB-A242	 U+FE68
 HB-A244	 U+FFE5
 HB-A246	 U+FFE0
 HB-A247	 U+FFE1
HB-A1C2	U+203E

3.6 Newly added symbol in HKCS

The original Big-5 standard does not include Big-5 0xA3E1 (€ Euro Sign). Microsoft added this symbol in Windows ME which was published in 2000. Since HKSCS does not include this symbol, it is added to HKCS-2015 through Horizontal Extension and encoded with U+20AC.

Section 4 Appendix

Table 1 HKCS Horizontal Extension List

Seq.	Big-5	U+ code	Big-5 Glyph	HK Preferred Glyph	Extension: U+ code
1	0xA769	U+544A	告	告	U+543F
2	0xA749	U+514C	兌	兑	U+5151
3	0xAEAE	U+6085	悅	悦	U+60A6
4	0xD1BE	U+6329	挩	挩	U+635D
5	0xD5A7	U+6553	敓	敓	U+655A
6	0xD5BF	U+68B2	梲	梲	U+68C1
7	0xD258	U+6D97	涗	涗	U+6D9A
8	0xB57C	U+7A05	稅	税	U+7A0E
9	0xB2E6	U+812B	脫	脱	U+8131
10	0xB8C0	U+86FB	蛻	蜕	U+8715
11	0xBBA1	U+8AAA	說	説	U+8AAC
12	0xBE55	U+92B3	銳	鋭	U+92ED
13	0xBE5C	U+95B1	閱	閲	U+95B2
14	0xB6FE	U+5ABC	媼	媪	U+5AAA
15	0xB759	U+614D	慍	愠	U+6120
16	0xBA72	U+6C33	氳	氲	U+6C32
17	0xE2BE	U+7185	熅	煴	U+7174
18	0xEAD5	U+7E15	縕	緼	U+7DFC
19	0xE3A6	U+8183	膃	腽	U+817D
20	0xC4AD	U+860A	蘊	藴	U+85F4
21	0xEEC1	U+8F40	轀	轀	U+8F3C
22	0xC1DF	U+919E	醞	醖	U+9196

Table 2 HKCS-2015 Character Table

The following is an example of a typical cell in Table 2:

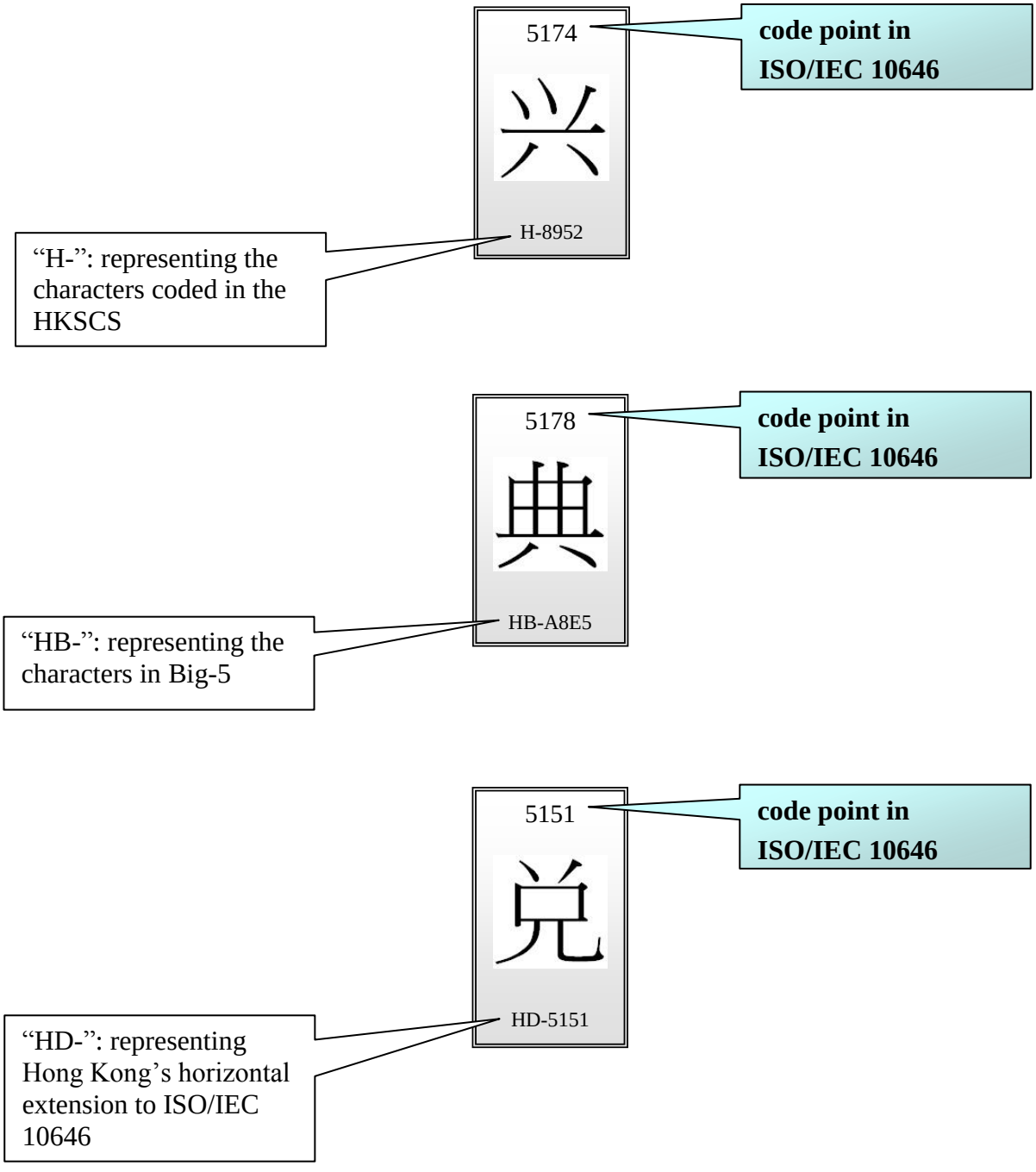


Table 3 HKCS-2015 Symbol Table

The following is an example of a typical cell in Table 3: