Hong Kong Supplementary Character Set - 2015

(Draft 7)

Office of the Government Chief Information Officer &
Official Languages Division, Civil Service Bureau
The Government of the Hong Kong Special Administrative Region

January 2017

Table of Contents

Preface

After the first release of the Hong Kong Supplementary Character Set (HKSCS) in 1999, there have been three updated versions. The HKSCS-2001, HKSCS-2004 and HKSCS-2008 were published with 116, 123 and 68 new characters added respectively. A total of 5 009 characters were included in the HKSCS-2008. These publications formed the foundation for promoting the adoption of the ISO/IEC 10646 international coding standard, and were widely supported and adopted by the IT sector and members of the public.

The ISO/IEC 10646 international coding standard is developed by the International Organization for Standardization (ISO) to provide a common technical basis for the storage and exchange of electronic information. It provides a unified standard for the coding of characters in all major languages in the world including traditional and simplified Chinese characters. Containing more than 80 000 Han characters, the ISO/IEC 10646:2014 provides computer platforms with comprehensive support. However, supporting a character set with over 80 000 Han characters will unnecessarily increase the cost and time of product development. Hence, vendors of font or input method software will select an appropriate number of characters in the light of the requirements of individual countries and regions in developing supporting products.

All the characters in the HKSCS-2015 have already been included in the ISO/IEC 10646 international coding standard. ISO/IEC 10646 will continue adding new characters that have not been included in its character set but are used in Hong Kong (i.e. the addition of new characters through vertical extension). Furthermore, for characters already included in the ISO/IEC 10646, information about the characters and glyphs commonly used in Hong Kong will be added to reflect the actual use of these characters locally (i.e. the addition of information to existing characters through horizontal extension). This will not only facilitate the development of vendor support for Chinese characters actually used in Hong Kong and the relevant localised technology, but will also reduce the time and cost of development, enabling the IT industry to develop more products suitable for Hong Kong.

Compared with the HKSCS-2008, the HKSCS-2015 has 23 more Chinese characters and one more symbol, further fulfilling the needs of local electronic communication in Chinese.

Coding Scheme as the Foundation of Electronic Communication

Information stored in a computer or transmitted in electronic communication is coded according to a pre-defined coding scheme. For information in Chinese, as early as the 1990s, there were different coding schemes including Big-5, GB (Guo Biao) and the ISO/IEC 10646 international coding standard. At that time, as the coding schemes did not cover all the Chinese characters commonly used in Hong Kong, users might need to create unencoded characters on their computers and assign internal codes for them in the user-defined area. Some characters were used in the names of persons and places while some in the Cantonese dialect. This worked well in stand-alone computers, but when computers were connected to each other, such user-defined characters might give rise to problems in communication and data exchange.

Common Chinese Language Interface

Under the Digital 21 Strategy for IT Development, the Government has developed an open and common Chinese language interface for the Hong Kong Special Administrative Region (HKSAR), with the objective of achieving more accurate electronic communication in Chinese. A pivotal element of the open and common Chinese language interface is the adoption of the ISO/IEC 10646 international coding standard.

Development of the HKSCS

To facilitate electronic communication within the Government, the then “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 common character set was later enhanced by the Hong Kong Special Administrative Region Government (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 had 4 702 characters and was also known as the HKSCS-1999 for aligning with the versions published afterwards.

The Government recognised the need for the public and government departments to include new characters in the HKSCS from time to time. In collaboration with CLIAC, the Government published the procedures and principles for the inclusion of characters in the HKSCS in April 2000. CLIAC meets regularly to consider applications for inclusion of characters in the HKSCS. Once approved, the new HKSCS characters will be submitted to the Ideographic Rapporteur Group, a working group under the ISO, for inclusion in the future releases of the ISO/IEC 10646 standard.

The HKSCS has subsequently been updated three times and a total of 5 009 characters were included in the HKSCS-2008. There were two code allocation schemes for each of the four versions of the HKSCS: one for the Big-5 which was used in Hong Kong at that time and the other for ISO/IEC 10646. This arrangement aimed at a gradual migration of computing platforms from Big-5 to ISO/IEC 10646.

With the increased popularity of platforms and products supporting ISO/IEC 10646, the Government promulgated the revised procedures and principles for the inclusion of characters in the HKSCS in April 2008. Since then, only code points of the characters in ISO/IEC 10646 will be provided and no Big-5 code points will be assigned for newly included HKSCS characters. Furthermore, the ISO/IEC 10646:2011 has included all characters in the HKSCS-2008, indicating a complete integration of HKSCS-2008 with ISO/IEC 10646 and marking a milestone for the full adoption of the ISO/IEC 10646.

In order to reflect the local use of Han characters in the ISO/IEC 10646, the CLIAC revised the procedures and principles for the inclusion of characters in the HKSCS in December 2015 and started to prepare for the HKSCS-2015.

The coding scheme and code table of the HKSCS-2015 are provided in this document. Other mapping tables of the HKSCS-2015, and the documents stipulating the procedures and principles for the inclusion of characters in the HKSCS can be found at the Common Chinese Language Interface website at:
http://www.ogcio.gov.hk/en/business/tech_promotion/ccli/hkscs/.

Acknowledgement

This specification was completed with invaluable assistance from CLIAC. We would like to thank the following committee members for their contribution: current members Ms. Michelle CHEUNG, Mr. CHEUNG To, Mr. Ricky CHUN, Dr. FUNG Suk Yee Roxana, Mr. Alan KAN, Dr. LEE Ka Kui, Miss LEUNG Wai Chu Judy, Professor LU Qin, M.H., Dr. MAN Ying Ha, Dr. TANG Pui Ling, Mr. Wilson WONG, and former committee members: Mr. CHIU Leung Fai Fritz, Dr. FONG Kin Kiu Ken, Mr. Raymond HUNG, Mr. LEUNG Shung Yu and Mr. Jonathan SHEA.

We would also like to thank the working group chaired by Professor LU Qin and Ms. XIONG Dan and Mr. Roy YU of the Professor LU's project team of the Polytechnic University for their great efforts. The working group consists of some CLIAC members and Dr. CHEUNG Kwan Hin, Mr. LAI Tat Kiu Alex, Mr. LEE Kin Hong, Dr. LUN Suen Caesar, Professor WONG Yiu Kwan and representative from the Education Bureau.

Section 1 Overview

1.1 This document provides the characters in the HKSCS-2015 and their corresponding code points in ISO/IEC 10646, and explains the overall coding architecture of the HKSCS-2015 in the ISO/IEC 10646 international coding standard. The HKSCS-2015 is fully compatible with the GCCS and the previous versions of the HKSCS.

1.2 The HKSCS-2015 contains 5 033 characters, including 5 009 characters from the HKSCS-2008, and the newly added 23 Chinese characters and one symbol.

1.3 The HKSCS-2015 is a coded character set. It is not meant to be a glyph standard. For glyph guidelines, please consult the "Reference Glyphs for Chinese Computer Systems in Hong Kong", which is available at: http://www.ogcio.gov.hk/en/business/tech_promotion/ccli/download_area/.

1.4 For the purpose of this document, the following definitions of terms 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 0000 to FFFF.
Block	A collection of characters that share 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 Han character in its abstract form as an image. It is independent of any specific image. The basic elements to form an ideograph are strokes, radicals, components and their relative positions.

Term	Definition
Character Set	A defined set of characters.
CJK Compatibility Ideographs	An area defined in the BMP (Plane 0) for compatibility with CJK Ideographs Blocks. This area is used to include the variants or duplicate characters already coded in CJK Ideograph Sources which would otherwise not be coded in ISO/IEC 10646. Code points are from F900 to FAFF. In ISO/IEC 10646, these variants and their corresponding standard characters are unified. However, they are assigned different code points in their respective CJK Ideograph Sources already. Therefore, this special area is defined to avoid having one character with multiple code points in CJK Ideographs Blocks and at the same time allow round-trip conversion for backward compatibility. Every Compatibility Ideograph has a corresponding standard character coded in CJK Ideographs Blocks.
CJK Compatibility Ideographs Supplement	An extended area defined in the Supplementary Ideographic Plane (SIP, Plane 2) for compatibility with CJK Ideographs Blocks. Code points are from 2F800 to 2FA1F. The Han characters in the compatibility blocks and Ideographs Supplement are collectively referred to as “compatibility characters” in this document.
CJK Ideographs Main Block	The first block assigned to the unified ideographs, including Chinese, Japanese and Korean characters. Code points are from 4E00 to 9FFF.
CJK Ideograph Source	The CJK ideographs in the ISO/IEC 10646 international coding standard are defined based on the original computer character standards of China, Japan, Korea and other countries and regions. The original computer character

Term	Definition
	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 coding standard in accordance with the unification rules. The ideographs are derived from the original character standards of China, Japan, Korea, and other countries and regions. As the first version of the standard includes ideographs mainly from China, Japan and Korea, the name “CJK” has been used ever since. In this document, these ideographs are also referred to as “Han characters”.
Code Point	An assigned hexadecimal code value to represent a character.
Coded Character Set	A character set established under a set of unambiguous rules. It defines the relationship between the characters of the set and their coded representation.
CJK Ideographs Extension Blocks	The blocks developed as extensions to 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 then “Hong Kong Government” in 1995 for exchanging and processing Chinese information within the Government.

Term	Definition
Horizontal Extension	This refers to the addition of information and source reference to the characters already included in the ISO/IEC 10646.
H-Column	Each code point of the CJK Unified ideographs has multiple glyphs and these glyphs are listed in individual columns. This multi-column format aims to support and define the characters used in a particular country or region. The Chinese characters used in Hong Kong are listed in the H-column.
Ideograph	Refers to a character in a writing system in which the scripts are not primarily used to represent sound, but to represent meaning. Chinese characters are 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.
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)	Plane 2 is assigned under the ISO/IEC 10646 coding framework for CJK ideograph extensions. Code points are from 20000 to 2FFFF.
Unification	The process of assigning one code point to two or more character glyphs which, though seemingly different, are actually variants representing the same element in data representation. Consequently, only one of the variants is

Term	Definition
	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 the HKSCS-2015

2.1 The HKSCS-2015 consists of 5 033 characters, including 4 602 Chinese characters and 431 symbols. All these characters have been reviewed for use on computer platforms. Unlike the HKSCS-2008, the HKSCS-2015 provides code points of the characters in the ISO/IEC 10646 only. The code table can be found at Table 1.

2.2 The HKSCS-2015 contains all the 4 579 Chinese characters and 430 symbols from the HKSCS-2008, and the newly added 23 Chinese characters and one symbol already included in the ISO/IEC 10646, so as to reflect the actual use of these characters in the HKSAR. These 23 Chinese characters are included in the CJK Ideographs Main Block and the symbol is in the block for currency symbols. The newly included characters are listed in Table 2.

2.3 According to the resolution made by the working group under the ISO/IEC JTC1/SC2, two characters from the HKSCS-2008 should have their code points in the ISO/IEC 10646 re-assigned as follows:

Glyph (source reference)	HKSCS-2008	HKSCS-2015
鰂 (H-9D73)	4CA4	9FD0
黎 (H-91B5)	3D1D	2A3ED

The new code point 9FD0 has been adopted in the ISO/IEC 10646:2014/Amendment 2:2016.

Code points 4CA4 and 3D1D are kept as compatibility points to enable computer systems yet to adopt the latest version of the ISO/IEC 10646 to continue using them.

2.4 The table below shows the relationship between the HKSCS and the character blocks of the ISO/IEC 10646 coding standard. The names of the blocks given are for ease of reference only and may not be the same as those used in the ISO/IEC 10646 international coding standard document.

ISO/IEC 10646 Character block	Number of characters in the HKSCS-2008	Number of characters in the HKSCS-2015
Symbols	430	431
CJK Ideographs Main Block	2 291	2 315
Extension A	574	572
Extension B	1 701	1 702
Extension C	1	1
Extension D	-	-
Extension E	-	-
Compatibility Ideograph Block	12	12
Total	5 009	5 033

2.5 The ISO/IEC 10646 document provides compatible characters for characters included in the CJK Compatibility Ideograph Block, in which 12 are HKSCS-2015 characters. These 12 characters and their corresponding characters are listed in Table 3 for reference.

2.6 As most of the existing computer platforms support ISO/IEC 10646, persons-in-charge are recommended to upgrade their systems to support ISO/IEC 10646 as soon as possible to enable more efficient and convenient use of the most comprehensive Chinese character set for communication and information exchange.

Section 3 HKSCS-2015 under the Architecture of the ISO/IEC 10646

3.1 Under the architecture of the ISO/IEC 10646 international coding standard, Han characters refer to the CJK unified ideographs. Each code point of the CJK Unified ideographs has multiple glyphs listed in individual columns. This multi-column format serves to support and define the characters needed in a particular country or region. The Chinese characters used in Hong Kong are listed in the H-column. Details of the ISO/IEC 10646 international coding standard are available at: http://standards.iso.org/ittf/PubliclyAvailableStandards/.

3.2 All the HKSCS-2015 characters have been included in the ISO/IEC 10646 international coding standard. Characters used in Hong Kong but not included in the standard will be added through vertical extension. For characters already in the standard, information will be added to specify which characters are used in Hong Kong and the glyphs used locally will be included through horizontal extension, so as to reflect the actual use of Chinese characters in Hong Kong.

3.3 Based on the HKSCS-2008, the HKSCS-2015 has newly included 23 Chinese characters already in the ISO/IEC 10646. These characters are denoted as "HD-XXXX" (where "XXXX" is the code point of the character in ISO/IEC 10646). The preferred glyphs used in Hong Kong and the source reference are also added to the H-column.

3.4 The HKSCS-2015 has also added one new symbol, which is already included in the ISO/IEC 10646. This symbol is denoted in the form of "HE-XXXX", where "XXXX" is the code point of the symbol in the ISO/IEC 10646.

3.5 For any HKSCS character to be added to the ISO/IEC 10646 in the future, the source reference will be given in the form of HC-0001 to HC-9999, denoting that the character is included in the ISO/IEC 10646 through vertical extension. The HKSCS-2015 does not include characters with source reference denoted in the form of "HC-".

3.6 Under the architecture of the ISO/IEC 10646, the H-column lists not only the characters in the HKSCS, but also those included in the Big-5 coding scheme. The source reference of such Big-5 characters is provided in the form of “HB0-XXXX”, “HB1-XXXX” and “HB2-XXXX ”, where “XXXX” is the code point of the character under the Big-5 coding scheme, denoting characters from the Big-5 symbol area, frequently used characters and less frequently used characters respectively. Information on the mapping of the Big-5 and Unicode is available at the following website:

http://www.unicode.org.

Table 1: Code Table of the HKSCS-2015

The HKSCS-2015 contains 5 033 characters, including 5 009 characters from the HKSCS-2008, and 24 newly added characters (23 Chinese characters and one symbol).

The following are examples of typical cells in the code table of the HKSCS-2015:

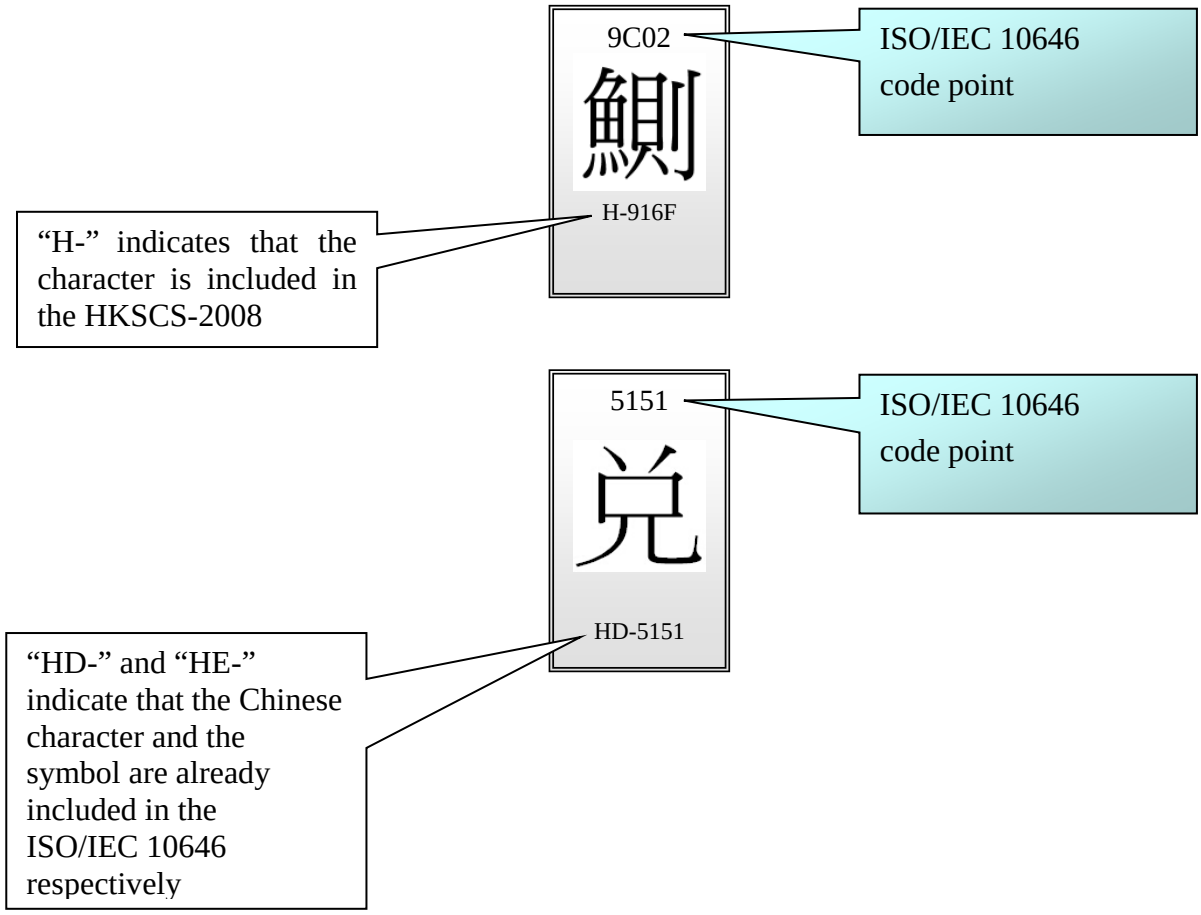

Table 2: Newly Included Characters in the HKSCS-2015

Compared with the HKSCS-2008, the HKSCS-2015 includes 23 more Chinese characters and one more symbol. These characters are included in the ISO/IEC 10646 through horizontal extension, reflecting their use in Hong Kong and further fulfilling the needs of the local electronic communication in Chinese. These 24 characters are listed below.

S/N	Glyph commonly used in Hong Kong and its ISO/IEC 10646 code point	Unifiable character and its ISO/IEC 10646 code point
1	兌 5151	兌 514C (corresponding Big-5 code: 0xA749)
2	悅 60A6	悦 6085 (corresponding Big-5 code: 0xAEAE)
3	挩 635D	挩 6329 (corresponding Big-5 code: 0xD1BE)
4	敓 655A	敓 6553 (corresponding Big-5 code: 0xD5A7)
5	梲 68C1	梲 68B2 (corresponding Big-5 code: 0xD5BF)
6	涀 6D9A	涀 6D97 (corresponding Big-5 code: 0xD258)
7	稅 7A0E	稅 7A05 (corresponding Big-5 code: 0xB57C)
8	脫 8131	脫 812B (corresponding Big-5 code: 0xB2E6)
9	蛻 8715	蛻 86FB (corresponding Big-5 code: 0xB8C0)
10	誽 8AAC	說 8AAA (corresponding Big-5 code: 0xBBA1)
11	鋭	銳

S/N	Glyph commonly used in Hong Kong and its ISO/IEC 10646 code point	Unifiable character and its ISO/IEC 10646 code point
	92ED	92B3 (corresponding Big-5 code: 0xBE55)
12	閱 95B2	閱 95B1 (corresponding Big-5 code: 0xBE5C)
13	媼 5AAA	媼 5ABC (corresponding Big-5 code: 0xB6FE)
14	慍 6120	慍 614D (corresponding Big-5 code: 0xB759)
15	氳 6C32	氳 6C33 (corresponding Big-5 code: 0xBA72)
16	煴 7174	煴 7185 (corresponding Big-5 code: 0xE2BE)
17	緼 7DFC	縕 7E15 (corresponding Big-5 code: 0xEAD5)
18	腽 817D	膃 8183 (corresponding Big-5 code: 0xE3A6)
19	蘊 85F4	蘊 860A (corresponding Big-5 code: 0xC4AD)
20	轀 8F3C	轀 8F40 (corresponding Big-5 code: 0xEEC1)
21	醞 9196	醞 919E (corresponding Big-5 code: 0xC1DF)
22	告 543F	告 544A (corresponding Big-5 code: 0xA769)
23	鱷 9C47	---
24	€ 20AC	---

During the formulation of the Reference Glyphs for Chinese Computer Systems in Hong

Kong, it is found that for some characters (characters S/Ns 1 to 22), there are differences between the glyphs commonly used in Hong Kong and those specified in the Big-5 code table. These glyphs used in Hong Kong are therefore included in the HKSCS-2015. As “鱇” (character S/N 23) is included in the International Ideographs Core and commonly used in Hong Kong together with “鮟”, a HKSCS-1999 character, to form the name of the food fish “鮟鱇”, and as the euro sign “€” is a commonly used currency symbol in Hong Kong, they are also included in the HKSCS-2015.

Table 3: Compatibility Characters in the HKSCS-2015

The HKSCS-2015 contains 12 characters which are in the CJK Compatibility Ideographs Block. Their corresponding characters are shown in the following table for reference.

S/N	Compatibility characters in the HKSCS-2015	Corresponding character
1	龜 F907	龜 9F9C
2	勇 2F825	勇 52C7
3	吆 2F83B	吆 5406
4	罟 2F840	罟 54A2
5	屮 2F878	中 5C6E
6	弋 2F894	弋 5F22
7	慈 2F8A6	慈 6148
8	晋 2F8CD	晉 6649
9	芳 2F994	芳 82B3
10	夔 2F9B2	夔 456B
11	蜨 2F9BC	蜨 8728
12	貫 2F9D4	貫 8CAB