LegCo Panel on Information Technology and Broadcasting

Hong Kong Supplementary Character Set

Introduction

This paper explains the objective of developing the Hong Kong Supplementary Character Set (HKSCS), the criteria for the inclusion of characters in the HKSCS and other related matters.

Objective of developing the HKSCS

2. The ISO 10646 and the BIG-5 coding scheme commonly used in Hong Kong do not encompass some of the Chinese characters which the Hong Kong Special Administrative Region (HKSAR) Government and the public need to use in electronic communication in Chinese. Owing to the absence of a common Chinese language interface, conflicts of internal codes arise in electronic communication. (Please refer to Annex 1 for details.) The objective of developing the HKSCS is to collate Chinese characters which computer users need for practical purposes and which are not at present included in any of the commonly used coding schemes. In doing so, we seek to provide a common Chinese language interface so as to facilitate accurate electronic communication in Chinese. Annex 1

3. In 1995, the Hong Kong Government developed a set of coded Chinese characters called the Government Common Character Set (GCCS) in the User-Defined Area of the Big-5 coding scheme to facilitate electronic communication among government departments. This character set, which included a number of Chinese characters unique to Hong Kong, was supplementary to the Big-5 basic character set. Initially, GCCS was intended for use within Government only. Subsequently, in order to enable the public to view government documents on the Internet, GCCS was made freely available on the Government's web site.

4. In April 1998, the Information Technology Services Department (ITSD) and the Official Languages Agency (OLA) jointly reviewed the GCCS. They collected additional characters from different sectors of the community with a view to developing an updated common character set. This updated character set was intended to facilitate the development of a common Chinese language interface for data exchange between government departments and the public. As the updated GCCS includes characters collected from different sectors of the community, it serves as a common character set for use by the general public. Thus, it was renamed as the HKSCS.

Development of the HKSCS and criteria for the inclusion of characters in HKSCS

5. In 1998, ITSD and OLA collected characters for Chinese computing from government departments and non-government institutions. OLA processed each of the characters received and eliminated the duplicate ones. The remaining characters were checked against authoritative dictionaries (including “Kangxi Zidian” (《康熙字典》), “Hanyu Da Zidian” (《漢語大字典》), “Hanyu Da Cidian” (《漢語大詞典》) and “Zhonghua Zihai” (《中華字海》)), Cantonese dialect dictionaries and relevant literature, and documents in the ISO 10646, in order to ascertain whether the characters could be found in dictionaries, whether they were commonly recognised Cantonese characters and whether they had already been included in the ISO 10646. As a result, some characters, the source or usage of which could not be verified, were deleted and about 1,700 new characters were obtained. Characters in the GCCS were also reviewed and it was found that 106 characters had to be eliminated. Consequently, the total number of characters contained in the HKSCS is 4,702.

6. ITSD has also set up the Chinese Language Interface Advisory Committee (CLIAC) to advise the Director of Information Technology Services on the submission of local characters for Chinese computing to

the International Organization for Standardization (ISO), the promotion of the adoption of ISO 10646 in Hong Kong, and the co-ordination with different sectors in Hong Kong on the need for additional characters for Chinese computing locally. The CLIAC comprises members from academic institutions, language and linguistic bodies, and the information technology and publishing industries. The member list of CLIAC is at Annex 2. The CLIAC and its two Working Groups have met on a number of occasions to discuss the criteria for the inclusion of characters into the HKSCS, the coding scheme and other aspects of the HKSCS. Where appropriate, the Information Infrastructure Advisory Committee (IIAC) under the Information Technology and Broadcasting Bureau is briefed and consulted on the discussion items of the CLIAC. Discussion papers of the CLIAC are also placed on the Government's web site for reference by the public, whose comments are welcome.

Annex 2

7. The basic principle for the inclusion of characters in the HKSCS is that each character should have a known source and is required by either Government or the public for information exchange. As the HKSCS will ultimately converge with the ISO 10646 character set, we also make reference to the Unification Rules of ISO 10646 when verifying the characters collected. This principle for the inclusion of characters in the HKSCS was discussed and endorsed by the CLIAC and its Working Groups.

8. The updated character set and the allocated codes were endorsed by the CLIAC in September 1999. The specifications of the character set were published on 28 September 1999, and were made available on the Government's web site on the same day. Based on the specifications published, software developers can proceed to develop compatible products such as fonts, input methods and handwriting recognition devices. For ease of reference by the public, ITSD has developed a set of font and input method which is available on the Government's web site (www.digital21.gov.hk/chi/hkscs).

Reasons for including characters used in Cantonese expressions in the HKSCS

9. Some characters used in Cantonese expressions have been included in the HKSCS. Most of these characters were submitted by the Judiciary, the Hong Kong Police Force, the Department of Justice, linguistics societies and academic institutions. Some of them can be found in Cantonese dictionaries or academic writings. One of the reasons for their inclusion is to facilitate the Judiciary, the Police Force and other law enforcement agencies to prepare records of proceedings and take statements. Since the records of proceedings and statements produced by the Judiciary and law enforcement agencies are in verbatim, they need to use these Cantonese characters to produce an accurate and complete record. In addition, the study of Cantonese is an academic subject on which many linguistic societies and tertiary institutions are conducting research. They may encounter difficulties when publishing their papers if the HKSCS did not include characters used in Cantonese expressions. In including these characters in the HKSCS, it is not our intention to encourage the public to use them. The HKSCS is just like a toolbox containing different tools from which different users may select and use the appropriate tools according to their own circumstances and needs.

Reasons for including some characters that cannot be found in dictionaries

10. Some characters in the HKSCS cannot be found in dictionaries. These characters are mainly from the databases of the Immigration Department, the Company Registry, Inland Revenue Department and Lands Department, and have been confirmed to be in use as names of persons, companies and places. These names may appear in various kinds of identity documents, contracts and legal documents and the exact characters used in the registered names must be used in such documents. Therefore, these characters must be included for practical reasons. To meet the needs of the public and Government for additional characters for

Chinese computing, and to avoid unnecessary additions to the HKSCS, the HKSAR Government and the CLIAC are considering further improvements to the existing mechanism for the collection, verification and management of additional characters for Chinese computing.

Mechanism for the management of the HKSCS

11. The CLIAC has established the mechanism for the management of the HKSCS. Under this mechanism, the CLIAC and its two Working Groups will be responsible for the following functions:

- to propose criteria for the verification and inclusion of additional characters in the HKSCS;
- to examine requests for the inclusion of additional characters;
- to allocate code points to additional characters;
- to publish and announce updated versions of the HKSCS; and
- to advise on the submission of characters to ISO for inclusion into the ISO 10646 standard and verification of these characters.

12. The above mechanism for the management of the HKSCS has been discussed and endorsed by the IIAC.

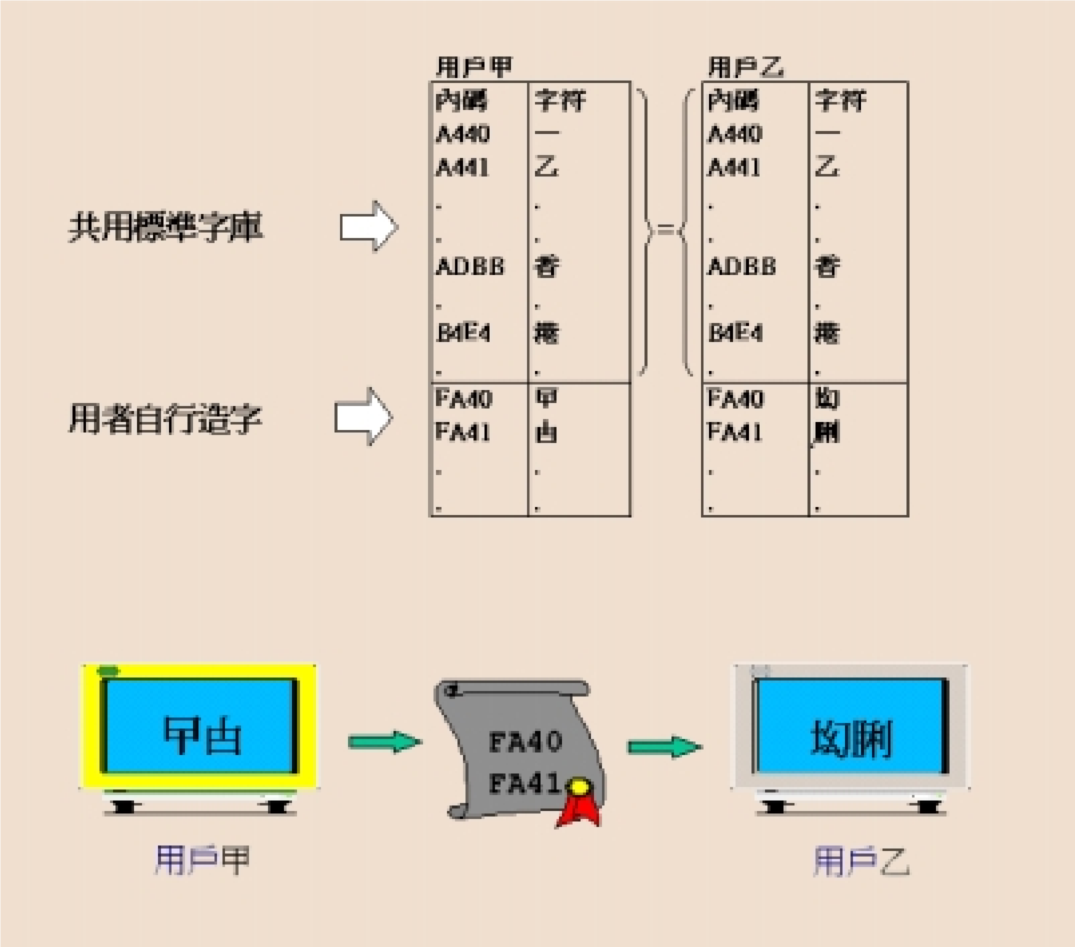
13. The detailed operational procedures of the mechanism are being considered by the Working Groups of the CLIAC. It is expected that these procedures will be finalized by April 2000.

Information Technology Services Department
January 2000

Annex 1

Conflicts of internal codes due to the absence of a common Chinese language interface

Information stored in a computer is coded according to a pre-defined coding scheme. For information in Chinese, there are different coding schemes, including "Big-5", "GB" and " ISO 10646". These coding schemes, however, do not cover all characters in use in Hong Kong. Among the characters not included in these schemes, many are unique to Hong Kong such as names of persons and places, and Cantonese expressions. To be able to use such characters on a computer, users need to assign internal codes for these characters in the user-defined area by themselves. This works well in stand-alone computers, but when computers are connected with each other, these user-defined characters may give rise to problems in communication and data exchange. For example, User A and User B have separately assigned the same internal codes for different characters in their own user-defined areas. As illustrated by the example below, the characters “曱甴” which appear in the message transmitted by User A will be displayed as “坳脷” on the screen of User B, causing problems in electronic communication.



用戶甲

內碼	字符
A440	一
A441	乙
.	.
.	.
ADBB	香
.	.
B4E4	港
.	.
FA40	甲
FA41	由
.	.
.	.

=

用戶乙

內碼	字符
A440	一
A441	乙
.	.
.	.
ADBB	香
.	.
B4E4	港
.	.
FA40	坳
FA41	脷
.	.
.	.

Annex 2

Membership of the Chinese Language Interface Advisory Committee

Director of Information Technology Services (Chairman)
Mr. CHAN Chun Leung
Mr. Dominic K. CHENG
Mr. Raymond CHENG
Dr. CHEUNG Kwan Hing
Dr. KEUNG Wing Ching
Mr. Francis LEE
Mr. Fan LOOK
Dr. Qin LU
Dr. LUKE Kang Kwong
Mr. Lawrence MO
Mr. Arics POON
Mr. Tony TAI
Mr. TSANG Hip Tai
Prof. Benjamin K TSOU
Mr. YAO Te Hwai
Representative of the Information Technology and Broadcasting Bureau
Representative of the Official Languages Agency