中文資訊交換碼

CCCII

Volume Ⅱ

Second Printing

第二冊　第二版

中華民國七十一年十月

0. INTRODUCTION

0.1 This set of standard codes for Chinese information exchange among data processing systems and within message transmission systems is a direct extension in terms of the number of characters coded of volume I of the Chinese Character Code for Information Interchange, which was published in April 1980. The errors, mostly in the placement of characters in the coding space, found in Volume I have been corrected in this publication. Since the basic structure of the coding space, the extension techniques, and the escape sequences for the identification process are the same as those used in Volume I, they will not be repeated here except when revisions are necessary.

0.2 The total number of characters included in this volume is 33,544. These characters consist of all the characters appearing in Volume I, the characters used in most data processing centers in Taiwan, and the characters in the bibliographic applications in China(Ref. table I, page 7). Also, the variant forms of the characters that appeared in this volume have been collected and coded. These variant forms, including the simplified forms of characters, are published in a separate volume, Part 2 of Volume II.

0.3 The code consisting of three 7-bit bytes vectors forms a finite 3-dimensional coding space, i.e.

it provides a total of 94 * 94 * 94 positions according to ISO 2022. It is called a space of 94 planes, with 94 sections in each plane and 94 positions in each section.

0.4 Since certain characters have variant and simplified forms the coding space has been subdivided into layers so that the character codes of the variant and/or simplified forms of a character can be used to trace the original character, Since such relationships between the characters and their variant forms can in many cases have important applications. Details of the layering will be explained in later sections of this publication.

0.5 In addition to the standard codes published in this volume, a cross-reference data base has been constructed. This data base contains the major indexing methods of Chinese characters currently being used in Taiwan, in the form of cross-reference tables that arranged according to the character sequences in this publication. Also, the popular phonetic transcription of the characters are stored in the data base. Furthermore, the image of every character appearing in Volume II of CCCII is contained in the data base in the form of a 32 * 32 dot matrix. This cross-reference data base is available on request.

1. SCOPE AND FIELD OF APPLICATION

1.1 The coded characeters include:

(1) 35 Chinese punctuation marks.

(2) 214 radicals.

(3) 41 Chinese numerical characters.

(4) 37 Chinese phonetic symbols and 4 tone marks.

(5) 4808 most frequently used Chinese characters.

(6) 17077 next frequently used characters.

(7) 11660 variant forms of the characters in(5) and (6).

This set of Chinese characters is denoted as CC1 (Chinese Character set 1). To assist in possible applications, this set is divided into two blocks. the first block, called CB1, contains the most fequently used 4808 Chinese characters. The second block, called CB2, contains the rest of the characters in Volume II. However, the characters in (1) to (4) are indispensable to both the two blocks of Chinese characters. It is believed that CB1 is sufficient for teaching, everyday use, newspaper printing, and populur cataloging purposes. CB2 together with CB1 is sufficient for Chinese bibliographic applications.

1.2 This publication consists of code tables for the characters mentioned in 1.1, a legend showing each graphic character other than Chinese characters and the escape sequence for this coding scheme. Explanatory notes are also included.

1.3 This character set is compatible with and may be in conjunction with any ISO character set which is based on the 7-bit code as specified by ISO 646.

2. IMPLEMENTATION

2.1 As mentioned in the introduction, the coding structure, the extension techniques, and the escape sequence for the identification process are all the same as used in Volume I. However, there are two major technical changes in this volume:

(1) The first section of each plane, as shown in Fig. 1, is reserved for the control codes that are required in handling Chinese character strings.

(2) The forming of layers for the purpose of storing variant forms of Chinese characters. This will be illustrated in the following section.

2.2 Started from plane 1, every six consecutive planes are grouped together to form a layer. Therefore, there are 16 layers, the last layer of which contains only four planes. This structure is shown in Fig. 2. The usage of these 16 layers is as follows:

(1) Layer 1, i.e. plane 1 through 6, is used to designate the graphic characters as mentioned in 1.1 and all other traditional forms of Chinese characters.

(2) Layer 2, i.e. plane 7 through 12, is used of designate the simplified form of Chinese characters that are used in mainland China.

(3) Layers 3 through 13 are used to designate

the variant forms of the Chinese characters that appeared in layer 1. The handling of these variant forms is discussed in the next section.

(4) Layer 16, i.e. the last layer, is used for designating characters of all other relevent languages.

(5) Layer 14 and 15 are reserved for other usage.

2.3 Handling of variant forms

A character having variant forms is a unique property of Chinese language. A character and its variant forms usually have exactly the same pronounciation and meaning, but differ in their stroke structure. Generally, a character and its variant forms are interchangable in writing. But, when used as identifiers to name persons, places or things,they are considered different characters, and should not be confused. In CCCII, the code assigned to the variant forms of a character is designed to have the same two furthest right bytes as the code of the corresponding original character. In other words, the variant form is placed in the same section and position as its corresponding original character, but in a different plane. By, such an arrangement, variant forms of a character can be identified and interchanged with the original form very easily when required. An example showing such an arrangement of variant forms of characters is given in Fig. 3.

2.4 The code assignment of graphic characters other

than Chinese characters in plane 1 is the same as in Volume I, and consequently not be repeated in this publication.

2.4 The ordering of Chinese characters in this volume follows the same convention as in Volume I.

2.5 The coding of each section is shown in its corresponding section table (see page 29). A CODE TABLE of CC1 of CCCII is also attached.

2.6 The escape sequence for identification of the designated code is the same as in Volume I, i.e. : ESC, 2/4, 4/1 For detailed specifications, please refer to section 5.3.9 of ISO 2022.

3. LEGEND

3.1 Symbols:

Notes:

(1) All the punctuation marks and special characters listed in the section tables and this code table are commonly used in Chinese data processing.

(2) The ASCII punctuation marks and characters are listed in the section tables and this code table, which exactly follows the positions assigned and their definitions specified in ISO 2022.

4. EXPLANATORY NOTES

4.1 Relevant ISO publications:

Table 1: Survey of characters for general data processing in Taiwan

item	Resource (Company Name or Book Title)	Number of characters (Approx.)	Applied fields
1	Taiwan Automation Co.	9,600	These systems have been applied to data processing works in the following areas: .Electricity .Libraries .Gas .small business companies .Telephone .Banking .Government Organizations .Water .Tax .Schools .Police
2	Wang Industrial Co.	10,499	
3	Fincial & Tax Center	11,000	
4	IPX Taiwan Ltd.	9,600	
5	Feng-chia University	16,000	
6	National police Administration	11,825	
7	Characters for Libraries	15,000	
8	Chiao-tung University	9,129	Basic character set for all Chinese data processing application and researches
9	The code book of the three corner coding method	8,800	The most popular and well adapted by users.
10	The Comprehensive Dictionary of Chinese Character. Index	9,600	
11	Chinese Characters for Telegraph Code	8,000	for data transmissions
12	Characters for Elementary school	4,600	for teaching
13	The Ministry of Education	12,701	These are the standard forms so far published

Table 2: The subset of CCCII with 2-Byte subcode length

subset	No. of characters	character set Name	publication	Sub-code length	compatibility	User defined positions	Typical Applications
CCCII-1	4808	The most frequently used characters (plane 1)	Vol.1 of CCCII	Two 7-bit Bytes	Fully ISO 646 Compatible	658	teaching, writing, newspaper printing, word processing
CCCII-2	22000	The character set for general data processing (plane 1,2 & 3)	Vol.2 of CCCII	two 8-bit Byte	Fully ISO 646 Compatible	658	Business Data Preconsing Library, Cencus applications
CCCII-3	33600	Complete first 4 planes	Not yet	two 8-bit Byte	Fully ISO 646 Compatible	658	For more specific applications

(1) ISO 646 7-bit coded character set for information processing interchange

(2) ISO 2022 code extension techniques for use with the ISO 7-bit coded character set.

(3) ISO 2375 data processing.....procedure for registration of escape squences

5. APPLICATION NOTES

5.1 The subsets of CCCII

The characters collected in Volume II include the characters used in most of the data processing centers in Taiwan, A list of the sources and number of characters used in each center is shown in Table 1. CCCII Volume II is based on the above collected data. However, since CCCII is designed to serve as a universal tool for Chinese information interchange, all symbols used in Chinese language and information handling will be collected and coded.From user's point of view, however applications do not need such a complete set of characters. In order to fulfill the above two criteria, CCCII is so designed that the user can select a proper subset of the code to make it suitable for his specific requirements

The subsets with typical applications, of CCCII with 2-byte subcode length are listed in Table 2. using these subsets, by agreement between the interchanging parties, only 2-byte are needed to represent a Chinese character. It should be notes that the code structure of all these subsets of CCCII is fully ISO 646 compatible.

5.2 The radix-94 code for internal storage

The first 6 planes in the first layer (fig. 2) have a total of 94*94*6=53,016 positions. Since this number is less than 2^{16}=65,536, the 3-byte code in the first layer can be compressed into one 16-bit or two 8-bit bytes as an unsigned binary number by applying radix-94 conversion. This conversion is an one-to-one, two way and unique conversion. Therefore, The resultant code can be used as an internal code for mass data storage to save storage space if the characters used are limited to those in the first layer.

The above mentioned conversion is done by using the following equation:

$$R94=\left[(B_3-33) \bmod 6\right] * 94^2 + (B_2 - 33) * 94 + (B_1 - 33)$$

where B_3 is the plane number,

B_2 is the section number,

B_1 is the position number, and

mod is an arithmetic function that will take the remainder after $(B_3 - 33)$ is divided by 6.

5.3 Code-length compression by commands.

The first section of each plane is reserved for commands used in Chinese data processing or defined by the user. For some applications if 3-byte code is to be used, the user may define commands to switch among planes and let the Chinese character code be two bytes. By this arrangement, the length of the code for a block of characters may be significantly reduced.

As an example, let us define two switch commands, as follows:

Command name	parameters	'Comment
Switch plane, locked	plane number	4 bytes more for each switch of a group of character
switch plane, unlocked	plane number	4 bytes more for each switch of a single character code.

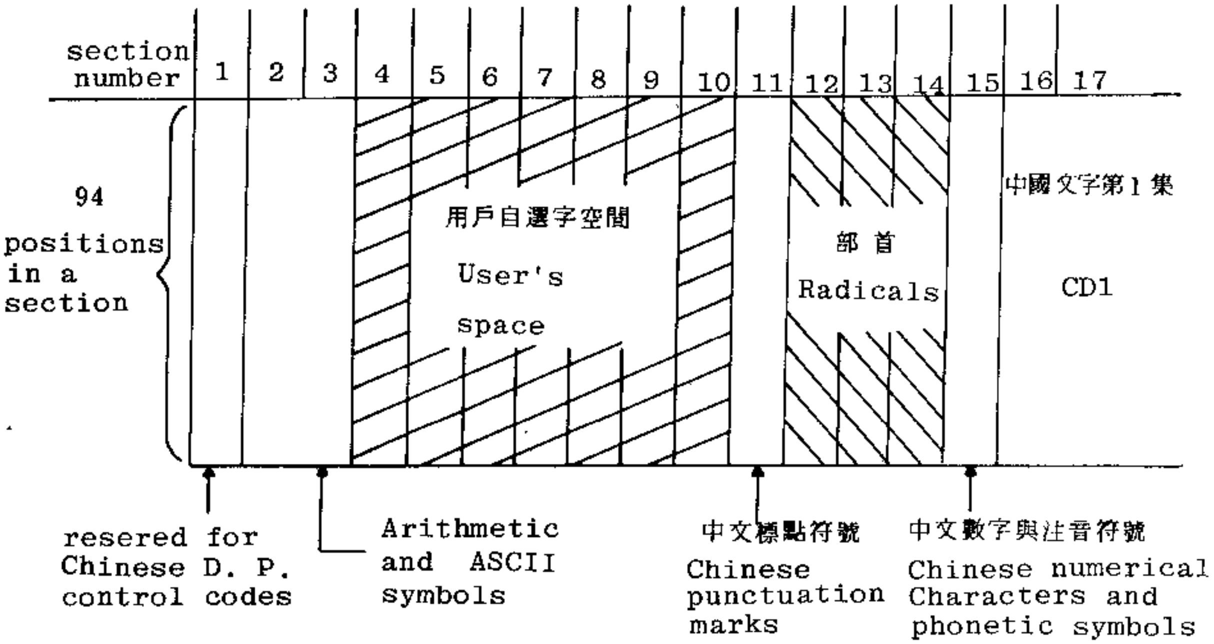
The third byte B_3 and the plane number may be omitted for all the characters. Assuming the command name and parameters are 2-bytes each, there is a 95% chance to use those characters in plane 1. Then, the weighted average of code length can be computed as follows, if the unlocked switch plane command only used:

average code length= $2 * 95\% + 6 * (1- 95\%)$

=1.9 + 0.3

=2.2 bytes

The 7-bit bytes used to code command names and parameters all belong to the set GO as defined in ISO 646. This causes the switching operation to be fully located within the processing operation of Chinese characters and hence fully ISO 646 compatible.


section number
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
94 positions in a section
用戶自選字空間
User's space
部 首
Radicals
中國文字第1集
CD1
resered for Chinese D. P. control codes
Arithmetic and ASCII symbols
中文標點符號
Chinese punctuation marks
中文數字與注音符號
Chinese numerical Characters and phonetic symbols

圖一第一層前15段之結構

fig. 1 The structure of the first 15 sections in plane 1

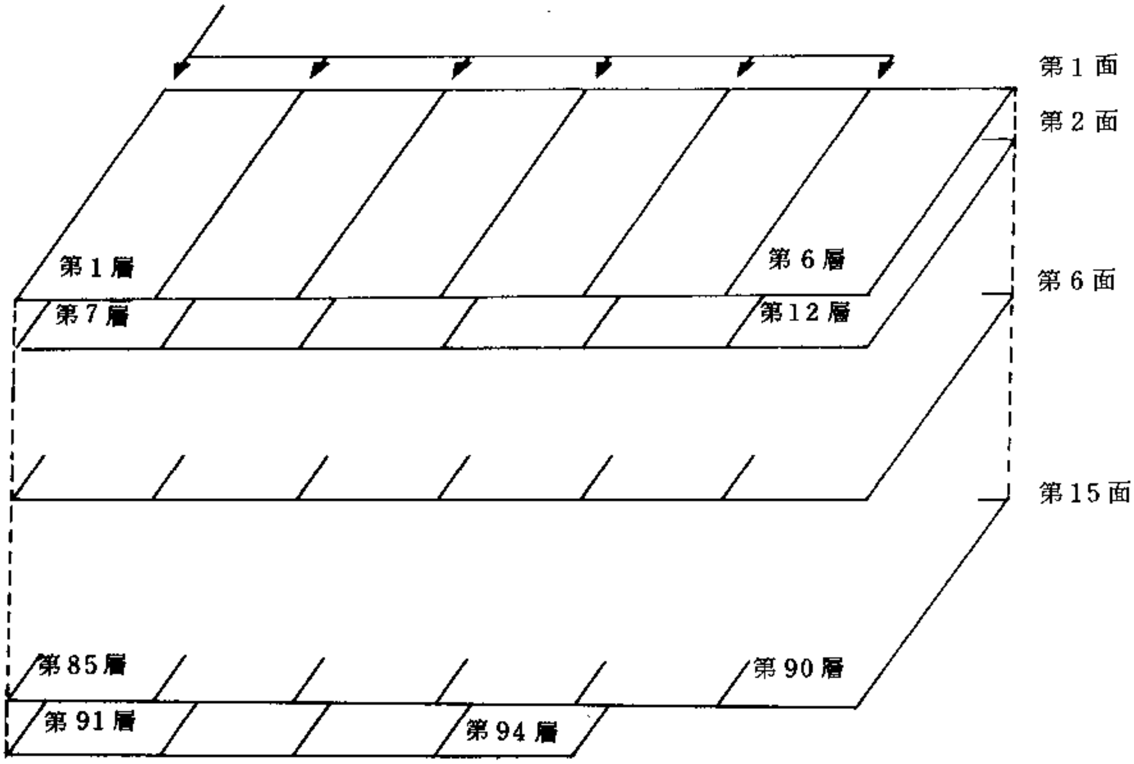
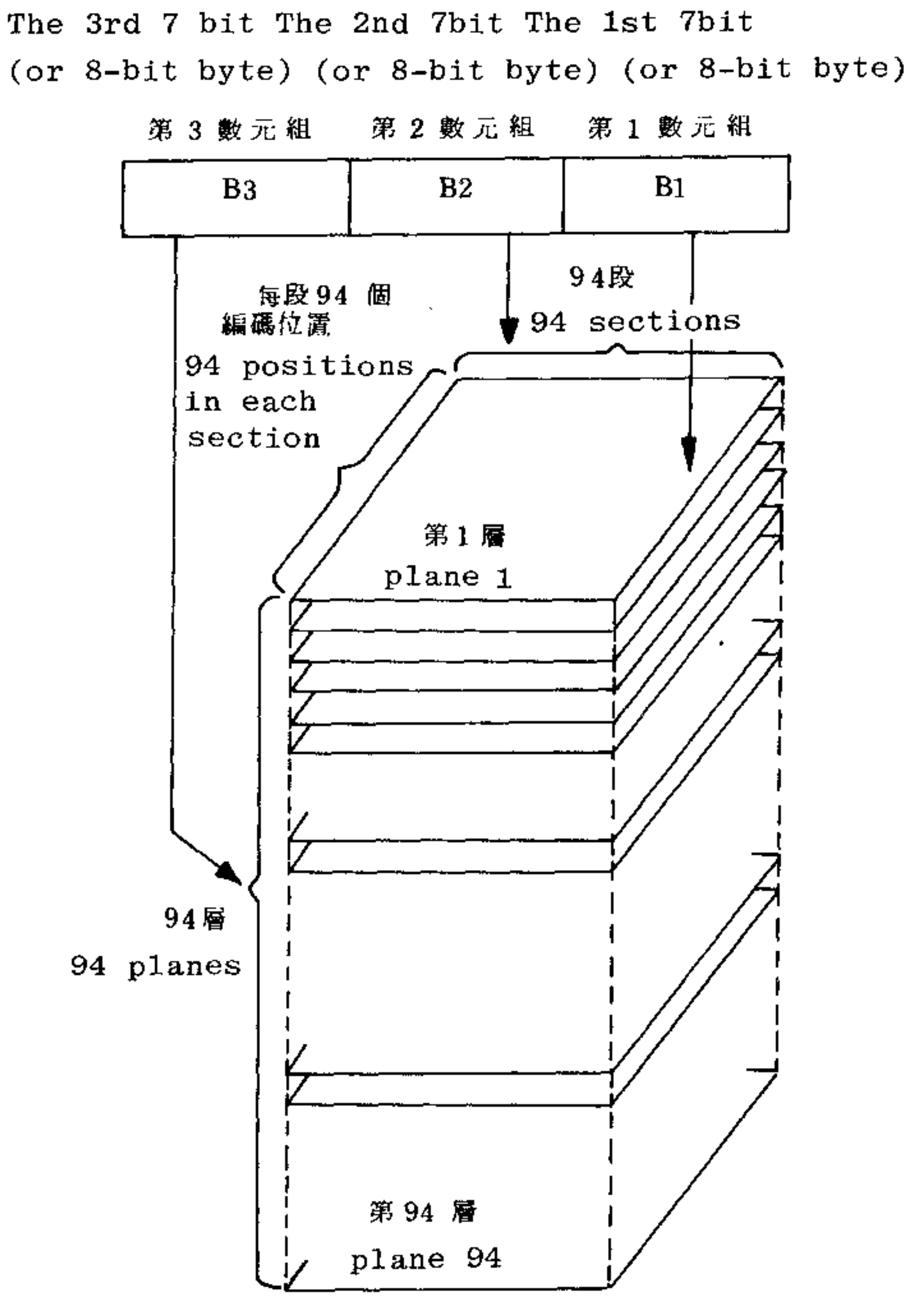


Fig.3


The 3rd 7 bit The 2nd 7bit The 1st 7bit
(or 8-bit byte) (or 8-bit byte) (or 8-bit byte)
第 3 數元組
第 2 數元組
第 1 數元組
B3
B2
B1
每段94 個
編碼位置
94 positions in each section
94段
94 sections
第1層
plane 1
94層
94 planes
第94 層
plane 94

圖二：三度空間94×94×94=830,584個編碼位置結構圖
Figure 2: The 3-dimensional Structure of the whole 94×94×94= 830,584 coding spaces

異體字形CCCII 編碼結構實例

		1 *3*	1 *3*	1 *3*	1 *3*	1 *3*	1 *3*	1 *3*	1 *3*	1 *3*	1 *3*	B2
		7 *1*	7 *1*	7 *1*	7 *1*	7 *1*	7 *1*	8 *2*	8 *2*	8 *2*	8 *2*	
		7 *7*	7 *7*	7 *7*	7 *7*	7 *7*	7 *7*	2 *2*	2 *2*	2 *2*	2 *2*	B1
B3		7 *7*	9 *9*	C *C*	A *A*	D *D*	3 *3*	6 *6*	2 *2*	4 *4*	5 *5*	
0 *2*	1 *1*	倖	倆	值	借	倚	們	倀	俱	個	候	normal form 此列爲通用體
0 *2*	7 *7*		俩	值			们	伥	俱	个		simplified form 以下四列爲同義異體
1 *2*	3 *D*	幸	俩		偺	倚				箇	侯	other variations 此列爲大陸簡體

An example of the table of variant forms
and their associated CCCII codes.