
Information Technology - Chinese Ideograms Coded Characters Set for Information Interchange –

Extension for the Basic Set

Introduction

1. Scope

.1 Abstract

This standard, as a mandatory code standard of GB 2311 Chinese ideograms encoding system, sets the primary Chinese ideogram characters and their hexadecimal representation of binary codes for information interchange, it is an extension of the GB 2312.

The following are set forth in this standard:

- Background for publishing this standard
- Repertoire collected in this standard
- Overall encoding structure and rules
- Code position arrangement of characters
- Allocation of code positions
- Attachments to this standard (Appendix A, Appendix B, Appendix C and Appendix D)

.1 Background

GB 2312-80, the primary collection of Chinese coded graphic characters published in 1981 as a national standard, covers only 6,763 Chinese characters. In 1995, GBK (Chinese Internal Code Specification) for GB Extension was published. It is the super set of GB and totally compatible with GB 2312-80. GBK expands its characters set to 20,902 characters.

In order to cover and process more Chinese characters and meet the requirement for Chinese customers (such as Banking and Posting etc) on super set of Chinese characters processing capability and keep the compatible with the existing GB 2312-80 and GBK encoding system, GB 18030 for super set of Chinese characters code standard is released.

In GB 18030, one-byte, two-byte and four-byte encoding systems are adopted. The total capability is over 1.5 millions of code positions. Currently, GB 18030 contains more than 27,000 Chinese characters which have been defined in the latest version Unicode 3.0. This standard provides utter solutions for the urgent needs of Chinese characters used in names and addresses.

.1 Objective

This standard makes concrete stipulations on one- and two-byte coded Chinese ideogram characters, and also makes stipulations on the systematic structure of four-byte encoding system.

This standard adopts the following definitions:

- Repertoire

A definitive set of characters expressed in coded character set.

- Character

One element in an element set used for organizing, control or expressing data.

- Coded Character

Character with its binary-coded representation.

- Reserved Zone

The zone in this standard that reserved for future requires for international standard.

The objectives of this standard are:

1) To provide more coded Chinese characters for meeting the requirement for Chinese information processing and interchange.
2) To ensure consistent implementation of Chinese characters used in both Chinese and other Minority Nationalities' languages (such as Tibetan and Mongolian).
3) To provide a perpetual coded standard for future expanding while compatible with former Nation Standards, such as GB 2312-80 and GB 13000.

.1 Application

This standard is applicable to representation, processing, interchange, storage, transmission, display, input and output of information expressed in Chinese ideogram characters.

All foreign and domestic IT companies must comply with this new standard to implement their new product.

.1 Effective Date

This standard takes effect from the date of issue – March 17, 2000, with a transition period until December 31, 2000.

All new Chinese products since January, 2001 in China market should support this standard.

2. Document Administration

2.1 Originating Area and Responsibility

This standard is proposed by the Ministry of Information Industry of People's Republic of China.

This standard is summed up by the Electronics Industrial Standardization Research Institute of the Ministry of Information Industry .

The following units joined the drafting of this standard: Electronics Industrial Standardization Research Institute of the MII, Computer Technology Research Institute of Beijing University, Founder Group of Beijing University, Beijing Founder & Suntenday Information Network Science and Technology Liability Company, Ltd., Stone Group Company, Software Institute of Chinese Academy of Science, Great Wall Software Company, Stone Rich-sight Company, China Software Company General, Kingsoft Company and Legend Group.

The major drafters of this standard are Chen Kunqiu, Huang Jiang, Hu Wangjin, Zhang Jianguo, and Chen Zhuang.

2.2 Authorization

This standard was approved and released by the Ministry of Information Industry and China State Bureau of Quality and Technical Supervision (CSBTS).

2.3 Compliance

All foreign and domestic IT companies must comply with PRC national standard (GB 18030-2000 or GB 2312-80) to implement their new product.

As a mandatory standard, China State Bureau of Quality and Technical Supervision will supervise the GB 18030 standard implementation status for new products in China market. If new product is found to be inconsistent with new standard, the related IT company will be punished according to articles in The Standardization Law of People's

Republic of China. It is suggest that IT companies send their new products to Chinese Software Product Testing Center for testing.

3. Related Document

3.1 Superseded Document

Chinese Characters Internal Code Specification for GB Extension (GBK), Version 1.0
This technical specification was promulgated and implemented in [1995] No. 229 Technical Sponsorship Letter jointly by the Standardization Division of the former China State Bureau of Technical Supervision, and the Science and Technology and Quality Supervision Division of the former Ministry of Electronics Industry.

3.2 Referenced External Documents

GB 2311-1990
PRC National Standard, Information Processing - Code Extension Technique for 7-bit and 8-bit coded Characters Set (eqv ISO4873: 1986)
GB 2312-1980
PRC National Standard, Code of Chinese Graphics Characters Set for Information Interchange, Primary Set
GB 11383-1989
PRC National Standard, Information Processing - Structure and Encoding Rules for 8-bit Code for Information Interchange (idt ISO 4873: 1986)
GB 12345-1990
PRC National Standard, Code of Chinese Graphics Characters Set for Information Interchange, Auxiliary Set
GB 13000.1-1993
PRC National Standard, Information Technology – Universal Multiple Octet Coded Character Set (UCS), Part One: Systematic Structure and Basic Multi-Language Plane (idt ISO /IEC 10646.1: 1993)

3.3 Referenced IBM Documents

3.4 Copyright Permission

Not applicable.

Requirements

4. External Standards

4.1 *PRC National Standard GB 2312-80*

GB 2312-80, PRC National Standard, Code of Chinese Graphics Characters Set for Information Interchange, Primary Set, published in March 1981, specifies a primary set of graphics characters with their binary-coded representation for Chinese information interchange. It applies to Chinese information-processing systems, communication systems and so on. It covers 682 non-Chinese characters and 6,763 Chinese characters, 7,445 graphic characters in total.

The non-Chinese characters include general characters, ordinal numbers, numerical characters, Latin alphabet, Japanese Kana, Greek alphabet, Russian alphabet, Chinese phonetic symbols and Chinese phonetic-annotated letters.

The Chinese Characters are divided into 2 levels, 3,755 of them are included in Level 1 and 3,008 in Level 2, 6,763 Chinese characters in total.

4.1 *Relationship of GB 18030-2000 with GB 2312-80 and GBK*

GB 18030-2000 is a superset of GB 2312-80 and GBK. Those characters that defined in GB and GBK have exactly same code assignment in GB 18030-2000.

4.1 *Relationship of GB 18030-2000 with Unicode/ISO 10646.1-1993*

The Unicode (idt PRC standard GB 13000.1-1993) is international standard for the universal character encoding scheme for written characters and text. It defines a consistent way of encoding multilanguagal text that enables the exchange of text data internationally and creates the foundation for global software. The Unicode standard is a superset of all characters in widespread use today. It contains the characters from major international and national standards as well as prominent industry character sets.

GB 18030-2000 contains all characters defined in Unicode, but they have totally different code assignment. Currently, GB 18030-2000 contains more than 27,000 Chinese characters which have been defined in the latest version of Unicode 3.0. In the future, more Chinese characters can be extended in GB 18030-2000.

5. Specification of Character Repertoire

Collected in this standard are coded one-byte, two-byte and four-byte characters.

5.1 The One-Byte Portion

In this standard, collected in the one-byte portion are all the 128 characters from 0x00 to 0x7F set in GB 11383, and the one-byte coded Euro symbol.

5.2 The Two-Byte Portion

In this standard, collected in the two-byte portion are as follows:

All the unified Chinese characters in the CJK (Chinese, Japanese and Korean) , set in GB 13000.1,

21 Chinese characters selected from the CJK compatible zone, set in GB 13000.1.

139 ideogram characters used in Taiwan region not collected in GB 2312 but collected in GB 13000.1,

31 other characters collected in GB 13000.1,

Non-Chinese symbols collected in GB 2312,

19 punctuation marks used in vertical alignment, set in GB 12345,

10 lower case Roman numbers not collected in GB 2312.

5 Chinese phonetic letters with tone, not collected in GB 2312 and □ and □,

Chinese number “O”,
13 descriptors of ideogram characters,
80 complementary Chinese characters and radicals/components,
the two-byte coded Euro symbol.

5.3 The four-byte portion

Collected in the four-byte portion of this standard are all characters set in GB 13000.1, including the unified Chinese characters in Extension A of CJK, except for the above two-byte characters.

6. Overall structure

In this standard, one-byte, two-byte and four-byte encoding systems are adopted. In this standard, any byte is composed of 8 binary bits, and any 8-bit value is expressed in hexadecimal annotation, from 0x00 to 0xFF.

In the one-byte portion, encoding structure and rules set in GB 11383 are adopted, using codes from 0x00 to 0x80. In the two-byte portion, two 8-bit binary strings are used to represent one character, whose first byte includes codes from 0x81 to 0xFE, and the end byte includes codes from respectively 0x40 to 0x7E and 0x80 to 0xFE. In the four-byte portion, codes from 0x30 to 0x39 are used as suffix of extended two-byte codes. This extended four-byte codes range from 0x81308130 to 0xFE39FE39, as shown in Table 1 and Fig. 1.

Table 1 Allocation of Code Range

Number of Bytes	Space of Code Positions				Number of Codes
One-byte	0x00-0x80				129 codes
Two-byte	First byte		Second byte		23,940 codes
	0x81~0xFE		0x40~0x7E 0x80~0xFE		
Four-byte	First byte	Second byte	Third byte	Fourth byte	1,587,600 codes
	0x81~0xFE	0x30~0x39	0x81~0xFE	0x30~0x39	

The codes of four-byte characters start from the fourth byte, the code-position is from 0x30 to 0x39; then the third byte, its code-position is from 0x81 to 0xFE; and then the second byte, its code-position is from 0x30 to 0x39; finally, the first byte, its code-position is from 0x81 to 0xFE, namely,

0x81308130 to 0x81308139;
0x81308230 to 0x81308239;
…
0x8130FE30 to 0x8130FE39;
0x81318130 to 0x81318139;
…
0x8131FE30 to 0x8131FE39;
…
0x82308130 to 0x82308139;
…
0x8230FE30 to 0x8230FE39;
…
0xFE308130 to 0xFE308139;
…

0xFE39FE30 to 0xFE39FE39.

Notes: Numbers that preceding by 0x are in hexadecimal system while those without 0x are in decimal system.

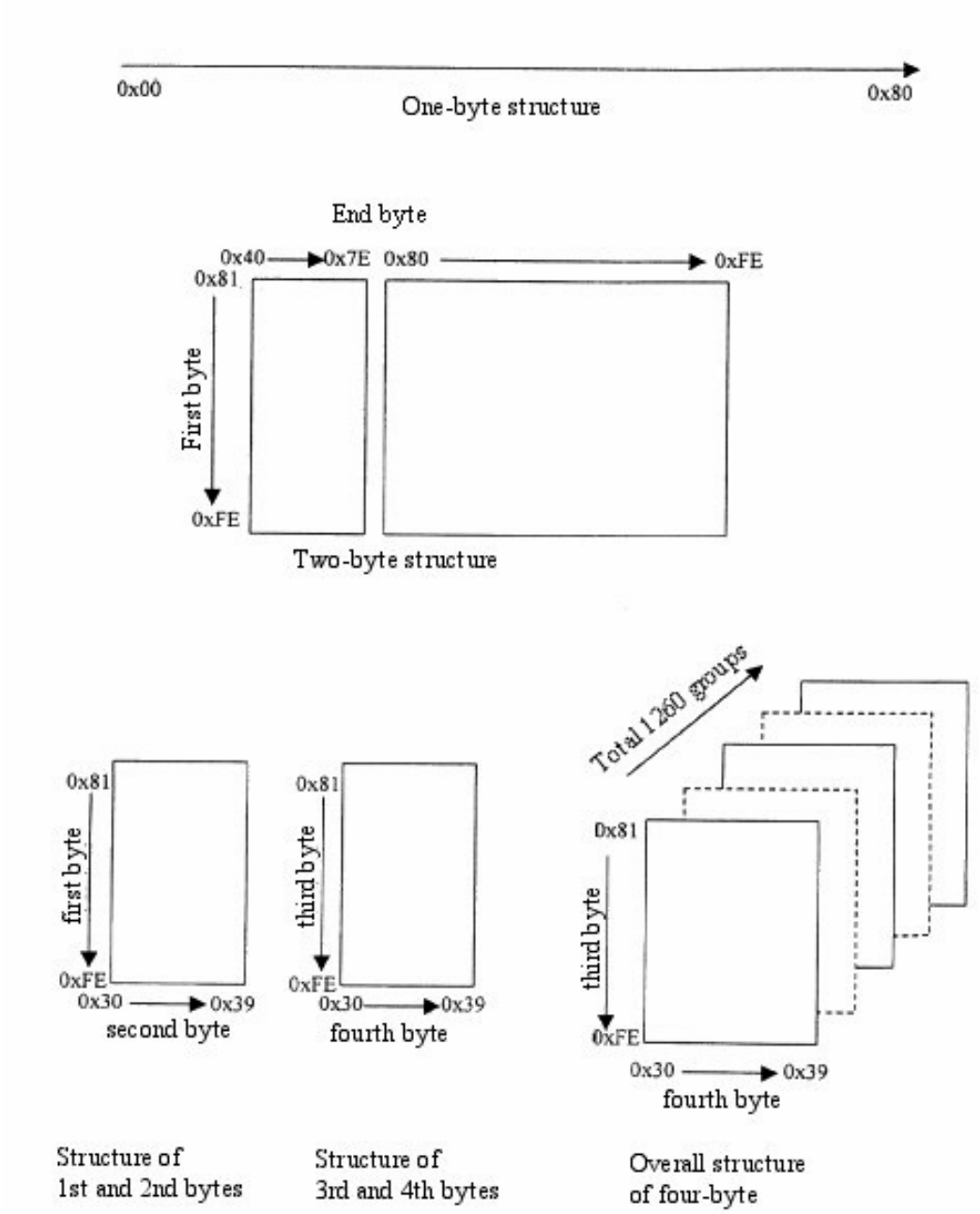

Fig. 1 Diagram of overall structure

7. Code Position Order of Characters

7.1 Order of characters in one-byte portion

In this standard, all characters in one-byte portion are arranged in the order corresponding to the characters set in GB 11383. The one-byte Euro symbol is place on 0x80, corresponding to position
0x20AC in GB 13000.1, as shown in Fig. 2.

7.2 Order of characters in two-byte portion

In this standard, the order of characters in two-byte portion is shown in Appendix A.

7.3 Order of characters in four-byte portion

50,400 codes from 0x81308130 to 0x8439FE39 correspond to all characters set in GB 13000.1 which are not included in the two-byte portion of this standard, they are sorted correspondingly in the order set in GB 13000.1, and the remaining code positions are reserved.

12,600 codes from 0x85308130 to 0x8539FE39 constitute reserved zone of this standard, for future extension of characters.

12,600 codes from 0x86308130 to 0x8F39FE39 constitute reserved zone of this standard for future extension of Chinese characters.

1,058,400 codes from 0x90308130 to 0xE339FE39 are correspondent to the 16 auxiliary plane in GB 13000, the order of characters is completely correspondent to the order in the 16 auxiliary plane in GB 13000, and the remaining code positions are reserved.

315,000 codes from 0xE4308130 to 0xFC39FE39 constitute a reserved zone in this standard, for future extension of standards.

25,200 codes from 0xFD308130 to 0xFE39FE39 is an user definable character (UDC) zone.

8. Allocation of Code Positions

8.1 Allocation of code positions in one-byte portion

For allocation of code positions in one-byte portion in this standard, refer to GB 11383, the Euro Symbol is placed at position 0x80, as shown in Fig. 2.

b8	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
b7	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
b6	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
b5	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
b4 b3 b2 b1	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
0 0 0 0 00			SP	0	@	P	`	p	€							
0 0 0 1 01			!	1	A	Q	a	q								
0 0 1 0 02			"	2	B	R	b	r								
0 0 1 1 03			#	3	C	S	c	s								
0 1 0 0 04			¥	4	D	T	d	t								
0 1 0 1 05			%	5	E	U	e	u								
0 1 1 0 06			&	6	F	V	f	v								
0 1 1 1 07			'	7	G	W	g	w								
1 0 0 0 08			(	8	H	X	h	x								
1 0 0 1 09			)	9	I	Y	i	y								
1 0 1 0 10			*	:	J	Z	j	z								
1 0 1 1 11		ESC	+	;	K	[	k	{								
1 1 0 0 12			´	<	L	\	l	|								
1 1 0 1 13			−	=	M	]	m	}								
1 1 1 0 14			.	>	N	^	n	~								
1 1 1 1 15			/	?	O	_	o	DEL								

Fig. 2 Diagram of code positions in one-byte portion

8.2 Allocation of code positions in two-byte portion

In this standard, arrangement of two-byte characters is respectively from 0x8140 to 0xFE7E and from 0x8180 to 0xFEFE, totally 23,940 code positions. Refer to Fig. 3 and Table 2.

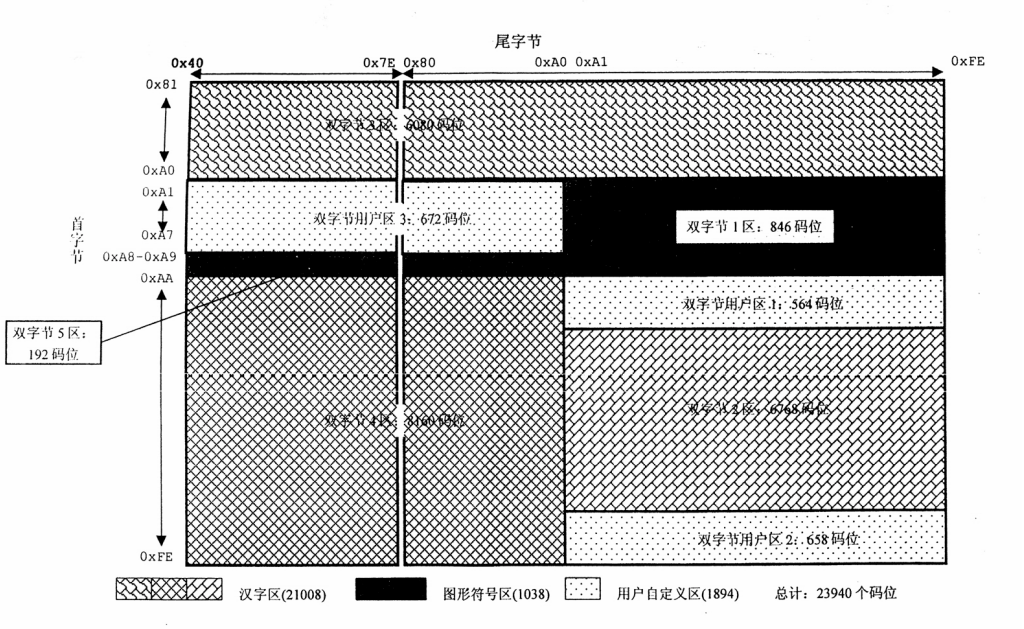

Fig. 3 Diagram of spatial structure of two-byte portion

Table 2 Arrangement of code positions in two-byte portion

Category	Name of zone	Range of code positions	Number of code positions	Number of characters	Type of characters
symbol zone	two-byte zone 1	A1A1-A9FE	846	718	Ideogram and symbol
	two-byte zone 5	A840-A9A0	192	166	Ideogram and symbol
Chinese character zone	two-byte zone 2	B0A1-F7FE	6768	6763	Chinese characters
	two-byte zone 3	8140-A0FE	6080	6080	Chinese characters
	two-byte zone 4	AA40-FEA0	8160	8160	Chinese characters
user definable character zone	two-byte UDC zone 1	AAA1-AFFE	564		
	two-byte UDCzone 2	F8A1-FEFE	658		
	two-byte UDC zone 3	A140-A7A0	672		

In this standard, in the Chinese character zone in two-byte portion (namely the two-byte zones 2, 3, 4), CJK unified Chinese characters are arranged in the front part while complementary Chinese characters are in the rear part. The coded Chinese characters set in GB 2312 are arranged in two-byte zone 2. The 21 CJK compatible coded Chinese characters selected from GB 13000.1 are arranged in two-byte zone 4, from 0xFD9C to 0xFDA0 and from 0xFE40 to 0xFE4F.

Complementary Chinese characters and 80 codes of Chinese radicals/components are arranged in two-byte zone 4.

139 ideogram characters used in Taiwan region collected in GB 13000.1 but excluded by GB 2312, Chinese number “O” and 13 descriptors of ideogram characters are arranged in the two-byte zone 5.

The non-Chinese symbols collected in GB 2312, 5 Chinese phonetic letters with tone (not collected in GB 2312) and □ and □, 10 lower case Roman numbers not collected in GB 2312, 19 punctuation marks used in vertical alignment in GB 12345, and the two-byte coded Euro symbol (at position 0xA2E3) are arranged in two-byte zone 1.

8.3 Allocation of code positions in four-byte portion

Allocation of code positions in four-byte portion is shown in article 7.3.