

San José, February 16, 2001

Feel free to distribute this text (version 1.4) including the author's e-mail address (<mailto:dmeyer@adobe.com>) and to contact him for corrections and additions. Please do not take this text as a literal translation, but as a help to understand the standard GB 18030-2000.

Insertions in brackets [] are used throughout the text to indicate corresponding sections of the published Chinese standard.

Thanks to Markus Scherer (IBM) and Ken Lunde (Adobe Systems) for initial critical reviews of the text.

Major changes & Mapping data information

Since the first writing of this text, quite a few changes have been introduced to the mappings between the GB 18030-2000 “coded character set” and Unicode. The biggest consequence of the changes introduced is probably the complete re-enumeration of all four-byte GB sequences for the Unicode BMP code points starting from U+0080.

Mapping data were re-released in late 2000 that reflect the changes as far as mapping data or larger sections of them are concerned. Based on the mapping data, I added short updates to the document. Where applicable, I indicated modifications through phrases like **now:** or **update:**. However, in order to continuously reflect the history of the standard, often both old and corrected values appear next to each other. Please refer to the latest mapping data (see below) in order to find accurate and up-to-date information.

Currently, developers are waiting for the second print of GB 18030 that is likely to document the changes. While, in this second print, the structural principles of the standard itself can be expected to remain valid, other changes in the mapping data might appear. Many sections of this document are still referring to the first print of the standard. These sections will remain unchanged until a new print will become available. If some sections are lacking sufficient clarity, it is probably the reflection of missing necessary explanations and definitions as they would be provided in a second print of the standard.

A good source to keep track of current and future mappings between GB 18030 and Unicode is the Web site of the open-source “International Components for Unicode” project, hosted on the servers at IBM/Cupertino. You can find the downloadable mapping data in XML format on the ICU Web site at <http://oss.software.ibm.com/icu/charset>. [For a description of the ICU project, see: <http://oss.software.ibm.com/icu/index.html>.]

SUMMARY, EXPLANATIONS, AND REMARKS:

CHINESE NATIONAL STANDARD GB 18030-2000: INFORMATION TECHNOLOGY –
CHINESE IDEOGRAMS CODED CHARACTER SET FOR INFORMATION INTERCHANGE –
EXTENSION FOR THE BASIC SET
(信息技术—信息交换用汉字编码字符集—基本集的扩充 *Xinxi Jishu – Xinxi Jiaohuan
Yong Hanzi Bianma Zifuji – Jibenji De Kuochong*)

March 17, 2000, was the publishing date of the Chinese national standard (国家标准 guojia biao zhun) GB 18030-2000 (hereafter: GB 18030). This standard tries to resolve issues resulting from the advent of Unicode, version 3.0. More specific, it attempts the combination of Unicode's extended character repertoire, namely the Unihan Extension A, with the character coverage of earlier Chinese national standards.

HISTORY

The People's Republic of China had already expressed her fundamental consent to support the combined efforts of the ISO/IEC and the Unicode Consortium through publishing a Chinese National Standard that was code- and character-compatible with ISO 10646-1/Unicode 2.1. This standard was named GB 13000.1. Whenever the ISO and the Unicode Consortium changed or revised their “common” standard, GB 13000.1 adopted these changes subsequently.

In order to remain compatible with GB 2312, however, which at the time of publishing Unicode/GB 13000.1 was an already existing national standard widely used to represent the Chinese “simplified” characters, the “specification” GBK was created. GBK is the second coded character set that contains a character repertoire very similar to that of GB 13000.1, but it uses a completely different encoding. GBK stands short for “*Guojia biao zhun kuozhan*”, the official title is “*Hanzi neima kuozhan guifan* 汉字内码扩展规范”, or: “Rules/Specifications defining the extensions of internal codes for Chinese ideograms” (Note: Throughout this text, the Chinese term “*Hanzi* 汉字” is translated as “Chinese ideograms”).

The significant property of GBK is that it leaves the characters and codes as defined in GB 2312 untouched and positions all additional characters around it. The additional characters are mainly those of the Unified Han portion of Unicode 2.1 that go beyond the character repertoire of GB 2312. Thus, code and character compatibility between GBK and GB 2312 is ensured while, at the same time, the complete Unicode Unified Han character set is made available. At the time when GBK was defined, other characters were added not available in Unicode at that point.

We will later see that both the code space and the character repertoire of GBK create the foundation for the almost completely compatible code space of GB 18030. In fact, there is text in the standard indicating that GB 18030 (the “standard”) is meant to replace (代替 *daiti*) GBK (the “specification”).

The two-byte code space allocated in GBK is defined as follows:

Standard area for symbols (符号标准区 *fu hao biao zhun qu*):

- GBK/1: 0xa1a1-0xa9fe (846 codes / 717 graphic symbols)
- GBK/5: 0xa840-0xa9a0 (192 / 166 graphic symbols)

Standard area for Chinese ideograms (汉字标准区 *Hanzi biao zhun qu*):

- GBK/2: 0xb0a1-0xf7fe (6,768 / 6,763 Hanzi)
- GBK/3: 0x8140-0xa0fe (6,080 / 6,080 Hanzi)
- GBK/4: 0xaa40-0xfea0 (8,160 / 8,160 Hanzi)

User-defined area (用户自定义区 *yong hu zi ding yi qu*):

- UDA 1: 0xaa1-0xffe (564 / 0)
- UDA 2: 0xf8a1-0xfefe (658 / 0)
- UDA 3: 0xa140-0xa7a0 (672 / 0)

Thus, GBK defines 23,940 code points containing 21,886 characters. At the same time, GBK provides mappings to the code points of Unicode 2.1. Those characters in GBK that, at the time of its publishing, were not included in Unicode are mapped to Unicode's Private Use Area (PUA) starting at code point U+e000. The same is true for those code points that remain empty in GBK.

When looking at the packed code space used to define GBK, it becomes obvious that there is no space left for a major addition. The 1,894 code points of GBK's three user-defined areas are not even close to provide sufficient space for the Unihan Extension A, which defines 6,582 new characters in plane 0 of Unicode, version 3.0, the Basic Multilingual Plane (BMP, this term originates in ISO 10646, but is now used in Unicode context, too).

RECENT DEVELOPMENTS

Per its definition, GB 18030 is a "coded character set" (编码字符集 *bian ma zi fu ji*) in that it defines not only its character repertoire, but also standardizes the characters' code points.

The significant properties of GB 18030 are:

- It incorporates Unicode's Unihan Extension A completely.
- It provides code space for all used and unused code points of Unicode's plane 0 (BMP) and its 16 additional planes, if these code points were not already included in GBK. Expressed differently: While being a code- and character compatible "superset" of GBK, GB 18030, at the same time, intends to provide space for all remaining code points of Unicode. Thus, it effectively creates a 1-to-1 relationship between parts of GB 18030 and Unicode's complete encoding space.
- In order to accomplish the Unihan incorporation and code space allocation for Unicode 3.0, GB 18030 defines and applies a four-byte encoding mechanism.

To illustrate the code mapping as described in the second point, here some examples how this is implemented in GB 18030:

- Looking at the standard's tables that show the mappings between Unicode and GB 18030, we can find the first gap in Unicode code points being mapped between 0x81308434 (**now:** 0x81308435) and 0x81308435 (**now:** 0x81308436). The omitted Unicode value is U+00a4 (currency sign). We find this character at GB 18030 code point 0xa1e8 (where it was positioned in GB 2312, GBK, and now in GB 18030).
- Two more significant gaps can be identified in the four-byte GB 18030 to Unicode mapping area. As could be expected, one of them reflects the fact that Unicode's CJK Unihan character repertoire (U+4e00 through U+9fa5) is the major part of GBK/GB 18030: U+4dff is mapped to 0x82358f33 (**now:** 0x82358F32), U+9fa6 is mapped to 0x82358f34 (**now:** 0x82358F33).

The other significant mapping gap is related to Unicode's Private Use Area (U+e000 through U+ffff). The last Unicode code point before the PUA U+dfef is mapped to 0x83389837 (**now:** U+d7ff maps to 0x8336C738, mappings to the surrogate area are not included anymore), the first one of those not used in the non-four-byte areas of GB 18030 U+e865 is mapped to 0x83389838 (**now:** 0x8336D030). The PUA code points in-between can be found in the following areas of GB 18030:

U+e000-U+e765: two-byte (user) area 1, 2, 3 (0xaa-0xaf, 0xf8-0xfe, 0xa1-0xa7)

U+e766-U+e7bb: two-byte area 1 (0xa1-0xa7)

U+e7bc-U+e7c6: two-byte area 5 (0xa8)

U+e7c7-U+e7e1: two-byte area 1 (0xa8)

U+e7e2-U+e7fd: two-byte area 5 (0xa9)

U+e7fe-U+e80f: two-byte area 1 (0xa9)

U+e810-U+e814: two-byte area 2 (0xd7) and

U+e815-U+e864: two-byte (user) area 4 (0xfe).

Update: Actually, a third large gap can be found in the re-released mapping data now that is caused by the removal of the Unicode Surrogate code points (U+d800 through U+dfef) from the four-byte mapping area.

SUMMARY OF THE MAIN PARTS OF GB 18030

The preface of the standard mentions:

- that GB 18030 is an extension of GB 2312;
- that GB 18030 replaces (代替 *daiti*) GBK, version 1.0;
- that GB 18030 was published by the People's Republic of China's Ministry of the Information Industry;
- which institutions and companies contributed to the drafting of this standard, namely: the Research Institute for Standardization in the Electronics Industry at the Ministry of the Information Industry, the Research Institute of Computer Technology at the Peking University, the Peking University Fangzheng Group, the Beijing Fangzheng Xin Tiandi Information and Networking Science and Technology Incorporated, the company Stone (*Sitong*) Group,

the Software Institute at the Chinese Academy of Sciences,
the company Changcheng Software,
the company Stone (*Sitong*) Lifang,
the companies Chinese Software, Jin Shan, and Lian Xiang.

Individual contributors mentioned are:

Chen Kunqiu, Huang Jiang, Hu Wanjin, Zhang Jianguo, and Chen Zhuang. [Preface]

The standard's main text defines:

1) Referenced standards are

- GB 2311-1990 (equivalent to ISO 2022:1986), “Information processing – ISO 7-bit and 8-bit coded character sets – Code extension techniques),
- GB 2312-1980 Basic Set, “Information processing – Chinese ideograms coded character set for information interchange – Basic set”,
- GB 11383-1989 (identical to ISO 4873:1986), “Information processing – 8-bit code for information interchange – Structure and rules for implementation”,
- GB 12345-1990, “Information processing – Code of Chinese ideogram set for information interchange – Supplementary set”, and
- GB 13000.1-1993 (identical to ISO/IEC 10646.1-1993), “Information technology – Universal multiple-octet coded character set (UCS) – Part 1: Architecture and Basic Multilingual Plane”. [2]

Again, being a “specification”, not a “standard”, GBK is not mentioned in this list.

2) The two guiding principles of GB 18030 are that it shall remain “encoding standard compatible” (内码标准兼容 *neima biaoazhun jianrong*) with GB 2312 and, with regard to the characters collected, that it shall “completely support all characters of the CJK Unified Hanzi character set of GB 13000.1 as well as all characters of the CJK Unified Hanzi Extension A”. [3]

3) Definition of terms

- the Chinese term “字汇 *zihui*” means “repertoire”;
- the Chinese term “字符 *zifu*” means “character”;
- the Chinese term “编码字符 *bianma zifu*” means “coded character”;
- the Chinese term “保留区 *baoliu qu*” means “reserved zone”. [4]

4) The character repertoire

a) in the single-byte area:

- all 128 characters as defined on positions 0x00 through 0x7f in GB 11383, plus the single-byte encoded currency sign “Euro” on position 0x80 (**now:** the “Euro” has been removed from that position, U+0080 maps to the first available code point of the four-byte area);

b) in the two-byte area:

- all CJK Unified Hanzi characters as defined in GB 13000.1;
- 21 selected Hanzi characters from the CJK compatibility area of GB 13000.1;

- 139 graphical characters as used in Taiwan and included in GB 13000.1, but not included in GB 2312;
- 31 additional characters included in GB 13000.1;
- the non-Hanzi symbols of GB 2312;
- 19 vertical punctuation symbols of GB 12345;
- 10 lowercase Roman numerals not included in GB 2312;
- 5 characters of the phonetic Hanyu Pinyin alphabet, plus the letters a and g, all of which were not included in GB 2312;
- the Chinese numeral “zero”;
- 13 ideographic description symbols (表意文字描述符 *biaoyi wenzi miaoshu fu*);
- 80 additional Hanzi and radicals (部首 *bushou*) or components (构件 *goujian*);
- the two-byte encoded currency symbol Euro.

With regard to the two-byte character repertoire as defined above in b), GB 18030 is effectively a full superset of GBK.

c) in the four-byte area:

- all characters of the CJK Unified Han Extension A as it exists in GB 13000.1, with the exception of those two-byte encoded characters that have already been described above. [5]

This basically implies that there will be no duplication of code points, a unique “entry” point for each character has been created or, in other words, the chance to create 1-to-n or n-to-1 mappings has been avoided.

5) Overall code structure

The standard applies a single-, two-, and four-byte character encoding method.

The single-byte portion applies the coding structure and principles of GB 11383 by using the code points 0x00 through 0x80 (**now:** 0x7f).

The two-byte portion uses two eight-bit binary sequences to express a character. The code points of the first (leading) byte range from 0x81 through 0xfe, the code points of the second (trailing) byte ranges from 0x40 through 0x7e and from 0x80 through 0xfe.

The four-byte portion uses the code points 0x30 through 0x39, which are vacant in GB 11383, as an additional means to extend the two-byte encodings, thus effectively increasing the number of four-byte codes to now include code points ranging from 0x81308130 through 0xfe39fe39 (as shown in Table 1 and Illustration 1). [6]

Table 1 and Illustration 1 of the standard illustrate the ordering and the counting sequence of the possible four-byte combinations, each of the bytes coming from one of four ranges that indicate the byte’s allowed value: 0x81 through 0xfe for the first and the third byte, 0x30 through 0x39 for the second and fourth byte. In order to achieve the intended code sequences, initially, the value of the fourth byte is increased, then the value of the third, the second, and the first.

Coverage of the code space as defined by the standard and expressed through four bytes:

```

0x81308130 – 0x81308139,
0x81308230 – 0x81308239,
...
0x8130fe30 – 0x8130fe39,
0x81318130 – 0x81318139,
...
0x8131fe30 – 0x8131fe39,
...
0x82308130 – 0x82308139,
...
0x8230fe30 – 0x8230fe39,
...
0xfe308130 – 0xfe30fe39,
...
0xfe39fe30 – 0xfe39fe39.

```

The number of code points created through applying these principles is 128 for the single-byte portion of GB 18030, 23,940 for the two-byte portion, and 1,587,600 for the four-byte portion. [6]

6) Sequential order of characters / Code point allocation

- a) The characters of the single-byte portion and their sequence is identical to the respective characters of GB 11383. The single-byte encoded currency symbol Euro is added at position 0x80 to represent position 0x20ac of GB 13000.1 (see Illustration 2). [7.1, 8.1] (**Now:** The currency symbol Euro has been removed from that position. The fullwidth Euro 0xa2e3 now maps to U+20ac)
- b) The characters of the two-byte portion are sequentially ordered as illustrated in Appendix A (as mentioned before, this sequential order is virtually identical to that of GBK, with the exception of the two-byte encoded version of the currency symbol Euro, increasing the character count in the two-byte area 1 – 0xa1a1-0xa9fe – by one (Euro) to 718; see Illustration 3 and Table 2). [7.2].

Listed here for completeness, the two-byte code space of GB 18030:

Standard area for symbols (符号标准区 *fu hao biao zhun qu*):

- Two-byte area 1: 0xa1a1-0xa9fe (846 codes / 718 graphic symbols, including the Euro at 0xa2e3, mapped to U+e76c, **now:** the Euro now maps to U+20ac)
- Two-byte area 5: 0xa840-0xa9a0 (192 / 166 graphic symbols)

Standard area for Chinese ideograms (汉字标准区 *Hanzi biao zhun qu*):

- Two-byte area 2: 0xb0a1-0xf7fe (6,768 / 6,763 Hanzi)
- Two-byte area 3: 0x8140-0xa0fe (6,080 / 6,080 Hanzi)
- Two-byte area 4: 0xaa40-0xfea0 (8,160 / 8,160 Hanzi)

User-defined area (用户自定义区 *yonghu zidingyi qu*):

- UDA 1: 0xaaa1-0xaffe (564 / 0)
- UDA 2: 0xf8a1-0xfefe (658 / 0)
- UDA 3: 0xa140-0xa7a0 (672 / 0)

These three areas contain a total of 23,940 code points located in the two code spaces ranging from 0x8140 through 0xfe7e and 0x8180 through 0xfefe, used by 21,887 characters.

In more detail, GB 18030 lists the following special groups of characters and their locations:

- the two-byte areas 2, 3, and 4 contain the CJK Unified Han characters first, they are followed by additional Hanzi;
- the encoded Hanzi of GB 2312 occupy the two-byte area 2;
- the code points 0xfd9c through 0xfda0 and 0xfe40 through 0xfe4f of two-byte area 4 contain 21 CJK compatibility characters selected from GB 13000.1;
- 80 additional Chinese characters and radicals or character components are also encoded in two-byte area 4;
- 139 graphical characters used in Taiwan and included in GB 13000.1, but not included in GB 2312, as well as the Chinese numeral “zero”, and 13 ideographic description characters are encoded in two-byte area 5.
- the non-Hanzi symbols of GB 2312, 5 characters of the phonetic Hanyu Pinyin alphabet, plus the letters a and g, all of which not included in GB 2312, 10 lowercase Roman numerals not included in GB 2312, 19 vertical punctuation symbols of GB 12345, and the currency symbol Euro are all encoded in two-byte area 1. [8.2]

The sequential ordering of the characters in the four-byte encoded is as follows:

- 50,400 code points ranging from 0x81308130 through 0x8439fe39 serve as code positions for the not yet included two-byte encoded characters of GB 13000.1, their sequential order is according to GB 13000.1, remaining code points will be reserved (剩余码位保留 *shengyu mawei baoliu*);
- 12,600 code points ranging from 0x85308130 through 0x8539fe39 represent a zone reserved for the use through future character extensions;
- 126,000 code points ranging from 0x86308130 through 0x8f39fe39 represent a zone reserved for the use through Hanzi character extensions;
- 1,058,400 code points ranging from 0x90308130 through 0xe339fe39 will be used to include the 16 additional planes of GB 13000.1, the sequential order of these code points will completely honor the respective code point sequences and ordering of the 16 additional planes of GB 13000, remaining code points will be reserved;
- 315,000 code points ranging from 0xe4308130 through 0xfc39fe39 represent a zone reserved for the use of future extensions of the standard;
- 25,200 code points ranging from 0xfd308130 through 0xfe39fe39 serve as a user-defined area. [7.3, 8.3]

7) Appendices

- Appendix A shows a table of the two-byte encoded characters and their Unicode values.
- Appendix B lists the ideographic description characters that were formerly encoded using code values from the PUA. Their code points are now located in the two-byte code area and range from GB 18030 0xa989 through 0xa995.
- Appendix C lists additional Hanzi and radicals or character components located in the two-byte code area ranging from 0xfe50 through 0xfea0, Unicode values are listed.
- Appendix D shows the complete range of the four-byte encoded characters and their respective Unicode values. The area occupied by the Unihan Extension A shows the characters in the code space ranging from 0x8139ef30 through 0x82358739 (**now:** U+3400/0x8139EE39 through U+4dff/0x82358F32) (the overall number of BMP code points here is 6,530, which equals the 6,582 characters of Unihan A minus 52 of its characters already covered in the two-byte area).
- Appendix E lists characters formerly encoded in GBK that were mapped to code points in Unicode's Personal Use Area. The encodings of these characters changed because they are now included in Unicode 3.0.

An explanation at the beginning of Appendix E mentions that Appendix A of GB 18030 (the list of all double-byte encoded characters) – for reasons of compatibility with its predecessor GBK – still lists mappings to the so-called “temporary codes” of GB 13000.1 (as they were included in GBK).

The characters affected are the 13 ideographic description characters, the letter “lowercase n grave”, 52 additional Chinese characters, and 13 Chinese radical variants (**now:** 14 Chinese radical variants: U+2e97 is part of Unicode's CJK Radicals Supplement and considered equivalent to GB 18030's 0xfe5e.) Listed are their earlier code points (“temporary codes” 临时代码 *linshi daima*) in GB 13000.1/Unicode and their current code points in GB 18030/Unicode 3.0.

8) Additional remarks

a) Typographical Errors

- 1) Page 294 (Appendix E): 0xa987 is listed to be the code point for the “Ideographic Variation Indicator”, the correct position should be 0xa989. This is correct in Appendix B.
- 2) Page 82 (Appendix A): For code point 0xa98a a Unicode mapping of U+e7e6 is listed. This represents a duplication of the mapping for 0xa95f, the correct value for 0xa98a should be U+e7e8 (compare Appendix E).
- 3) Page 11 (Appendix A): For the code point sequence 0xa5fb through 0xa5fd Unicode mappings of U+e681 through U+e683 are listed. This represents a duplication of the mappings for 0xa57b through 0xa57d, the correct values for 0xa5fbff should be U+e781 through U+e783 (cp. Appendix E).

b) The “added” Variation Indicator mapping / The “missing” 0x80 mapping

On page 149 of Appendix D, which lists the mappings of the standard's four-byte area to Unicode, we find the mapping U+303e for 0x8139a634. This contradicts the general remarks made in Appendix E for the 79 characters that are now part of the BMP.

All of these characters appear in the two-byte area, their Unicode code points should not appear in the four-byte area as mapped. This particular mapping seems to have been added erroneously.

Update: The former mapping from U+303e to 0x8139a634 has been corrected.

Given the standard's code point count as expected for the four-byte area (41,388, see the following remarks below) logic tells us that there is probably a Unicode mapping missing somewhere. And indeed, the standard does not provide such a mapping for the code point U+0080. As for a possible explanation: The single-byte area has been extended to include 0x80 (the "Euro" sign). It is precisely this inclusion of 0x80, which makes GB 18030 leave the coverage of GB 11383/ISO 4873 behind and – at the same time – become incompatible to Unicode at this code point. The creators of GB 18030 realized this: There is no mapping provided for U+0020AC (the "Euro" sign) in the four-byte area, which leads to the assumption that this mapping is implicitly reserved for 0x80. Hence, a mapping for Unicode U+0080 should have been included as the first mapping entry for code point 0x81308130. The question remains to be answered which mapping from U+0080 to the four-byte area should be created to fill this "hole". A "natural" solution would be to assign U+0080 to the first available four-byte code point after the code point used to map U+ffff, i.e. 0x8432eb38. This way, the complete re-mapping of the four-byte area could be avoided (U+0080 would have to be the first character of it, mapped to 0x81308130).

In this context, it should be allowed to add the general remark that explicit mappings for GB 18030's single-byte area to Unicode would have been helpful.

Update: The code point 0x80 (occupied by the "Euro" sign) has been removed from the standards's one-byte area. As could be expected, an empty code U+0080 is now the first code point that maps to GB 18030's four-byte area, starting at 0x81308130. Of course, this had a significant influence on the subsequent mappings, which all had to be re-enumerated.

c) Appendix E (pp. 294ff.)

Leaving aside the situation as described in b), which effectively does not change the overall count of code points in the four-byte area ("one added, one omitted") another observation.

The BMP has 65,536 code points. The tables on pages 94 through 293 of GB 18030 show 41,388 BMP code points in the four-byte area, additional BMP code points can be found in the one-byte (129) and two-byte area (23,940), totaling 65,457. The 79 code points seemingly "missing" to completely cover the BMP are "hidden" in Appendix E. Neither the "temporary" nor the "final codes" of the 79 characters listed in Appendix E are part of the mappings in Appendix D (the four-byte area). The characters themselves appear in two-byte area 5 (0xa989 through 0xa995), there mapped to their "temporary" codes, consistent with the remarks in Appendix E. This explains the difference between the total number of code points in the BMP and the actual number of code points used in GB 18030 to cover the BMP.

Again, remarks c) and d) go hand in hand, they are examined separately in these two paragraphs for reasons of clarity.

d) Illustration 2 of GB 18030 (p. 6)

The illustration shows the one-byte code space of GB 18030 (0x00 through 0x80). It is unclear why the currency symbol “yuan”/“yen” is used to represent the code position 0x24. Throughout GB 18030 the standard GB 11383 is quoted as the standard that serves as the model how this particular code space shall be used, code- as well as character-wise (see [5] of GB 18030, describing the character repertoire). Looking at the original GB 11383-1989, however, shows that apparently – at least its 1989 edition – no decision was made as to which character should represent code point 0x24. Instead, a separate paragraph (7.4.2) explains that either the “yuan sign” or the “currency sign” might be used, depending on intention or context. (Even the “dollar sign” is mentioned as potential character representing the code point.)

There is reason for the assumption that the “yen sign” better be replaced by the “dollar sign”. We find the “fullwidth dollar sign” at 0xa1e7, mapped to U+ff04, and the “fullwidth yen sign” at 0xa3a4, mapped to U+ffe5, both seems correct. In addition to that, we find the Unicode mapping for the “yen sign” U+00a5 used to map GB 18030’s 0x81308435. No mapping for U+0024, the “dollar sign” can be found. Given that only those characters should appear in the four-byte area that have not been included in the one- or two-byte areas, this supports the assumption that the “dollar sign” might be the intended representation of 0x24 in the context of GB 18030. GB 18030 stresses that GB 11383 is identical to ISO 4873:1986. However, ISO 4873 underwent changes since 1986. It is now identical to recent versions of ECMA 43, which – in its current version and in that code space (G0) – is defined as being identical to US-ASCII with the “dollar sign” at U+0024.

Update: There is reason to assume that it will indeed be the Dollar sign which will have code point 0x24 as its “home”.

e) Question: Is it necessary to encode Unicode’s surrogate area (U+d800 through U+dfef) in the four-byte encoding space of GB 18030?

As described, GB 18030 provides one-, two-, and four-byte code points for the complete Unicode character repertoire. This includes Unicode’s surrogate area in GB 18030’s four-byte code range from 0x8336c830 through 0x83389837. There are at least two reasons that make this attempt seem redundant:

Firstly, GB 18030 provides sufficient code points to achieve the same result that the surrogate pairs were intended to achieve.

Secondly, using two of GB 18030’s four-byte codes as pairs in the same way the “high” and “low” surrogates of Unicode are intended to be used, will effectively lead to 8-byte character codes. Even inside GB 18030, an encoding result identical to that of using surrogate pairs can be achieved with the help of only four bytes. Moreover, it seems to be against the standard’s own specification, if 8-byte codes are used to represent a character. The question seems void whether Unicode’s surrogate area should appear in GB 18030’s four-byte code space at all.

Update: The re-released mapping data indicate that mappings to Unicode’s surrogate area will not be included in future versions of the standard. No mappings for U+d800 through U+dfef can be found anymore in the four-byte encoding space of GB 18030.

Version history:

- 1.1: First public version
- 1.2: Corrections in the paragraphs describing the code space of GBK2K
- 1.3: Minor corrections, adding section e) in the Additional Remarks section
- 1.4: Updated version, minor corrections, describing the major changes in the re-released mapping data, added graphic explaining GB 18030-2000's code space, replaced GBK2K with GB 18030