GB 18030

GB 18030 is a Chinese government standard described as *Information technology—Chinese coded character set* and defines the required language and character support necessary for software in China. **GB18030** is the registered Internet name for the official character set of the People's Republic of China (PRC) superseding GB2312.[1] As a Unicode Transformation Format[a] (i.e. an encoding of all Unicode code points), it is compatible with legacy encodings including GB2312, CP936,[b] and GBK 1.0, GB18030 supports both simplified and traditional Chinese characters.

In addition to the "GB18030 character encoding", this standard contains requirements about which scripts must be supported, font support, etc.[2]

GB 18030

MIME	GB18030
Alias(es)	Code page 54936
Standard	GB 18030-2005, GB 18030-2000
Language(s)	zh
Preceded by	GBK, GB2312

Contents

History

The GB18030 character set is formally called "Chinese National Standard GB 18030-2005: Information technology—Chinese coded character set". **GB** abbreviates Guójiā Biāozhǔn (国家标准), which means *national standard* in Chinese. The standard was published by the China Standard Press, Beijing, November 8, 2005. Only a portion of the standard is mandatory.[2] Since May 1, 2006, support for the mandatory subset is officially required for all software products sold in the PRC.

An older version of the standard, known as "Chinese National Standard GB 18030-2000: Information Technology—Chinese ideograms coded character set for information interchange—Extension for the basic set", was published on March 17, 2000. The encoding scheme stays the same in the new version, and the only difference in GB-to-Unicode mapping is that GB 18030-2000 mapped the character `A8 BC` (ḿ) to a private use code point U+E7C7, and character `81 35 F4 37` (without specifying any glyph) to U+1E3F (ḿ), whereas GB 18030-2005 swaps these two mapping assignments.[3]:534 More code points are now associated with characters due to update of Unicode, especially the appearance of CJK Unified Ideographs Extension B. Some characters used by ethnic minorities in China such as Mongolian characters and Tibetan characters (GB 16959-1997 and GB/T 20542-2006), have been added as well, which accounts for the renaming of the standard.

Different Unicode mappings between GB 18030 versions

GB byte sequence	Unicode code point	
	GB 18030-2000	GB 18030-2005
A8 BC (ḿ)	U+E7C7	U+1E3F ḿ
81 35 F4 37	U+1E3F ḿ	U+E7C7

Compared with its ancestors, GB 18030's mapping to Unicode has been modified for the 81 characters that were provisionally assigned a Unicode Private Use Area code point (U+E000–F8FF) in GBK 1.0 and that have later been encoded in Unicode.[4] This is specified in Appendix E of GB 18030.[3]:534[5]:499 There are 24 characters in GB 18030-2005 that are still mapped to Unicode PUA.[6]

Private use characters in GB-to-Unicode mappings

GB byte sequence	Unicode code point (blue = private use)		
	GBK 1.0[7][3]:534	GB 18030-2005[6]	Unicode 4.1
A6 D9[8]:108		U+E78D	U+FE10 ’
A6 DA		U+E78E	U+FE12 °
A6 DB		U+E78F	U+FE11 `
A6 DC		U+E790	U+FE13 :
A6 DD		U+E791	U+FE14 ;
A6 DE		U+E792	U+FE15 !
A6 DF		U+E793	U+FE16 ?
A6 EC		U+E794	U+FE17 ﹗
A6 ED		U+E795	U+FE18 ﹘
A6 F3		U+E796	U+FE19 ⋮
A8 BC	U+E7C7	U+1E3F ḿ	
A8 BF	U+E7C8	U+01F9 ǹ	
A9 89	U+E7E7	U+303E 〾	
A9 8A	U+E7E8	U+2FF0 ⿰	
A9 8B	U+E7E9	U+2FF1 ⿱	
A9 8C	U+E7EA	U+2FF2 ⿲	
A9 8D	U+E7EB	U+2FF3 ⿳	
A9 8E	U+E7EC	U+2FF4 ⿴	
A9 8F	U+E7ED	U+2FF5 ⿵	
A9 90	U+E7EE	U+2FF6 ⿶	
A9 91	U+E7EF	U+2FF7 ⿷	
A9 92	U+E7F0	U+2FF8 ⿸	
A9 93	U+E7F1	U+2FF9 ⿹	
A9 94[8]:173	U+E7F2	U+2FFA ⿺	
A9 95	U+E7F3	U+2FFB ⿻	
FE 50	U+E815	U+2E81 ⺁	
FE 51	U+E816		U+20087 𠂇
FE 52	U+E817		

		U+20089 𠂉
FE 53	U+E818	U+200CC 𠃌
FE 54	U+E819	U+2E84 ⺄
FE 55	U+E81A	U+3473 㑳
FE 56	U+E81B	U+3447 㑇
FE 57	U+E81C	U+2E88 ⺈
FE 58	U+E81D	U+2E8B ⺋
FE 59	U+E81E	U+9FB4 龴
FE 5A	U+E81F	U+359E 㖞
FE 5B	U+E820	U+361A 㘚
FE 5C	U+E821	U+360E 㘎
FE 5D	U+E822	U+2E8C ⺌
FE 5E	U+E823	U+2E97 ⺗
FE 5F	U+E824	U+396E 㥮
FE 60	U+E825	U+3918 㤘
FE 61	U+E826	U+9FB5 龵
FE 62	U+E827	U+39CF 㧏
FE 63	U+E828	U+39DF 㧟
FE 64	U+E829	U+3A73 㩳
FE 65	U+E82A	U+39D0 㧐
FE 66	U+E82B	U+9FB6 龶
FE 67	U+E82C	U+9FB7 龷
FE 68	U+E82D	U+3B4E 㭎
FE 69	U+E82E	U+3C6E 㱮
FE 6A	U+E82F	U+3CE0 㳠
FE 6B	U+E830	U+2EA7 ⺧
FE 6C	U+E831	U+215D7 𡗗
FE 6D	U+E832	U+9FB8 龸
FE 6E	U+E833	U+2EAA ⺪
FE 6F	U+E834	U+4056 䁖
FE 70	U+E835	U+415F 䅟
FE 71	U+E836	U+2EAE ⺮

FE 72	U+E837	U+4337 纷
FE 73	U+E838	U+2EB3 𠂢
FE 74	U+E839	U+2EB6 羊
FE 75	U+E83A	U+2EB7 羊
FE 76	U+E83B	U+2298F 戋
FE 77	U+E83C	U+43B1 糶
FE 78	U+E83D	U+43AC 糕
FE 79	U+E83E	U+2EBB 耒
FE 7A	U+E83F	U+43DD 胯
FE 7B	U+E840	U+44D6 劳
FE 7C	U+E841	U+4661 褵
FE 7D	U+E842	U+464C 褪
FE 7E	U+E843	U+9FB9 龷
FE 80	U+E844	U+4723 讶
FE 81	U+E845	U+4729 谦
FE 82	U+E846	U+477C 賵
FE 83	U+E847	U+478D 赙
FE 84	U+E848	U+2ECA 𠫊
FE 85	U+E849	U+4947 鏐
FE 86	U+E84A	U+497A 钑
FE 87	U+E84B	U+497D 钹
FE 88	U+E84C	U+4982 镉
FE 89	U+E84D	U+4983 镏
FE 8A	U+E84E	U+4985 镱
FE 8B	U+E84F	U+4986 镼
FE 8C	U+E850	U+499F 閟
FE 8D	U+E851	U+499B 閞
FE 8E	U+E852	U+49B7 阭
FE 8F	U+E853	U+49B6 阮
FE 90	U+E854	U+9FBA 卓
FE 91	U+E855	U+241FE 𤇾
FE 92	U+E856	

		U+4CA3 齾
FE 93	U+E857	U+4C9F 鲏
FE 94	U+E858	U+4CA0 鲝
FE 95	U+E859	U+4CA1 鲡
FE 96	U+E85A	U+4C77 齾
FE 97	U+E85B	U+4CA2 鳣
FE 98	U+E85C	U+4D13 鸆
FE 99	U+E85D	U+4D14 鸇
FE 9A	U+E85E	U+4D15 鸉
FE 9B	U+E85F	U+4D16 鸈
FE 9C	U+E860	U+4D17 鸗
FE 9D	U+E861	U+4D18 鸘
FE 9E	U+E862	U+4D19 鸝
FE 9F	U+E863	U+4DAE 龑
FE A0	U+E864	U+9FBB 纞

As a national standard

The mandatory part of GB 18030-2005 consists of 1 byte and 2 byte encoding, together with 4 byte encoding for CJK Unified Ideographs Extension A. The corresponding Unicode code points of this subset, including provisional private assignments, lie entirely in the BMP.[3]:3 These parts correspond to the fully mandatory GB 18030-2000.[2]:2

Most major computer companies had already standardised on some version of Unicode as the primary format for use in their binary formats and OS calls. However, they mostly had only supported code points in the BMP originally defined in Unicode 1.0, which supported only 65,536 codepoints and was often encoded in 16 bits as UCS-2.

In a move of historic significance for software supporting Unicode, the PRC decided to mandate support of certain code points outside the BMP. This means that software can no longer get away with treating characters as 16 bit fixed width entities (UCS-2). Therefore, they must either process the data in a variable width format (such as UTF-8 or UTF-16), which are the most common choices, or move to a larger fixed width format (such as UCS-4 or UTF-32). Microsoft made the change from UCS-2 to UTF-16 with Windows 2000.

Mapping

GB 18030 defines a one (ASCII), two (extended GBK), or four-byte (UTF) encoding. The two-byte codes are defined in a lookup table, while the four-byte codes are defined sequentially (hence algorithmically) to fill otherwise unencoded parts in UCS. GB 18030 inherits the bad aspects of GBK, most notably needing special code to safely find ASCII characters in a GB18030 sequence.

GB 18030 encoding[3]:3[5]:252[9]

GB 18030				code points[c]	Unicode
byte 1 (MSB)	byte 2	byte 3	byte 4		
00 – 7F				128	0000 – 007F
80				—	invalid[d]
81 – FE	40 – FE except 7F[e]			23 940	0080 – FFFF except D800 – DFFF[f]
81 – 84	30 – 39	81 – FE	30 – 39	39 420	
85				— (12 600)	*reserved for future character extension*
86 – 8F				— (126 000)	*reserved for future ideographic extension*
unassigned				—	D800 – DFFF[g]
90 – E3	30 – 39	81 – FE	30 – 39	1 048 576	1 0000 – 10 FFFF
E4 – FC				— (315 000)	*reserved for future standard extension*
FD – FE				— (25 200)	*user-defined*
FF				—	invalid
Total				**1 112 064**	

The one- and two-byte code points are essentially GBK with the euro sign, PUA mappings for unassigned/user-defined points, and vertical punctuations. The four byte scheme can be thought of as consisting of two units, each of two bytes. Each unit has a similar format to a GBK two byte character but with a range of values for the second byte of 0x30–0x39 (the ASCII codes for decimal digits). The first byte has the range 0x81 to 0xFE, as before. This means that a string search routine that is safe for GBK should also be reasonably safe for GB18030 (in much the same way that a basic byte-oriented search routine is reasonably safe for EUC).

This gives a total of 1,587,600 (126×10×126×10) possible 4 byte sequences, which is easily sufficient to cover Unicode's 1,112,064 (17×65536 − 2048 surrogates) assigned, reserved, and noncharacter code points.

Unfortunately, to further complicate matters there are no simple rules to translate between a 4 byte sequence and its corresponding code point. Instead, codes are allocated sequentially (with the first byte containing the most significant part and the last the least significant part) **only** to Unicode code points that are not mapped in any other manner. For example:

```
U+00DE (Þ) → 81 30 89 37
U+00DF (ß) → 81 30 89 38
U+00E0 (à) → A8 A4
U+00E1 (á) → A8 A2
U+00E2 (â) → 81 30 89 39
U+00E3 (ã) → 81 30 8A 30
```

An offset table is used in the WHATWG and W3C version of GB 18030 to efficiently translate code points.[10] ICU[9] and glibc use similar range definitions to avoid wasting space on large sequential blocks.

Support

Encoding

Windows 2000 can support the GB18030 encoding if GB18030 Support Package[11] is installed. Windows XP can support it natively The open source PostgreSQL database supports GB18030 through its full support for UTF-8, i.e. by converting it to and from UTF-8. Similarly Microsoft SQL Server supports GB18030 by conversion to and from UTF-16.

More specifically, supporting the GB18030 encoding on Windows means that **Code Page 54936** is supported by `MultiByteToWideChar` and `WideCharToMultiByte` Due to the backward compatibility of the mapping, many files in GB18030 can be actually opened successfully as the legacy Code Page 936, that is GBK, even if the Code Page 54936 is not supported. However, that is only true if the file in question contains only GBK characters. Loading will fail or cause corrupted result if the file contains characters that do not exist in GBK (see § Technical details for examples).

GNU glibc's gconv, the character codec library used on most Linux distributions, supports GB 18030-2000 since 2.2,[12] and GB 18030-2005 since 2.14;[13] glibc notably includes non-PUA mappings for GB 18030-2005 in order to achieve round-trip conversion.[14] GNU libiconv, an alternative iconv implementation frequently used on non-glibc UNIX-like environments like Cygwin, supports GB 18030 since version 1.4.[15]

Glyphs

The GB18030 Support Package for Windows contains SimSun18030.ttc, a TrueType font collection file which combines two Chinese fonts, SimSun-18030 and NSimSun-18030. The SimSun 18030 font includes all the characters in Unicode 2.1 plus new characters found in the Unicode CJK Unified Ideographs Extension A block, but despite its name, it does not contain glyphs for all GB 18030 characters, as all (about a million) Unicode code points up to U+10FFFF can be encoded as GB 18030. GB 18030 compliance certification only requires correct handling and recognition of glyphs in the mandatory (two-byte) Chinese part.[2]:4

Other CJK font families like HAN NOM[16] and Hanazono Mincho[17] provide wider coverage for Unicode CJK Extension blocks than SimSun-18030 or even Simsun (Founder Extended), but they don't support all code points defined in Unicode 5.0.0 either

See also

- Guobiao code
- CJK
- Chinese character encoding
- Comparison of Unicode encodings

Notes

a. Note that GB18030 omits surrogates; see #Mapping.
b. with the exception of the euro sign which is given a single byte code of 0x80 in Microsoft's later versions of CP936/GBK and a two byte code of A2 E3 in GB18030
c. Including the 66 Unicode noncharacters
d. ICU seems to erroneously consider this code point valid, which is in neither versions of the published standards. WHATWG assigns this byte to U+20AC (GBK Euro Sign) in its general-use gbk/gb18030 decoder.
e. For a finer division of this range see GBK (character encoding) § Encoding
f. Some code points are encoded with two bytes (upper row), the others with four bytes (lower row). U+FFFF is encoded as `84 31 A4 39` on page 239 of the 2005 standard, although the standard gives as far a`84 39 FE 39` for BMP mapping.
g. These are surrogate code points, they have no meaning outside of UTF-16 encoding.

References

1. Anthony Fok (2002-03-15). "Application of IANA Charset Registration for GB18030" (https://www.iana.org/assignments/charset-reg/GB18030) IANA Character Set Registrations Retrieved 2016-12-05.

2. CESI (2009-07-08). "GB18030 符合性问与答" (https://archive.org/details/GB18030-compliance-faq) [GB18030 compliance FAQ]. *CESI Certification Center*. Archived from the original (http://www.cc.cesi.cn/UpLoadFolder/OtherFile/200907/2009070816133686.doc) on 2016-09-28. Retrieved 2016-10-12. "*Page 4* 同时达到以下两个要求的产品，为符合GB 18030-2005强制部分的产品：①产品可以正确输入、输出、处理GB 18030-2005强制部分规定的全部汉字字符；②产品可以正确识别GB 18030-2005强制性部分规定的全部汉字字符对应的编码。[A product compliant with the mandatory part of GB 18030 must be able to correctly a) input, output and process all Chinese characters defined in the mandatory set; b) recognize encodings for characters in the mandatory set.]
3. Standardization Administration of China (SAC) (2005-11-18). *GB 18030-2005: Information Technology—Chinese coded character set* (https://archive.org/details/GB18030-2005)
4. "Unicode FAQ on GB 18030" (https://ssl.icu-project.org/docs/papers/unicode-gb18030-faq.html). *ICU Project*. Retrieved 10 September 2016.
5. Standardization Administration of China (SAC) (2000-03-17). *GB 18030-2000: Information Technology—Chinese coded character set for information interchange — Extension for the basic set* (https://archive.org/details/GB18030-2000).
6. Lunde, Ken (2006). "L2/06-394 Update on GB 18030:2005" (http://www.unicode.org/L2/L2006/06394-gb18030-2005.txt). Unicode Technical Committee Document Registry. Retrieved 28 September 2016.
7. "Group:GBK外字" (http://zht.glyphwiki.org/wiki/Group:GBK%E5%A4%96%E5%AD%97). *GlyphWiki*. Retrieved 11 September 2016.
8. Lunde, Ken (December 2008). *CJKV Information Processing* (https://books.google.com/books?id=SA92uQqTB-AC) O'Reilly Media, Inc. ISBN 978-0-596-51447-1. Retrieved 11 September 2016.
9. Authoritative mapping table between GB18030-2000 and Unicode (http://source.icu-project.org/repos/icu/data/trunk/charset/data/xml/gb-18030-2000.xml) ICU – International Components for Unicode. 2001-02-21. Accessed 2016-09-04.
10. "Encoding Standard # gb18030-index" (https://encoding.spec.whatwg.org/#index-gb18030-ranges-pointer) *WHATWG*. Retrieved 2016-09-24.
11. Microsoft. "GB18030 Support Package" (https://web.archive.org/web/20120605011449/http://www.microsoft.com/en-us/download/details.aspx?id=5503) Archived from the original (http://www.microsoft.com/downloads/details.aspx?FamilyID=fc02e2e3-14bb-46c1-afee-3732d6249647&DisplayLang=en) on 2012-06-05.
12. Drepper, Ulrich. "GB18030 iconv module for glibc" (https://sourceware.org/git/gitweb.cgi?p=glibc.git;a=commit;h=bc0413276ea17dc2a255ce726e9b9a504438980f) *glibc git*. Retrieved 29 November 2016.
13. Drepper, Ulrich. "Update GB18030 to 2005 version" (https://sourceware.org/git/gitweb.cgi?p=glibc.git;a=commit;h=ee30c380b8f7c9253c87103c58c5201268d30181) *glibc git*. Retrieved 29 November 2016.
14. Weimer, Florian; O'Donell, Carlos. "Status of GB18030 tables (#19575)" (https://sourceware.org/bugzilla/show_bug.cgi?id=19575). *Sourceware Bugzilla*. Retrieved 29 November 2016.
15. "NEWS - libiconv.git - libiconv" (http://git.savannah.gnu.org/cgit/libiconv.git/tree/NEWS?id=v1.14) *git.savannah.gnu.org*. Retrieved 2016-10-13.
16. VietUnicode. "/hannom" (https://sourceforge.net/projects/vietunicode/files/hannom/) *sourceforge.net*. Retrieved 2016-10-13.
17. "Hanazono fonts" (http://fonts.jp/hanazono/) *fonts.jp*. Retrieved 2016-10-13.

External links

- IANA Charset Registration for GB18030
- English language summary of GB 18030-2000
- Introduction to GB18030 including evolution from GB2312 and GBK (Sun/Internet Archive)
- ICU data
  - GB18030: A mega-codepage (IBM DeveloperWorks)
  - Authoritative mapping table between GB18030-2000 and Unicode
  - ICU Converter Explorer: GB18030
- Unicode charts
  - Unicode CJK Unified Ideographs Extension A (PDF, 1.5MB)
  - Unicode CJK Unified Ideographs Extension B (PDF, 13 MB)

- GB18030 Support Package for Windows 2000/XP including Chinese, Tibetan, Yi, Mongolian and Thai font by Microsoft (Internet Archive)
- SIL's freeware fonts, editors and documentation
