GB18030 编码

GB18030 编码采用单字节、双字节、四字节分段编码方案，具体码位见下文。

GB18030 向下兼容 GBK 和 GB2312 编码。

国家标准 GB18030《信息技术 中文编码字符集》是我国继 GB2312-1980 和 GB13000-1993 之后最重要的汉字编码标准，是我国计算机系统必须遵循的基础性标准之一。 GB18030 有三个版本：GB18030-2000、GB18030-2005 和 GB18030-2022。GB18030-2000 是 GBK 的取代版本，它的主要特点是在 GBK 基础上增加了 CJK 统一汉字扩充 A 的汉字。GB18030-2005 的主要特点是在 GB18030-2000 基础上增加了 CJK 统一汉字扩充 B 的汉字。GB18030-2022 则主要增加了 CJK 统一汉字扩充 C、D、E、F 区的汉字。

GB18030-2000 编码标准《信息技术 汉字编码字符集 基本集的扩充》是由信息产业部和国家质量技术监督局在 2000 年 3 月 17 日联合发布的，并且将作为一项国家标准在 2001 年的 1 月正式强制执行。GB18030-2000 仅规定了常用非汉字符号和 27533 个汉字（包括部首、部件等）的编码。

GB18030-2005《信息技术 中文编码字符集》是以汉字为主并包含多种我国少数民族文字的超大型中文编码字符集，其中收入汉字 70000 余个。在 GB18030-2000 的基础上增加了 42711 个汉字和多种我国少数民族文字的编码（如藏、蒙古、傣、彝、朝鲜、维吾尔文等）。增加的这些内容是推荐性的，原 GB18030-2000 中的内容是强制性的，市场上销售的产品必须符合。故 GB18030-2005 为部分强制性标准，自发布之日起代替 GB18030-2000。

GB18030-2022《信息技术 中文编码字符集》在 2005 版基础上再增加了一万多个汉字，使得汉字总数达到 87887 个，全面覆盖了《通用规范汉字表》中的汉字。收录的少数民族文字包括：藏文、滇东北苗文、彝文、傈僳文、朝鲜文、西双版纳新傣文、西双版纳老傣文、维吾尔文、哈萨克文、柯尔克孜文、蒙古文、德宏傣文等。

GB18030-2000 字汇

GB18030-2000 标准收录的字符分别以单字节、双字节和四字节编码。

1、单字节部分

本标准中，单字节的部分收录了 GB 11383 的 0x00 到 0x7F 全部 128 个字符及单字节编码的欧元符号。

2、双字节部分

本标准中，双字节的部分收录内容如下：

GB 13000.1 的全部 CJK 统一汉字字符。

GB 13000.1 的 CJK 兼容区挑选出来的 21 个汉字。

GB 13000.1 中收录而 GB 2312 未收录的我国台湾地区使用的图形字符 139 个。

GB 13000.1 收录的其它字符 31 个。

GB 2312 中的非汉字符号。

GB 12345 的竖排标点符号 19 个。

GB 2312 未收录的 10 个小写罗马数字。

GB 2312 未收录的带音调的汉语拼音字母 5 个以及α 和g 。

汉字数字“〇”。

表意文字描述符 13 个。

增补汉字和部首/构件 80 个。

双字节编码的欧元符号。

3 、四字节部分

本标准的四字节的部分，收录了上述双字节字符之外的，包括 CJK 统一汉字扩充 A 在内的 GB 13000.1 中的全部字符。

GB18030-2005 字汇

GB18030-2005 标准收录的字符分别以单字节、双字节或四字节编码。

1、单字节部分

本标准中，单字节的部分收录了 GB/T 11383-1989 的 0x00 到 0x7F 全部 128 个字符。

2、双字节部分

本标准中，双字节的部分收录内容如下：

GB 13000.1 - 1993 的全部 CJK 统一汉字字符。

GB 13000.1 - 1993 的 CJK 兼容区挑选出来的 21 个汉字。

GB 13000.1 - 1993 中收录而 GB 2312 未收录的我国台湾地区使用的图形字符 139 个。

GB 13000.1 - 1993 收录的其它字符 31 个。

GB 2312 中的非汉字符号。

GB 12345 的竖排标点符号 19 个。

GB 2312 未收录的 10 个小写罗马数字。

GB 2312 未收录的带音调的汉语拼音字母 5 个以及α 和g 。

汉字数字“〇”。

表意文字描述符 13 个。

对 GB 13000.1 - 1993 增补的汉字和部首/构件 80 个。

双字节编码的欧元符号。

3、四字节部分

本标准的四字节的部分，收录了上述双字节字符之外的，GB 13000 的 CJK 统一汉字扩充 A、CJK 统一汉字扩充 B 和已经在 GB13000 中编码的我国少数民族文字的字符。

GB18030-2005 最主要的变化是增加了 CJK 统一汉字扩充 B。它还去掉了单字节编码的欧元符号 0x80）。

GB18030 有 1611668 个码位，在 GB18030-2005 中定义了 76556 个字符。

随着我国汉字整理和编码研究工作的不断深入，以及国际标准 ISO/IEC 10646 的不断发展，GB18030 所收录的字符将在新版本中增加。

GB18030-2022 字汇

GB18030-2022 标准收录的字符分别以单字节、双字节或四字节编码。

1、单字节部分

本标准中，单字节的部分收录了 GB/T 11383-1989 的 0x00 到 0x7F 全部 128 个字符。

2、双字节部分

双字节部分采用两个八位二进制位串表示一个字符，其首字节码位从 0x81~0xFE，尾字节码位分别是 0x40~0x7E 和 0x80~0xFE。

3、双字节部分

四字节部分采用 GB/T 11383-1989 未采用的 0x30~0x39 作为对双字节编码扩充的后缀，编码范围为 0x81308130~0xFE39FE39。四字节字符的第一个字节编码范围为 0x81~0xFE；第二个字节编码范围为 0x30~0x39；第三个字节编码范围为 0x81~0xFE；第四个字节编码范围为 0x30~0x39。即：

0x81308130~0x81308139；

0x81308230~0x81308239；

……

0x8130FE30~0x8130FE39；

0x81318130~0x81318139；

……

0x8131FE30~0x8131FE39；

……

0x82308130~0x82308139；

……

0x8230FE30~0x8230FE39；

……

0xFE308130~0xFE308139；

……

0xFE39FE30~0xFE39FE39。

GB18030-2000 汉字

如下表所示，GB18030-2000 收录了 27533 个汉字：

类别	码位范围	码位数	字符数	字符类型
双字节部分	第一字节 0xB0-0xF7	6768	6763	汉字
	第二字节 0xA1-0xFE			
	第一字节 0x81-0xA0	6080	6080	汉字
	第二字节 0x40-0xFE			
	第一字节 0xAA-0xFE	8160	8160	汉字
	第二字节 0x40-0xA0			
四字节部分	第一字节 0x81-0x82	6530	6530	CJK 统一汉字扩充 A
	第二字节 0x30-0x39			
	第三字节 0x81-0xFE			
	第四字节 0x30-0x39			

27533 就是 6763+6080+8160+6530 。双字节部分的 6763+6080+8160=21003 个汉字就是 GBK 的 21003 个汉字。

在 Unicode 中，CJK 统一汉字扩充 A 有 6582 个汉字，为什么这里只有 6530 个汉字？

这是因为在 GBK 时代，双字节部分已经收录过 CJK 统一汉字扩充 A 的 52 个汉字，所以还余 6530 个汉字。

GB18030-2005 汉字

如下表所示，GB18030-2005 收录了 70244 个汉字：

类别	码位范围	码位数	字符数	字符类型
双字节部分	第一字节 0xB0-0xF7 第二字节 0xA1-0xFE	6768	6763	汉字
	第一字节 0x81-0xA0 第二字节 0x40-0xFE	6080	6080	汉字
	第一字节 0xAA-0xFE 第二字节 0x40-0xA0	8160	8160	汉字
四字节部分	第一字节 0x81-0x82 第二字节 0x30-0x39 第三字节 0x81-0xFE 第四字节 0x30-0x39	6530	6530	CJK 统一汉字扩充 A
	第一字节 0x95-0x98 第二字节 0x30-0x39 第三字节 0x81-0xFE 第四字节 0x30-0x39	42711	42711	CJK 统一汉字扩充 B

70244 就是 6763+6080+8160+6530+42711。

GB18030-2022 汉字

如下表所示，GB18030-2022 收录的汉字：

双字节部分	2 区： 首字节 0xB0~0xF7 尾字节 0xA1~0xFE	6768	6763	汉字
	3 区： 首字节 0x81~0xA0 尾字节 0x40~0x7E 和 0x80~0xFE	6080	6080	汉字
	4 区： 首字节 0xAA~0xFE 尾字节 0x40~0x7E 和 0x80~0xA0	8160	8145	汉字
四字节部分	0x8139EE39~0x82358738	6530	6530	CJK 统一汉字扩充 A

	0x82358F33~0x82359636	74	66	CJK 统一汉字
	0x95328236~0x9835F336	42711	42711	CJK 统一汉字扩充 B
	0x9835F738~0x98399E36	4149	4149	CJK 统一汉字扩充 C
	0x98399F38~0x9839B539	222	222	CJK 统一汉字扩充 D
	0x9839B632~0x9933FE33	5762	5762	CJK 统一汉字扩充 E
	0x99348138~0x9939F730	7473	7473	CJK 统一汉字扩充 F

双字节 4 区删除了 6 个重复编码的汉字构件和 9 个重复编码的汉字。

上表计算后的字符数为 87901 个，媒体报道给出的汉字数量为 87887 个，可能上表中包含的汉字构件未统计为汉字，未核实。

GB18030 码位分配

GB18030 编码采用单字节、双字节和四字节三种方式对字符编码。

- 单字节部分采用 GB/T 11383 的编码结构与规则，使用 0x00 至 0x7F 码位（对应 ASCII 码位）。
- 双字节部分，首字节码位从 0x81 至 0xFE，尾字节码位分别是 0x40 至 0x7E 和 0x80 至 0xFE。
- 四字节部分采用 GB/T 11383 未采用的 0x30 到 0x39 作为对双字节编码扩充的后缀，这样扩充的四字节编码，其范围为 0x81308130 到 0xFE39FE39。其中第一、三个字节编码码位均为 0x81 至 0xFE，第二、四个字节编码码位均为 0x30 至 0x39。