汉字机内码、国标码和区位码定义区别

1、国家标准汉字代码体系

汉字字数繁多，属性丰富，因而汉字代码体系也较复杂，包括：

(1)汉字机内码。它们是汉字在计算机汉字系统内部的表示方法，是计算机汉字系统的基础代码。

(2)汉字交换码。它们是国标汉字(如机内码)进行信息交换的代码标准。

(3)汉字输入码。它们是在计算机标准键盘上输入汉字用到的各种代码体系。

(4)汉字点阵码。它们是在计算机屏幕上显示和在打印机上打印输出汉字的代码体系。

(5)汉字字形控制码。为了打印各种风格的字体和字形所制定的代码。

这些代码系统有的必须有统一的国家标准，有的则不要求统一。近年来我国已经制定系列汉字信息处理方面的国家标准，今后将继续完善，并与国际上求得统一。

2、国家标准汉字交换码（国标码）

我国制定了“中华人民共和国国家标准信息交换汉字编码”，标准代号为GB2312—80，这种编码又称为国标码。在国标码的字符集中共收录了一级汉字3755个，二级汉字3008个，图形符号682个，三项字符总计7445个。

国标码是指1980年中国制定的用于不同的具有汉字处理功能的计算机系统间交换汉字信息时使用的编码。国际码是二字节码, 用两个七位二进制数编码表示一个汉字。目前国标码收入6763个汉字, 其中一级汉字（最常用）3755个, 二级汉字3008个, 另外还包括682个西文字符、图符。一级汉字为常用字，按拼音顺序排列，二级汉字为次常用字，按部首排列。国标码的范围是 2121H—7E7EH。

3、区位码：

国标码是一个四位十六进制数，区位码是一个四位的十进制数，每个国标码或区位码都对应着一个唯一的汉字或符号，但因为十六进制数我们很少用到，所

以大家常用的是区位码，它的前两位叫做区码，后两位叫做位码。

在国标 GB2312—80 中规定，所有的国标汉字及符号分配在一个 94 行、94 列的方阵中，方阵的每一行称为一个“区”，编号为 01 区到 94 区，每一列称为一个“位”，编号为 01 位到 94 位，方阵中的每一个汉字和符号所在的区号和位号组合在一起形成的四个阿拉伯数字就是它们的“区位码”。区位码的前两位是它的区号，后两位是它的位号。用区位码就可以唯一地确定一个汉字或符号，反过来说，任何一个汉字或符号也都对应着一个唯一的区位码。汉字“母”字的区位码是 3624，表明它在方阵的 36 区 24 位，问号“?”的区位码为 0331，则它在 03 区 3l 位。

所有的汉字和符号所在的区分为以下四个组：

(1)01 区到 15 区。图形符号区，其中 01 区到 09 区为标准符号区，10 区到 15 区为自定义符号区。

01 区到 09 区的具体内容如下：

1)01 区。一般符号 202 个，如间隔符、标点、运算符、单位符号及制符；

2)02 区。序号 60 个，如 1．~20.、(1)~(20)、①~⑩及(一）~（十）；

3)03 区。数字 22 个，如 0—9 及 X 一 XII，英文字母 52 个，其中大写 A—Z、小写 a—z 各 26 个；

4)04 区。日文平假名 83 个；

5)05 区。日文片假名 86 个；

6)06 区。希腊字母 48 个；

7)07 区。俄文字母 66 个；

8)08 区。汉语拼音符号 a—z26 个；

9)09 区。汉语拼音字母 37 个。

(2)16 区到 55 区。一级常用汉字区，包括了 3755 个一统汉字。这 40 个区中的汉字是按汉语拼音排序的，同音字按笔划顺序排序。其中 55 区的 90 一 94 位未定义汉字。

(3)56 区到 87 区。二级汉字区，包括了 3008 个二级汉字，按部首排序。

(4)88 区到 94 区。自定义汉字区。

第 10 区到第 15 区的自定义符号区和第 88 区到第 94 区的自定义汉字区可由用户自行定义国标码中未定义的符号和汉字。

4、国家标准汉字机内码（内码）

汉字的机内码是指在计算机中表示一个汉字的编码。机内码与区位码稍有区别。如上所述，汉字区位码的区码和位码的取值均在 1~94 之间，如直接用区位码作为机内码，就会与基本 ASCII 码混淆。为了避免机内码与基本 ASCII 码的冲突，需要避开基本 ASCII 码中的控制码(00H~1FH)，还需与基本 ASCII 码中的字符相区别。为了实现这两点，可以先在区码和位码分别加上 20H，在此基础上再加 80H(此处“H”表示前两位数字为十六进制数)。经过这些处理，用机内码表示一个汉字需要占两个字节，分别　称为高位字节和低位字节，这两位字节的机内码按如下规则表示：

高位字节＝区码+20H+80H(或区码+A0H)

低位字节=位码+20H+80H(或位码+AOH)

由于汉字的区码与位码的取值范围的十六进制数均为 01H~5EH(即十进制的 01~94)，所以汉字的高位字节与低位字节的取值范围则为 A1H~FEH(即十进制的 161~254)。

例如，汉字“啊”的区位码为 1601，区码和位码分别用十六进制表示即为 1001H，它的机内码的高位字节为 B0H，低位字节为 A1H，机内码就是 B0A1H。

5、汉字的输入码

在计算机标准键盘上，汉字的输入和西文的输入有很大的不同。西文的输入，击一次键就直接输入了相应的字符或代码，“键入”和“输入”是同一个含义。但是在计算机上进行汉字输入时，“键入”是指击键的动作即键盘操作的过程，而“输入”则是把所需的汉字或字符送到指定的地方，是键盘操作的目的。目前已有多种汉字输入方法，因此就有多种汉字输入码。汉字输入码是面向输入者的，使用不同的输入码其操作过程不同，但是得到的结果是一样的。不管采用何种输入方法，所有输入的汉字都以机内码的形式存储在介质中，而在进行汉字传输时，又都以交换码的形式发送和接收。

国标 GB2312—80 规定的区位码和沿用多年的电报码都可以作为输入码。这类汉字编码和输入码是一一对应的，具有标准的性质，它们编码用的字符是 10 个阿拉伯数字，每个汉字的码长均为等长的四个数码。

其他编码的种类很多，可从以下几点加以讨论：

(1)编码类型。可分为拼音码、字形码、音形结合码等类型。

(2)编码规则。不同的编码方案有很大的不同，有的规则简单，学习起来较容易记忆，有的规则复杂，较难记亿。

(3)编码字符集。有用字母键的，有用数字键的，有用字母键加数字键的，或者用了更多的键作编码字符集的。

(4)编码长度。它与编码字符集的大小有关，字符集越大，编码长度越短。采用 26 个字母的编码，其码长一般为四位。

(5)对应关系。除上面提到的区位码和电报码为一一对应的无重码编码外，其他现有的编码方案均有一定数量的重码。所谓重码即一码对应多字。有许多编码为了增加输入的灵活性，同一汉字用多个码来对应，例如双音编码。

(6)单宁和词汇的编码。现有的编码方案，为了提高效率，除了单字外还规定了词汇的编码，甚至使用者可以自行增加词汇库中的词汇，但在提高效率的同时也增加了记亿和操作的复杂性。

(7)码表的类型和大小。从汉字输入码到机内码的转换一股需要在机内检索码表。如果输入码和机内码存在简单的函数关系，有公式可以计算，如区位码等编码就不需要码表，其他没有简单函数关系的编码就需要码表。码表大小与数据结构、单字数量、词汇数量等因素有关。国标血 2312—80 规定的 6763 个一、二级汉字，备类编码的码表从几千字节到几万字节。随着词汇旦的增加，有的码表达到了若干兆字节。

6、汉字的点阵码

汉字的显示和输出，普遍采用点阵方法。由于汉字数量多且字形变化大，对不同字形汉字的输出，就有不同的点阵字形。所谓汉字的点阵码，就是汉字点阵字形的代码。存储在介质中的全部汉字的点阵码又称为字库。

16x16 点阵的汉字其点阵有 16 行，每一行上有 16 个点。如果每一个点用一个二进制位来表示，则每一行有 16 个二进制位，需用两个字节来存放每一行上的 16 个点，并且规定其点阵中二进制位 0 为白点，1 为黑点，这样一个 16X16 点阵的汉字需要用 2×t6 即 32 个字节来存放。依次类推，24×24 点阵和 32×32 点阵的汉字则依次要用 72 个字节和 128 个字节存放一个汉字，构成它在字库中

的字模信息。

要显示或打印输出一个汉字时，计算机汉字系统根据该汉字的机内码找出其字模信息在字库中的位置，再取出其字模信息作为字形在屏幕上显示或在打印机上打印输出。

汉字机内码、国标码和区位码三者之间的关系：

区位码（十进制）的两个字节分别转换为十六进制后加 20H 得到对应的国标码；机内码是汉字交换码（国标码）两个字节的最高位分别加 1，即汉字交换码（国标码）的两个字节分别加 80H 得到对应的机内码；区位码（十进制）的两个字节分别转换为十六进制后加 A0H 得到对应的机内码。

（1）区位码先转换成十六进制数表示

（2）国标码＝区位码的十六进制表示＋2020H

（3）机内码＝国标码＋8080H=区位码+A0A0H

举例：

以汉字“大”为例，“大”字的区内码为 2083

1、区号为 20，位号为 83

2、将区位号 2083 转换为十六进制表示为 1453H

3、1453H＋2020H＝3473H，得到国标码 3473H

4、3473H＋8080H＝B4F3H，得到机内码为 B4F3H