**几种字符集的介绍**

**摘 要：**学习xml与web技术这门课需要我们深入了解几种字符集的相关知识，这样可以帮助我们更好的学习如何在制作xml的时候更好的使用这些字符集。这对我们软工专业的学习有着至关重要的作用。

本文主要介绍了几种字符集例如Unicode，big5，GB2312的来历，产生原因和它们所解决的问题等相关方面的内容。这样有助于我们在接下来的对xml与web技术学习和其他软工方面的学习起到事半功倍的作用。

**关键词：**Unicode；UTF-8；UTF-16；GB2312；big5

**目 录**

摘要 I

1.Unicode 1

1.1 Unicode的来源.................................................................................................................2

1.2 Unicode的作用.................................................................................................................2

1.3 Unicode的字节数.............................................................................................................3

1.4 Unicode兼容的字符.........................................................................................................3

1.5 Unicode与其他字符的相互转换.....................................................................................3

1. UTF-8/16 3

2.1 UTF-8/16的来源...............................................................................................................3

2.2 UTF-8/16的字节数...........................................................................................................4

1. GB2312 4

3.1 GB2312的来源.................................................................................................................4

3.2 GB2312的收录.................................................................................................................4

3.3 GB2312的字节结构.........................................................................................................4

1. Big5 5

4.1 Big5历史发展...................................................................................................................5

4.2 Big5解决的问题...............................................................................................................5

4.3 Big5使用方法...................................................................................................................5

1. 结论 5

**1.Unicode**

**1.1 Unicode的起源**

Unicode（[统一码](https://baike.baidu.com/item/%E7%BB%9F%E4%B8%80%E7%A0%81" \t "https://baike.baidu.com/item/_blank)、万国码、单一码）是计算机科学领域里的一项业界标准,包括字符集、编码方案等。Unicode 是为了解决传统的字符编码方案的局限而产生的，它为每种语言中的每个字符设定了统一并且唯一的[二进制](https://baike.baidu.com/item/%E4%BA%8C%E8%BF%9B%E5%88%B6" \t "https://baike.baidu.com/item/_blank)编码，以满足跨语言、跨平台进行文本转换、处理的要求。1990年开始研发，1994年正式公布。

Unicode 是为了解决传统的[字符编码](https://baike.baidu.com/item/%E5%AD%97%E7%AC%A6%E7%BC%96%E7%A0%81" \t "https://baike.baidu.com/item/_blank)方案的局限而产生的，例如ISO 8859所定义的字符虽然在不同的国家中广泛地使用，可是在不同国家间却经常出现不兼容的情况。很多传统的[编码](https://baike.baidu.com/item/%E7%BC%96%E7%A0%81/80092" \t "https://baike.baidu.com/item/_blank)方式都有一个共同的问题，即容许电脑处理双语环境（通常使用[拉丁字母](https://baike.baidu.com/item/%E6%8B%89%E4%B8%81%E5%AD%97%E6%AF%8D" \t "https://baike.baidu.com/item/_blank)以及其本地语言），但却无法同时支持多语言环境（指可同时处理多种语言混合的情况）。

Unicode 编码包含了不同写法的字，如“ɑ/a”、“户/户/戸”。然而在汉字方面引起了一字多形的认定争议。

在文字处理方面，统一码为每一个字符而非字形定义唯一的代码（即一个整数）。换句话说，统一码以一种抽象的方式（即数字）来处理字符，并将视觉上的演绎工作（例如字体大小、外观形状、字体形态、文体等）留给其他软件来处理，例如网页浏览器或是文字处理器。

几乎所有电脑系统都支持基本拉丁[字母](https://baike.baidu.com/item/%E5%AD%97%E6%AF%8D" \t "https://baike.baidu.com/item/_blank)，并各自支持不同的其他编码方式。Unicode为了和它们相互兼容，其首256字符保留给ISO 8859-1所定义的字符，使既有的西欧语系文字的转换不需特别考量；并且把大量相同的字符重复编到不同的字符码中去，使得旧有纷杂的编码方式得以和Unicode编码间互相直接转换，而不会丢失任何信息。举例来说，[全角](https://baike.baidu.com/item/%E5%85%A8%E8%A7%92" \t "https://baike.baidu.com/item/_blank)格式区段包含了主要的拉丁字母的全角格式，在中文、日文、以及韩文字形当中，这些字符以全角的方式来呈现，而不以常见的半角形式显示，这对竖排文字和等宽排列文字有重要作用。

在表示一个Unicode的字符时，通常会用“U+”然后紧接着一组十六进制的数字来表示这一个字符。在[基本多文种平面](https://baike.baidu.com/item/%E5%9F%BA%E6%9C%AC%E5%A4%9A%E6%96%87%E7%A7%8D%E5%B9%B3%E9%9D%A2" \t "https://baike.baidu.com/item/_blank)（英文为 Basic Multilingual Plane，简写 BMP。它又简称为“零号平面”, plane 0）里的所有字符，要用四位十六进制数（例如U+4AE0，共支持六万多个字符）；在零号平面以外的字符则需要使用五位或六位十六进制数了。旧版的Unicode标准使用相近的标记方法，但却有些微的差异：在Unicode 3.0里使用“U-”然后紧接着八位数，而“U+”则必须随后紧接着四位数。

**1.2 Unicode的作用**

Unicode的主要作用就是能够使计算机实现跨语言、[跨平台](https://baike.baidu.com/item/%E8%B7%A8%E5%B9%B3%E5%8F%B0" \t "https://baike.baidu.com/item/_blank)的文本[转换](https://baike.baidu.com/item/%E8%BD%AC%E6%8D%A2" \t "https://baike.baidu.com/item/_blank)及处理。因为很多传统的[编码](https://baike.baidu.com/item/%E7%BC%96%E7%A0%81/80092" \t "https://baike.baidu.com/item/_blank)方式都有一个共同的问题，即容许电脑处理双语环境（通常使用[拉丁字母](https://baike.baidu.com/item/%E6%8B%89%E4%B8%81%E5%AD%97%E6%AF%8D" \t "https://baike.baidu.com/item/_blank)以及其本地语言），但却无法同时支持多语言环境（指可同时处理多种语言混合的情况），所以不同国家间却经常出现不兼容的情况。而有了Unicode之后就能够帮助计算机跨语言、[跨平台](https://baike.baidu.com/item/%E8%B7%A8%E5%B9%B3%E5%8F%B0" \t "https://baike.baidu.com/item/_blank)的文本[转换](https://baike.baidu.com/item/%E8%BD%AC%E6%8D%A2" \t "https://baike.baidu.com/item/_blank)及处理，从而使各种不同平台的计算机都能兼容这些文本。

**1.3 Unicode的字节数**

[通用字符集](https://baike.baidu.com/item/%E9%80%9A%E7%94%A8%E5%AD%97%E7%AC%A6%E9%9B%86" \t "https://baike.baidu.com/item/_blank)（Universal Character Set, UCS）是由ISO制定的ISO 10646（或称ISO/IEC 10646）标准所定义的标准字符集。UCS-2用两个字节编码，UCS-4用4个字节编码。

**1.4 Unicode兼容的字符**

Unicode兼容的字符集包括中文、英文、日文、韩文等等的各国文字，还有数字和标点符号等。

**1.5 Unicode与其他字符的相互转换**

Unicode转换为UTF-8时，可以将Unicode二进制从低位往高位取出二进制数字，每次取6位，如上述的二进制就可以分别取出为如下示例所示的格式，前面按格式填补，不足8位用0填补。

注：Unicode转换为UTF-8需要的字节数可以根据这个规则计算：如果Unicode小于0X80（Ascii字符），则转换后为1个字节。否则转换后的字节数为Unicode二进制位数减1再除以5。

1. **UTF-8与UTF-16**

**2.1 UTF-8/16的来源**

UTF-8（8-bit Unicode Transformation Format）是一种针对Unicode的可变长度字符编码，又称万国码。由Ken Thompson于1992年创建。现在已经标准化为RFC 3629。UTF-8用1到6个字节编码Unicode字符。用在网页上可以统一页面显示中文简体繁体及其它语言（如英文，日文，韩文）。

UTF-16是[Unicode](https://baike.baidu.com/item/Unicode" \t "https://baike.baidu.com/item/_blank)字符编码五层次模型的第三层：字符编码表（Character Encoding Form，也称为 "storage format"）的一种实现方式。即把Unicode字符集的抽象码位映射为16位长的整数（即[码元](https://baike.baidu.com/item/%E7%A0%81%E5%85%83" \t "https://baike.baidu.com/item/_blank)）的序列，用于数据存储或传递。Unicode字符的码位，需要1个或者2个16位长的码元来表示，因此这是一个变长表示。

事实证明，对可以用[ASCII](https://baike.baidu.com/item/ASCII" \t "https://baike.baidu.com/item/_blank)表示的[字符](https://baike.baidu.com/item/%E5%AD%97%E7%AC%A6" \t "https://baike.baidu.com/item/_blank)使用UNICODE并不高效，因为UNICODE比ASCII占用大一倍的空间，而对ASCII来说高[字节](https://baike.baidu.com/item/%E5%AD%97%E8%8A%82" \t "https://baike.baidu.com/item/_blank)的0对他毫无用处。为了解决这个问题，就出现了一些中间格式的[字符集](https://baike.baidu.com/item/%E5%AD%97%E7%AC%A6%E9%9B%86" \t "https://baike.baidu.com/item/_blank)，他们被称为通用转换格式，即[UTF](https://baike.baidu.com/item/UTF" \t "https://baike.baidu.com/item/_blank)（Unicode Transformation Format）。常见的UTF格式有：UTF-7, UTF-7.5, UTF-8,[UTF-16](https://baike.baidu.com/item/UTF-16" \t "https://baike.baidu.com/item/_blank), 以及 [UTF-32](https://baike.baidu.com/item/UTF-32" \t "https://baike.baidu.com/item/_blank)。

**2.2 UTF-8/16的字节数**

如果UNICODE[字符](https://baike.baidu.com/item/%E5%AD%97%E7%AC%A6" \t "https://baike.baidu.com/item/_blank)由2个[字节](https://baike.baidu.com/item/%E5%AD%97%E8%8A%82" \t "https://baike.baidu.com/item/_blank)表示，则编码成UTF-8很可能需要3个字节。而如果UNICODE字符由4个字节表示，则编码成UTF-8可能需要6个字节。用4个或6个字节去编码一个UNICODE字符可能太多了，但很少会遇到那样的UNICODE字符。

UTF-16比起[UTF-8](https://baike.baidu.com/item/UTF-8" \t "https://baike.baidu.com/item/_blank)，好处在于大部分字符都以固定长度的字节 (2字节) 储存，但UTF-16却无法兼容于[ASCII](https://baike.baidu.com/item/ASCII" \t "https://baike.baidu.com/item/_blank)编码。

1. **GB2312**

**3.1 GB2312的来源**

《信息交换用汉字编码字符集》是由中国国家标准总局1980年发布，1981年5月1日开始实施的一套国家标准，标准号是[GB 2312](https://baike.baidu.com/item/GB 2312" \t "https://baike.baidu.com/item/_blank)—1980。

GB2312编码适用于汉字处理、汉字通信等系统之间的信息交换，通行于中国大陆；新加坡等地也采用此编码。中国大陆几乎所有的中文系统和国际化的软件都支持GB 2312。

基本集共收入汉字6763个和非汉字图形字符682个。整个字符集分成94个区，每区有94个位。每个区位上只有一个字符，因此可用所在的区和位来对汉字进行编码，称为[区位码](https://baike.baidu.com/item/%E5%8C%BA%E4%BD%8D%E7%A0%81" \t "https://baike.baidu.com/item/_blank)。

把换算成十六进制的区位码加上2020H，就得到[国标码](https://baike.baidu.com/item/%E5%9B%BD%E6%A0%87%E7%A0%81" \t "https://baike.baidu.com/item/_blank)。国标码加上8080H，就得到常用的计算机机内码。1995年又颁布了《汉字编码扩展规范》（GBK）。GBK与GB 2312—1980国家标准所对应的内码标准兼容，同时在字汇一级支持ISO/IEC10646—1和GB 13000—1的全部中、日、韩（CJK）汉字，共计20902字。

**3.2 GB23212的收录**

GB 2312标准共收录6763个汉字，其中一级汉字3755个，二级汉字3008个；同时，GB 2312收录了包括[拉丁字母](https://baike.baidu.com/item/%E6%8B%89%E4%B8%81%E5%AD%97%E6%AF%8D" \t "https://baike.baidu.com/item/_blank)、[希腊字母](https://baike.baidu.com/item/%E5%B8%8C%E8%85%8A%E5%AD%97%E6%AF%8D" \t "https://baike.baidu.com/item/_blank)、日文平假名及片假名字母、俄语西里尔字母在内的682个[全角字符](https://baike.baidu.com/item/%E5%85%A8%E8%A7%92%E5%AD%97%E7%AC%A6" \t "https://baike.baidu.com/item/_blank)。

GB 2312的出现，基本满足了汉字的计算机处理需要，它所收录的汉字已经覆盖中国大陆99.75%的使用频率。

对于人名、古汉语等方面出现的罕用字，GB 2312不能处理，这导致了后来[GBK](https://baike.baidu.com/item/GBK" \t "https://baike.baidu.com/item/_blank)及GB 18030汉字字符集的出现。

**3.3 GB2312的字节结构**

在使用GB2312的程序中，通常采用EUC储存方法，以便兼容于ASCII。浏览器编码表上的“GB2312”，通常都是指“[EUC-CN](https://baike.baidu.com/item/EUC-CN" \t "https://baike.baidu.com/item/_blank)”表示法。

每个汉字及符号以两个字节来表示。第一个[字节](https://baike.baidu.com/item/%E5%AD%97%E8%8A%82" \t "https://baike.baidu.com/item/_blank)称为“高位字节”（也称“区字节）”，第二个字节称为“低位字节”（也称“位字节”）。

“高位字节”使用了0xA1-0xF7(把01-87区的区号加上0xA0)，“低位字节”使用了0xA1-0xFE(把01-94加上 0xA0)。 由于一级汉字从16区起始，汉字区的“高位字节”的范围是0xB0-0xF7，“低位字节”的范围是0xA1-0xFE，占用的码位是 72\*94=6768。其中有5个空位是D7FA-D7FE。

例如“啊”字在大多数程序中，会以两个[字节](https://baike.baidu.com/item/%E5%AD%97%E8%8A%82" \t "https://baike.baidu.com/item/_blank)，0xB0（第一个字节） 0xA1（第二个字节）储存。[区位码](https://baike.baidu.com/item/%E5%8C%BA%E4%BD%8D%E7%A0%81" \t "https://baike.baidu.com/item/_blank)=区字节+位字节（与区位码对比：0xB0=0xA0+16,0xA1=0xA0+1）。

1. **Big5（大五码）**

**4.1 Big5历史发展**

大五码是由资策会于1984年策划制定，拥有13053个中文字、408个字符以及33个控制字元的字集，是我国早期中文电脑的业界标准，也是中文社群最常用的电子汉字字集标准。而后随着电脑扩充需要，业界各操作系统开发商推出了不同版本的大五码，为统一标准，经济部标准检验局在2003年委托财团法人中国数位化技术推广基金会修改了大五码编码字元表，重整为Big5-2003版本。

**4.2 Big5解决的问题**

Big5是使用[繁体中文](http://zh.wikipedia.org/wiki/%E7%B9%81%E4%BD%93%E4%B8%AD%E6%96%87)（正体中文）社区中最常用的电脑[汉字](http://zh.wikipedia.org/wiki/%E6%B1%89%E5%AD%97" \o "汉字)[字符集](http://zh.wikipedia.org/wiki/%E5%AD%97%E7%AC%A6%E9%9B%86" \o "字符集)标准，共收录13,053个汉字。中文码分为[内码](http://zh.wikipedia.org/wiki/%E5%85%A7%E7%A2%BC)及[交换码](http://zh.wikipedia.org/wiki/%E4%BA%A4%E6%8F%9B%E7%A2%BC)两类，Big5属中文内码，知名的中文交换码有[CCCII](http://zh.wikipedia.org/wiki/CCCII" \o "CCCII)、[CNS11643](http://zh.wikipedia.org/wiki/CNS11643" \o "CNS11643)。Big5虽普及于[台湾](http://zh.wikipedia.org/wiki/%E5%8F%B0%E7%81%A3)、[香港](http://zh.wikipedia.org/wiki/%E4%B8%AD%E5%9C%8B%E9%A6%99%E6%B8%AF" \o "中国香港)与[澳门](http://zh.wikipedia.org/wiki/%E4%B8%AD%E5%9C%8B%E6%BE%B3%E9%96%80" \o "中国澳门)等繁体中文通行区，但长期以来并非当地的国家标准，而只是[业界标准](http://zh.wikipedia.org/wiki/De_facto" \o "De facto)。[倚天中文系统](http://zh.wikipedia.org/wiki/%E5%80%9A%E5%A4%A9%E4%B8%AD%E6%96%87%E7%B3%BB%E7%B5%B1)、[Windows](http://zh.wikipedia.org/wiki/Windows" \o "Windows)等主要系统的字符集都是以Big5为基准，但厂商又各自增加不同的造字与造字区，派生成多种不同版本。[2003年](http://zh.wikipedia.org/wiki/2003%E5%B9%B4)，Big5被收录到CNS11643中文标准交换码的附录当中，取得了较正式的地位。这个最新版本被称为Big5-2003。

**4.3 Big5使用方法**

1、首先安装BIG5码字体,安装方法：打开“控制面版”，找到“字体”打开，将需要的字体文件复制到字体文件夹中即可，安装完毕最好重新启动电脑。其它字体安装方法同上。（华康、金梅、超世纪等繁体字库都属于此类大五码字体，字体扩展名为TTF、或TTC）。

2、在word里把需要的汉字输入工作区内，然后将文字全选，选择“工具”菜单中的“语言”，内有“中文简繁转换”选择“简体中文转换为繁体中文”点确定，如果word中没有这个项目，说明你还没有安装，请首先找office安装光盘安装此项目。

3、如果你想在photoshop中使用，请在word中把要用的文字用上述步骤转换简体中文为繁体中文后，复制文字段，打开 photoshop，新建图像，点击文字输入工具，注意：先选择好字体（通常BIG5码字体名称都显示为字母和乱码，在字体列表最下边），然后用“粘贴” 命令，所要的字就出来了。一定注意，先选择字体再粘贴。

4、其他的繁体字库（BIG5码）98里也许不可以用，但在2000或xp、2003中按同样的方法也可以使用。

1. **结论**

在计算机的发展中，出现了许多的字符集。这些字符集兼容着不同的字符，每个字符集所应用的平台和作用都不一样。了解这些字符集对我们以后的web开发有着非常重要的帮助。

**参考网站**

1. 维基百科
2. 百度百科
3. [国家标准《信息交换用汉字编码字符集·第二辅助集》和《信息交换用汉字编码字符集·第四辅助集》的研制](https://baike.baidu.com/redirect/626eoBfpU6Uk2PWq7t3AE0I8XnvYWQQ709atYZQiCQ7yjtaxbB28IDUjw3QeBAt96mgc2MMfPS6ptFoqRwQ43vmkVoVtvpcmdj6d17qZkhctiD8uuX9N" \t "https://baike.baidu.com/item/_blank)