中文字符编码标准+Unicode+Code Page

2011 年 10 月 21 日 下午 4:38 crifan 已有3026人围观 我来说几句

中文字符编码标准+Unicode+Code Page

Version: 2011-10-21

Author: crifan

Mail: green-waste (at) 163.com

此文的PDF版本可以去这下载：

中文字符编码标准+Unicode+Code Page

http://bbs.chinaunix.net/thread-3610023-1-1.html

【中文字符编码标准】

1. GB2312，CP936，GBK，GB18030，GB13000

（1）GB2312

1980年，中国制定了GB2312-80，一共收录了 7445 个字符，包括 6763 个汉字和 682 个其它符号。

GB2312-80，简称为GB2312。

在 Windows 中的代码页（Code Page）是 CP936。

（2）GB13000

1993年，国际标准Unicode 1.1版本推出，收录中国大陆、台湾、日本及韩国通用字符集的汉字，总共有20,902个。

中国大陆订定了等同于Unicode 1.1版本的“GB 13000.1-93”，简称为GB13000。

GB13000，显然包含的GB2312已有的文字和其他很多为包含的文字，如GB 2312-80推出以后才简化的汉字（如“啰”），部分人名用字（如中国前总理朱镕基的“镕”字），台湾及香港使用的繁体字，日语及朝鲜语汉字等。

（3）GBK
微软，对GB2312-80的扩展，即利用GB 2312-80未使用的编码空间，收录所有的GB 13000.1-93和Unicode 1.1之中的汉字全部字符，制定了GBK编码。
GBK 收录了 21886 个符号，它分为汉字区和图形符号区。汉字区包括 21003 个字符。

GBK 作为对 GB2312 的扩展，在现在的 Windows 系统中仍然使用代码页 CP936 表示，但是同样的 936 的代码页跟一开始的 936 的代码页只支持 GB2312 编码不同，现在的 936代码页支持 GBK 的编码，GBK 同时也向下兼容GB2312 编码。

所以，技术编码上，GBK兼容旧的GB2312，但是编码方式和GB13000不同，不兼容GB13000，但是所包含文字上，算是和GB13000相同。

（4）GB18030
GBK自身并非国家标准，只是曾由国家技术监督局标准化司、电子工业部科技与质量监督司公布为“技术规范指导性文件”。
原始GB13000一直未被业界采用，2000年，国家出了标准GB18030-2000，简称GB18030，技术上兼容GBK而非GB13000，取代了 GBK1.0，成了正式的国家标准。
该标准收录了 27484 个汉字，同时还收录了藏文、蒙文、维吾尔文等主要的少数民族文字。
现在的PC平台必须支持 GB18030 ，对嵌入式产品暂不作要求。所以手机、MP3 一般只支持 GB2312。
GB18030 在 Windows 中的代码页是 CP54936。

这么多汉字编码标准的关系，总结起来就是：
2. 各种中文字符编码标准的关系
（ 中国大陆的标准）GB 13000.1-93 =（国际标准）Unicode 1.1

（中国大陆标准）GB2312-80
= 简称GB2312
= Windows系统中的 原先的CP936

（微软制定的）GBK
=（微软在编码方面）对 GB2312 的扩展
=（微软在所包含字符方面上包含了） GB 13000.1-93 + 其他部分汉字+ 台湾和香港的繁体 + 日语 + 朝鲜汉字
= Unicode 1.1 + 其他部分汉字+ 台湾和香港的繁体 + 日语 + 朝鲜汉字

对于GBK：
- ✍ 在编码方面：向下兼容GB2312，但是和GB 13000不同

✍ 在内容方面：等价于GB13000

微软中 现在的新的CP936
= GBK
=兼容旧的GB2312

在技术编码方面上，演化顺序为，ASCII -> GB2312 -> GBK -> GB18030。
后者对之前的，都是支持之前的编码，即向下兼容，即同一个字符，在这些编码中，都是同样的值，后面的标准，支持更多的字符。
区分中文编码的方法是高字节的最高位不为 0。
按照程序员的称呼，GB2312、GBK 到 GB18030 都属于双字节字符集 (DBCS)。

图表 1 中文字符相关编码标准

编码标准	别名	标准所属	包含字符
ASCII		国际通用	
GB2312	微软Windows中以前的CP936	中国大陆	6763 个汉字和 682 个其它符号
Unicode 1.1		国际通用	20,902个字符
GB13000		中国大陆	20,902个字符
GBK	微软Windows中现在的CP936	微软	21886 个符号
GB18030	微软Windows中的CP54936	中国大陆	27484 个汉字+其他少数民族字符

【Unicode】
世界上有很多个国家（和地区），每个国家都有自己对应的字符编码。
如果把各种文字编码形容为各地的方言，那么Unicode就是世界各国合作开发的一种语言。
在这种语言环境下，不会再有语言的编码冲突，在同屏下，可以显示任何语言的内容，这就是Unicode的最大好处。
那么Unicode是如何编码的呢？其实非常简单。
就是将世界上所有的文字用 2 个字节统一进行编码。可能你会问， 2 个字节最多能够表示65536个编码，够用吗？
韩国和日本的大部分汉字都是从中国传播过去的，字型是完全一样的。
比如： “文”字，GBK和SJIS中都是同一个汉字，只是编码不同而已。
那样，像这样统一编码， 2 个字节就已经足够容纳世界上所有的语言的大部分文字了。
Unicode的学名是“Universal Multiple-Octet Coded Character Set”，简称为UCS。
现在用的是UCS-2，即 2 个字节编码，而UCS-4是为了防止将来 2 个字节不够用才开发的。UCS-2也称为基本多文种平面。
UCS-2转换到UCS-4只是简单的在前面加 2 个字节0。UCS-4则主要用于保存辅助平面，例如

Unicode 4.0中的第二辅助平面：

20000-20FFF – 21000-21FFF – 22000-22FFF – 23000-23FFF – 24000-24FFF – 25000-25FFF – 26000-26FFF – 27000-27FFF – 28000-28FFF – 29000-29FFF – 2A000-2AFFF – 2F000-2FFFF

总共增加了16个辅助平面，由原先的65536个编码扩展至将近100万编码。