汉字内码扩展规范
——汉字编码的又一发展

电子部四所（100007） 林 宁 毛永刚

摘 要 从汉字编码的发展背景引出汉字内码扩展规范制定的必要性及重要性；阐述了该规范的技术要点、体系结构及将产生的影响和作用。

关键词 汉字编码 内码 扩展 信息化 GBK

一、发展背景

世界正进入信息时代。信息化已成为世界各国发展经济的必然趋势和迫切任务。实现信息化的四个关键要素是计算机、通信、信息和人，而这四个关键要素则集中体现在所采用的信息交换及信息处理技术上。

1980年我国颁布了GB2312《信息交换用汉字编码字符集》基本集。该标准共收入了6763个汉字及常用符号，奠定了中文信息处理的基础。为了满足更广泛的应用需求，继GB2312之后又颁布了GB7589、GB12345、GB13131等5个辅助集标准，共收了包括繁体字在内的近四万个汉字。80年代初第一个汉字操作系统CCDOS的推出不仅代表了中文平台开发的技术革命，而且也奏响了中文DOS“战国纷争”的序曲。从1985到1995的十年中间，我国中文平台得到了极大的发展，积极地介入了国际软件技术革命的潮流之中，从产品、技术和功能上来讲也日臻成熟，如联想DOS、UCDOS5.0及被行家认为“有飞跃性发展”的天汇3.0。同时也带来了建立在平台之上的与中文有关的软件的大繁荣，如各种词典、机器翻译工具、中文字处理、排版软件，各种著名西文应用软件的本地化产品，如中文AUTOCAD、汉化的Turbo C等。而大量西文优秀软件可以不经本地化处理直接运行在各种中文平台之上，可以象处理西文一样处理中文。这些汉字系统都是以GB2312—80为内码标准，并作了少量扩充开发出来的。由此可见信息交换用编码标准在我国软件产业及信息技术的发展过程所发挥的巨大作用。

随着我国对外开放的扩大及社会经济、文化领域国际间的交流与合作的加强，以及海峡两岸贸易往来的日益增加，特别是全球信息高速公路的建立与发展，信息处理应用对字符集编码提出了多文种、大字量、多用途的要求，要求简繁并存和增加汉字的呼声更加强烈。

1993年，经过ISO/IEC JTC1/SC2/WG2的组织，以及各成员国的积极努力，历经近十年终于出版发行了ISO／IEC10646—1《信息技术通用多八位编码字符集（UCS）》。该标准立足于多个八位字节，包括了世界上近百种文字及各种符号，各文种字符地位平等，其中汉字收录了包括中日韩20 902个汉字，我国包括GB2312在内的几大汉字标准作为原子集均收入在该标准中。目前该标准下一版的制

定工作尚在进行中，我国的藏文、蒙文及彝文等文字将逐步扩充进去。该标准将为多文种信息处理及信息交换打下坚实的基础，是未来系统发展的必然趋势。

由于GB2312仅仅是内码的基础，具体系统实现中各厂商之间仍有内码不兼容的情况，内码的混乱潜伏在系统与系统界面上。多种汉字内码共存的危害性随着计算机应用水平的提高和网络的发展已暴露出来。另外，GB2312的汉字编码不能满足INTERNET上的应用，GB2312-80已难于适应信息化发展的新要求，尽管ISO／IEC10646能够满足更为广泛的需求，但它是一个新的体系结构，与现有系统GB2312不兼容，要过渡到这个新的体系还将有一段时间。因此，基于GB2312研究制定汉字内码扩展规范已成为当务之急。

全国信息技术标准化技术委员会(以下简称信标委)从1995年5月开始会同国内主要系统开发厂商研究汉字内码扩展规范(简称为GBK)方案，经过3个月的紧张讨论与反复磋商，8月份完成方案的总体设计，12月份完成规范的制定工作。1995年12月15日国家技术监督局标准化司和电子工业部科技与质量监督司联合发文将《汉字内码扩展规范(GBK)》作为技术规范指导性文件发布和实施。

二、GBK的要点

1. GBK制定原则

·与GB2312信息处理交换码所对应的、事实上的内码标准兼容；

·在字汇一级支持ISO／IEC10646-1和GB13000-1的全部中日韩（CJK）汉字；

·除了GB2312和GB12345中包括的全部非汉字符号外，本规范还涵盖台湾中文标准交换码TCA-CNS11643（与其对应的内码为Big5）。

注：ISO／IEC 10646和BG13000－1完全等同。

2. 总体结构

GBK总体结构采用8140-FEFE的矩形区域，剔除xx7F一条线，共23 940个码位，参见图1。

标准编码区分位如下五个区：

GBK／1：A1A1-A9FE，846个码位，717个图形符号

GBK／2：BDA1—F7FE，6768个码位，6763个汉字

GBK／3：8140-AOFE，6080个码位，6080个汉字

GBK／4：AA40-FEAO，8160个码位，8160个汉字

GBK／5：A840-A9A0，192个码位，166个图形符号

用户自定义区分为如下三个区：

用户区／1：AAA1-AFFE，564个码位

用户区／2：F8A1-FEFE，658个码位

用户区／3：A140-A7A0，672个码位

3. GBK字序

·GB2312的汉字依然按照原有Ⅰ级字、Ⅱ级字，分别按拼音、部首／笔画排列；

·ISO／IEC 10646-1的其他CJK汉字，按UCS代码大小顺序排列；

·追加的80个汉字与部首／部件，与上述两类字汇分开，按康熙字典页码字位单独排列。

4. GBK与其他标准的关系

GBK汉字区和图形符号区的所有字符，都与ISO／IEC10646-1的编码字符一一对应；52个追加汉字（简化字总表中未收入的简化汉字及其对应繁体字）、28个部首／构件以及13个结构符均暂时对应于ISO／IEC10646－1的专用区（E000-F8FE）；带音调的拼音字母与ISO／IEC10646-1中A区的拉丁编码字符相对应。GBK与GB2312完全兼容，并且是Big5的并集和超集。GBK与其