·实践平台·

古籍数字化中的汉字录入与显示

徐 健 (中山大学资讯管理系 广东广州 510275)
肖 卓 (中山大学图书馆 广东广州 510275)

摘 要: 文章针对古籍数字化工作中大量繁难汉字录入和显示困难的问题,从计算机汉字输入与显示的基本原理入手,从五个方面提出了具体解决方案,较好地解决了古籍繁难文字处理的难题,对提高古籍数字化工作效率具有一定的借鉴意义。

关键词: 古籍数字化 汉字处理 输入法 Unicode

中图分类号: G255 文献标识码: A 文章编号: 1003-6938(2006)06-0079-04

Chinese Characters' Input and Display in Ancient Books' Digitalization

Xu Jian (Department of Information Management, Sun Yet-sen University, Guangzhou, Guangdong, 510275)
Xiao Zhuo (Library of Sun Yet-sen Unviersity, Ggangzhou, Guangdong, 510275)

Abstract: Concerned with the problems of input and display difficulties of complex Chinese characters, the paper aims at putting forward a concrete solution to the problem from five aspects on the basis of basic principles of the computational input and display of Chinese characters. The solution is of referential value to improve the efficiency of digitalization of ancient books.

Key words: ancient books' digitalization; Chinese characters processing; input method; unicode

CLC number: G255 Document code: A Article ID: 1003-6938(2006)06-0079-04

1 前言

在图书馆古籍数字化工作中,经常需要进行针对繁体汉字的计算机录入与显示。由于我国历史上出现过的汉字总数有8万多(也有6万多的说法),其中多数为异体字和罕用字",[1]在计算机汉字输入与显示方面,其实现难度远比英文输入和显示所用的26个字母及相关符号要大得多。因此,汉字显示时出现乱码、一些非常用字无法输入等问题在工作中时有发生,影响了古籍数字化工作的效率。针对这一问题,笔者进行了较为深入的研究,并从五个方面指出对计算机系统进行合理的配置,就能够较好地解决大量繁体汉字的显示和输入问题。为了更好地理解配置过程,我们先从计算机显示与输入汉字的原理以及相关标准说起。

2 计算机显示与输入汉字的相关标准

计算机只能直接处理和保存以二进制数字形式存在的信息,因此,所有字符必须经过编码后才能被计算机处理。计算机使用的缺省编码方式就是计算机的内码。早期的计算机使用7位二进制数的ASCII编码,它包括了常用的大小写英文字母、数字、标点符号以及控制字符等128个字符。

由于汉字数量多,用一个字节的128种状态不能全部表示出来,因此在1980年我国颁布的《信息交换用汉字编码字符集——基本集》,即国家标准GB2312-80方案中规定用两个字节的十六位二进制表示一个汉字,用以表示6763个常用汉字和682个其它符号。

GB2312支持的6763个常用汉字只占我国汉字数量极少的一部分,仅能够基本满足一般输入工作需要。1995年我

收稿日期: 2005-12-12; 责任编辑: 党燕妮

国颁布了汉字扩展规范 GBK1.0,它收录了 21886 个符号,分为汉字区和图形符号区,汉字区包括 21003 个字符。

2000 年我国发布的《信息技术信息交换用汉字编码字符集基本集的扩充》(GB18030),是取代 GBK1.0 的正式国家标准。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。[2]

为了解决世界范围内的信息交换、处理和显示问题,Unicode 学术学会发布了 Unicode 标准。在 Unicode 之前,全球有数百种独立编码系统,但是没有一种可以包含足够的字符,而且这些编码系统也会互相冲突。也就是说,两种编码可能使用相同的数字代表两个不同的字符,或使用不同的数字代表相同的字符。任何一台特定的计算机(特别是服务器)都需要支持许多不同的编码,数据通过不同的编码或平台保存,时时有损坏的危险。Unicode 给每个字符提供了一个惟一的数字,不论是什么平台、什么程序、什么语言。Unicode 标准已经被工业界许多公司所采用,我们使用的 Windows 2000 以及 Windows XP 操作系统都是基于 Unicode 字符集的,所有最新的浏览器和许多其他产品也都支持它。

目前 Unicode 的最新版本是 Unicode 4.1.0,字符总数达到了 9 万多个,其中中、日、韩汉字总数达到 7 万多个。[3]

3 计算机显示与输入汉字的原理

目前计算机大多使用 Windows 操作系统,这里给出 Windows 2000/XP 操作系统下的汉字显示原理图,如图 1 所示。

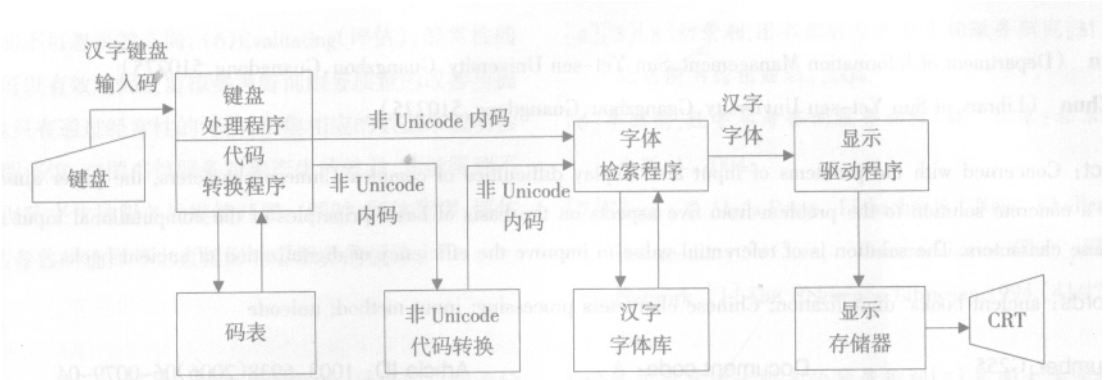

图 1 Windows 2000/XP 汉字输入及显示原理图

在输入汉字时,通过键盘输入的汉字编码,首先要经代码转换程序转换成汉字机内代码。转换时要用输入码到对应的码表中检索机内码。如果所用的输入法支持 Unicode,则得到 Unicode 内码;如果所用的输入法不支持 Unicode,则在得到非 Unicode 内码后,须经非 Unicode 代码转换程序转换得到 Unicode 内码。操作系统是以这样的 Unicode 内码形式保存字符数据的。当需要显示时,字体检索程序利用 Unicode 内码检索汉字字体库,查出相应的汉字字体并送显示驱动程序,在显示器上显示出来。

Windows 98 及以前的操作系统在汉字输入及显示原理上与 Windows 2000/XP 略有不同,并未使用 Unicode 内码作为操作系统的内码进行处理和保存。除此之外,汉字的处理流程基本上是类似的。

从图 1 可知,汉字输入时需要代码转换程序将汉字输入码转换为机内码,而代码转换程序是输入法程序的一部分。也就是说,输入法程序决定了汉字输入码和机内码的对应关系。一方面,如果输入法没有定义某个汉字机内码与输入码的对应关系,那么这个汉字就无法输入到计算机中,另一方面,机内码的编码方案也决定着其所能表示的不同汉字的数量。例如,GB2312(1980 年)一共收录了 6763 个汉字和 682 个其它符号,而最新的 Unicode 字符集收录了多达 7 万个汉字。由此可见,要想从根本上解决大量繁体汉字的输入问题,必须从两个方面着手,即:使用支持 Unicode 字符集的操作系统和选择支持 Unicode 字符集的输入法。

在汉字显示过程中,操作系统的字体检索程序根据字符的 Unicode 内码,在汉字字体库中找到相应字体,并以图形方式显示在显示器上。汉字字体库是这一过程的关键环节。如果汉字字体库不够大,一些不常用汉字的内码就有可能找不到与之对应的字体,那么显示在屏幕上的将会是乱码。因此,安装大的汉字字体库,并在编辑软件中选择合适的字体库,是解决汉字正确显示问题的根本途径。

4 汉字的正确显示与录入应注意的问题

在古籍数字化工作中,应从以下几方面着手来解决汉字

的显示与录入存在的问题。

4.1 Windows版本

目前比较常用的Windows操作系统的版本从低到高依次为:Windows 98、Windows 2000和Windows XP。Windows 98是基于ANSI字符集开发的,因此在系统内部处理内码时都以一个字节为处理单位。如果需要处理汉字,就需要相应的附加模块来识别和处理汉字的两字节内码。这种系统级别上的先天不足,导致大量的汉字内码需要相应模块进行转换,从而降低系统运行效率。除此之外,这种单字节和双字节混排的编码方式为汉字及其它字符提供的编码空间也比较有限。

Windows 2000/XP是使用Unicode进行开发的,各种字符内码基本上以两个字节为编码单位,所以它从根本上解决了Windows 98存在的字符处理效率低的问题。Unicode统一了现存的绝大多数字符集,为各种字符提供了统一的编码,这样无论何种文字都可以在Windows 2000/XP系统中正常显示,而且是同屏显示。例如,我们可以在同一文档中编辑中文、日文、韩文等各种字符,但不会出现乱码。

对于古籍数字化工作而言,我们最关心的是能否处理数量众多的生僻汉字。Unicode编码了7万多个中、日、韩汉字,完全有能力满足绝大部分汉字处理、保存的要求。因此,笔者建议安装Windows 2000/XP中文版操作系统,从而彻底解决字符内码冲突的问题。

4.2 设置正确的区域

在"控制面板->区域和语言选项->区域选项"中选择合适的选项。这个选项影响到某些程序显示日期、时间、货币和数字的方式。例如,安装了台湾或香港的软件,为了显示正确,就应设置该项目为"中文(台湾)"或"中文(香港特别行政区)"。此外,有些输入法(例如海峰五笔输入法)要求区域语言设置为"中文(中国)"。因为在其它的区域语言设置下,输入法默认调用的系统IME函数不同,所以不能正常使用。

4.3 设置正确的非Unicode程序语言

最新开发的软件产品基本上都是支持Unicode字符集的。如果在工作中需要使用较早前开发的非Unicode编码的程序,则需要在"控制面板->区域和语言选项->高级->非Unicode程序的语言"中选择与该程序的语言版本相匹配的语言。通过选择原始语言,才可以正确显示程序中的菜单和对话框。例如,如果需要使用旧的由香港开发的繁体中文程序,则应选择"中文(香港特别行政区)"选项。

另外,在"控制面板->区域和语言选项->高级->代码页转换表"中列出了当前计算机上所安装的代码页转换表。通过添加相应转换表,Windows能够解释非Unicode程序中使用的字母、汉字以及其他字符,并在它们与Unicode之间进行转换。如果需要使用基于某个代码页而编写的程序,则必须添加该代码页的转换表。

4.4 选择合适的输入法

输入法对于汉字输入而言是至关重要的,它决定了汉字的输入能力。我们知道,汉字输入时需要输入法的代码转换程序将汉字输入码转换为机内码,其间用到输入法自带的码表。这个码表定义了汉字输入码和机内码的对应关系。对于某个汉字而言,如果在码表中能够找到它的输入码和机内码的对应关系,则这个汉字是可以输入的;反之,如果码表中没有定义该字的对应关系,这个字将无法通过该输入法输入计算机。

目前因特网上可以获取的中文输入法种类繁多,特色各异。笔者通过因特网查询了最为常用的若干种中文输入法在汉字输入能力方面的特点,见表1。

表1 中文输入法汉字输入能力比较

输入法名称	最新版本	支持的最大字符集	可输入汉字数量	支持包	软件类别
王码五笔	98版	GB18030	2万7千多汉字		免费
万能五笔输入法	2005版	GBK	2万余汉字		免费
海峰五笔	86+98标准通用版V8.0	UNICODE的CJK、CJK-ExtA、CJK-ExtB	7万余汉字	通用Unicode字体支持(UniFont.exe)	免费
新概念五笔	2004企业版(UNICODE SuperCJK)	UNICODE的CJK、CJK-ExtA、CJK-ExtB	7万余汉字	1、微软超大字符集Surrogate支持包。 2、Windows Longhorn新细明体-ExtB字体(MingLiU-ExtB.ttc)	共享

仓颉输入法	第五代世纪版	UNICODE 的 CJK、CJK-ExtA、CJK- ExtB	7 万余汉字	1、微软超大字符集 Surrogate 支持包。 2、Windows Longhorn 新细明体- ExtB 字体(MingLiU-ExtB.ttc)	免费
晨曦五筆輸入法	簡繁捆綁版 7.0 正式版	GBK	2 万余汉字		共享
智能陈桥五笔	5.601	GB18030	2 万 7 千多汉字	微软 GB18030 支持包(GBEXTSUB.msi)	免费
郑码输入法	5.0	GBK	20902 个汉字		免费
微软拼音输入法	2003 加强版	GBK	2 万余汉字		免费
紫光华宇拼音输入法	3.0 版	GBK	2 万余汉字		免费
全拼输入法	5.0 版	GBK	2 万余汉字		免费

注: 输入法所需的支持包都可从因特网上获得。

从上表可以看出, 不同的中文输入法在汉字输入能力上的确存在比较大的差别。当然, 仅从汉字输入能力上来评判一种输入法的优劣未免有失偏颇。因为一般而言, 可输入汉字的数量越大, 输入码的重码率也就越高, 而对重码的选择降低了该输入法的输入效率。因此, 针对不同性质的工作, 应选择合适的输入法, 以达到较高的输入效率。在古籍数字化工作中, 由于经常会进行生僻汉字的输入, 支持 Unicode、可输入汉字数量庞大的输入法就成为最佳选择。

4.5 安装支持 Unicode 字符集的字体

如果说输入法决定了汉字的输入能力, 那么字体包就决定了汉字的显示能力。汉字的 Unicode 内码通过字体检索程序, 在汉字字体库中找到相应字体, 从而以图形方式显示出来。Windows 2000/XP 下默认安装的字体库已经能够支持 GBK 字符集 2 万余个汉字的显示。但对于超出这 2 万余个汉字的其它汉字而言, 由于没有与之对应的字体, 那么显示在屏幕上的将会是乱码。解决这一问题的办法是安装超大字符集支持包。超大字符集支持包可以从微软网站免费获得。还需注意的是, 安装超大字符集通常只是扩充了某一种或几种字体集。进行汉字输入和显示时, 需要在编辑软件中选择已被扩充的字体集作为显示字体, 否则也将会出现某些字无法显示的问题。

5 结束语

通过以上五方面的正确设置, 就可以输出 Unicode 所包含的多达 7 万个汉字了。如果还有个别不能输入的汉字, 则需要通过 windows 自带的 TrueType 造字程序造字。当然, 如果希望别的电脑也能正确显示这些造出的字, 在文档被拷贝到目标机器的时候, 包含有新造字的字体也应安装在目标机器中。

参考文献:

[1] 中文维基百科的编者和开发者.自由的百科全书——维基百科条目的精选[EB/OL].http://zh.wikipedia.org/wiki/Wikipedia:维基读本/2005 年第一期, 2005, 1(1):5.

[2] 全国信息技术标准化技术委员会. GB18030 介绍及其与相关标准的比较[EB/OL]. http://www.nits.gov.cn/sc2/jishufile1.asp, 2005.

[3] Unicode 学术学会. 什么是 Unicode(统一码)[EB/OL]. http://www.unicode.org/standard/translations/s - chinese.html, 2005.

[4] 闫凡蕾, 林仲湘, 李龙.古籍电子化中生僻汉字的处理[J].华侨大学学报(自然科学版), 2003, 24(3):331-334.

作者简介: 徐健(1977-), 男, 广东省中山大学资讯管理系讲师; 肖卓(1978), 女, 广东省中山大学图书馆特藏部助理馆员。