

东方多语种文献计算机管理初探

周 宁

计算机在文献管理中的应用开辟了文献信息生产与消费的新时代。东方多语种的文献管理是世界上多语种文献信息处理的难点。文章系统地论述了该领域研究的应用现状和需要解决的几个主要问题。分别对东方多语种字符统一编码、输入/输出接口、专用软件设计与标准化等问题进行了具体探讨。

在当今信息时代,文献信息处理已从手工管理发展到计算机自动处理的新阶段。无论是出版社编辑出版图书、期刊,还是办公室编印文件,计算机都发挥了重要作用。近两年中,报社在出版发行多语种报纸方面,计算机已成功地扮演了主角。它以快速、方便、高质量、高效率赢得了广大用户。

计算机在文献管理方面的应用,开始于60年代后期。当时主要是英文文献信息处理,后来发展到包括法语、德语、意大利语、西班牙语等西文多语种文献信息处理。东方多语种文献信息的计算机管理,发展于70年代中期至80年代初。它主要是解决汉字、日文、朝鲜文文献的计算机管理^①。在多语种文献信息管理中,西文多语种系统由于其字符集小(字符的个数 $n < 255$)、输入方法是单键直接键入字符,处理对象为单字节、输出方式单一,因而简便易行。而东方多语种文献管理系统则由于其字符集庞大(字符个数 $m > 50000$)、用多字节编码才能输入一个字符,处理的对象是多字节,输出方式多种多样,因而它是世界上多语种文献信息处理的难点。

80年代,我国在少数民族文字计算机处理方面发展迅速。主要是在微机上开发了一批单语种系统^②。例如,蒙文系统、藏文系统、维吾尔文系统、朝鲜文系统等。目前用得好的是编辑、排版、印刷系统。例如,新疆日报在每天出版汉字、蒙文、维文、哈文版四种报纸。他们已经废除了铅字排版,全部采用计算机编辑、排版。其他报社近两年也先后实现了报纸编辑排版的电子化。这说明我国在研究多语种文献信息处理方面已取得了可喜成绩。由于这些系统都是面向局部的、分割的单语种系统,因而它不适合图书馆、情报所、档案馆多语种文献信息的综合管理。这样,研制面向全局的多语种综合处理系统就成了我们面临的一项重要任务。

一、多语种文献综合管理所遇到的问题

我们知道,图书馆、情报所、档案馆等单位的计算机管理系统,必然是一个面向全局的多语种文献综合管理系统。绝对不是一批单语种文献信息处理系统。因而,上面所介绍的面向局部的单语种系统是不能胜任的。显然,面向全局的多语种综合管理需要有相应的支撑环

境。这一环境就是：计算机系统要有足够的编码空间供多语种字符统一编码用；系统配备有相应的多语种字符输入方法；字符的内、外码对照表及键盘重定义；并配有相应的输出方式的支持。在综合处理多语种文献信息时，由于字符编码的唯一性而能自动识别文献的语种。

仅就上述支撑环境而言，现有的单语种系统都不具备这些条件。从国内目前在微机上实现的蒙文系统、朝文系统、日文系统、藏文系统、维吾尔文系统而言，除了朝文系统外，一般都能处理汉字，并保存有完整的一、二级汉字库。实际上，这些系统(除朝鲜文系统外)就其处理功能可以称为蒙、汉文系统，日、汉文系统，藏、汉文系统，维吾尔文、汉文系统等。它们分别以GB2312—80标准为基础进行了扩充与复盖，均用一个 94×94 的平面为编码空间。由于GB2312—80已安排1—9区存放特殊符号、拼音字母、日文假名、俄文字母、希腊字母和制表字符等，而16—55区为一级汉字，56—87区为二级汉字。这样，94个区就已用81个区，还剩下13个区可作为扩充部分。它可扩充1200多个字符。在现有的蒙汉文系统、日汉文系统、维汉文系统、藏汉文系统中，由于各个系统所要扩充的常用字符均未超出1200个，因而，一个 94×94 的编码空间为它们分别实现提供了可能。

而朝鲜文系统，由于其字符多，无法在1200个字符编码上实现。因而，它采用了全部复盖或部分复盖的方法。《信息交换用朝鲜文编码字符集(GB12052—89)》中有5258个字符编码。它占用的编码空间刚好是16—72区，几乎复盖了GB2312—80中的一、二级汉字。由此看来，在微机上只用一个 94×94 编码平面，要同时实现朝、汉文系统是困难的。有的系统采用了折中方案。它以GB2312—80为基础，保留一级汉字，而将GB12052—89朝鲜文中3384个常用字符复盖二级汉字实现朝、汉文处理系统。这样做虽部分解决了兼容性问题，由于两个标准都残缺不全，字符数太少，因此很难达到实用水平。

上述一系列问题，提示我们必须扩展编码空间，将多语种字符统一编码，才能实现对多语种文献的综合管理。

二、两种编码方案

多语种文献计算机综合管理系统须有一整套相应的支撑环境。首先，必须有统一的多文种字符编码方案。

鉴于现行的计算机系统均为8bit编码，而即将问世的新一代计算机系统则采用16bit编码^③，所以多语种字符的编码方案必须有两套^④。一套适用于现行的计算机系统，而另一套配备给新一代计算机系统。

根据编码平面的容量，我们把前者称为“小方案”，后者称为“大方案”^⑤。

“小方案”是专为现行的计算机系统设计的。它有两个 94×94 编码平面。一个是‘1—1’平面，它以GB2312—80为基础，扩充维吾尔文、哈萨克文、柯尔克孜文字符(见图1(a)所示)。另一个是‘1—0’平面。它专门存放日文、朝鲜文、蒙文、藏文等字符集。具体布局见图1(b)所示。

大方案是为新一代计算机系统而设计的，它有两个 256×256 编码平面。一个为16bit编码平面(见图2(a)所示)，即基本的多语种字符编码平面。它是多语种常用字符集。另一个为32bit编码平面。它存放的是古字、偏字和不常用的多语种字符集(如图2(b)中所示)。

两种编码方案都考虑到与现行标准的兼容性。在现行的计算机系统中，我国大陆使用双字节编码，而我国港、台地区采用三字节编码，有的用四字节编码。为了达到数据资源共享，可以实现不同编码方案之间的字符代码转换。同时，只要进行大、小编码方案的代码转换就

1	专用字符集(3个区)
2	
3	
4	日文假名字符(2个区)
5	
6	希腊文字符(1个区)
7	俄文字符
8	制表字符
9	西文字符
10	维吾尔文、哈萨克文、柯尔克孜文 字符
15	
16	
55	
56	
87	一级汉字
88	二级汉字
94	扩充的汉字库

图 1 (a) '1-1'平面编码区布局
(基本字符集编码)

1	
1	
14	
15	
16	
17	
48	
49	
50	
75	
76	
82	
94	

1	蒙文字符编码区
1	
14	
15	
16	
17	
48	
49	
50	
75	
76	
82	
94	

1	GB2312-80中没有的常用CJK字符 (汉字)
1	
14	
15	
16	
17	
48	
49	
50	
75	
76	
82	
94	

1	傣文字符
1	
14	
15	
16	
17	
48	
49	
50	
75	
76	
82	
94	

1	朝鲜文字符集
1	
14	
15	
16	
17	
48	
49	
50	
75	
76	
82	
94	

1	藏文字符集
1	
14	
15	
16	
17	
48	
49	
50	
75	
76	
82	
94	

1	其他文种字符
1	
14	
15	
16	
17	
48	
49	
50	
75	
76	
82	
94	

图 1 (b) '1-0'平面字符编码
(辅助字符集编码)

OO	ISO 通用字符编码区
OO	
O1	
AO	
A1	
A2	
A3	
A4	
A6	
A7	
A8	
AB	
FF	

图 2 (a) 基本的多语种字符编码平面
(16 bit 编码)

OO	CJK辅助字符集编码区
OO	
9F	
AO	
AB	
A9	
FF	
FF	

图 2 (b) 辅助的多语种字符编码平面
(32 bit 编码)

可以实现新、老计算机系统之间的数据共享。由于小方案编码空间较小,一般来说,小方案的代码转换为大方案相应的字符代码比较容易,而逆操作要视具体情况而定。它只能在小方案所能容纳的范围内实现大方案代码向小方案编码之间的转换。

三、多语种字符的输入接口

字符的输入是信息处理的基础和前提。

多语种字符的输入方法由各语种按其字符个数多少和字形特点及规律确定。一般有两大类:

1. 小字符集的输入接口设计

小字符集一般指一个语种的字符个数在160个以下的字符集。它以现行的通用文字和有关的标准为依据。例如,《信息交换用蒙古文16×12、16×8、16×4点阵字模集及点阵数据集》(GB7422·1—7422·2—87)中规定蒙古文有95个字符。《信息处理 信息交换用维吾尔文编码图形字符集》(GB12050—89)中规定,现行的维吾尔文字有142个字符。

又如,俄文有66个字符,傣文、越南文字符的个数也较少。

这些小字符集的语种其字符输入只须对键盘重定义就可以实现。在输入模块中,将标准键盘上的字符与某一语种的字符集建立一个对照表就行了。在输入时,接口软件将从键盘上接收到的标准键符代码立即转换成相应语种所对应的字符进行显示和代码存贮。这就实现了一个语种的字符输入。不同语种的输入状态,通过组合功能键来进行切换。一旦切换到某语种的输入状态,输入模块立即激活了该语种的字符集对照表和相应的程序段,使系统处于接收该语种字符状态。

2. 大字符集输入接口的实现方法

很多语种都是大字符集。如,汉字、日文、朝鲜文、藏文等。

汉字的输入方法很多,如五笔型、联想、区位、首尾、拼音、快速、双拼、全拼等。它们都是将汉字的输入码与机内码建立一个对照表,输入程序自动完成输入码向机内码的转换。这些都是我们早已熟悉的常用方法。

日文的输入分日文假名(平假名和片假名)和日文汉字两类。我国制定的相应标准中,规定日文假名是169个㊦字符(而OCLC CJK 350系统所用的日文假名为174个㊦字符)。它可以按小字符集方法处理,重新定义键盘即可。而日文汉字的输入可用类似于中文汉字的输入方法实现。

朝鲜文字的输入可按其字形结构输入。根据字形可将一个朝鲜文字符分解成1—3个部件。按照结构分类建立输入码与机内码的对照表。每一结构部件在标准键盘上有相应的键符对应。由于输入码为可变长,所以键入一个字符的输入码后加一个空格键以示结束。这样,每输入一个朝鲜文字符,可以通过一键加一空格、或两键加一空格、或三键加一空格来完成。朝鲜文的输入方法与汉字的字形码相似。

藏文的输入要比汉字、日文和朝鲜文简单。藏文字符分别由34个字母、3个上加字和5个下加字的不同组合构成。因此,藏文字符可以拆分成1—4个部件。输入藏文字符时即按字符的字形编码来实现。一个字符的输入码由1—4个键符组成。不同字符输入码的分隔也是用一个空格来区分。这样,输入一个藏文字符可用一键加一空格、或两键加一空格、三键加一空格、四键加一空格实现。

各语种的输入状态由组合功能键来激活,切换非常方便。

四、三种输出方式

由于历史原因,多语种文字的书写要求与排列方式很不一致。大多数语种的字符要求从左至右横行排列,少数文种要求从右至左横行书写,个别语种的文字要求从上至下直行书写。为了与传统习惯一致,计算机系统也必须设计三种输出方式。

东方多语种文献信息三种输出方式如下:

1. 从左至右横行输出方式

汉字、日文、朝鲜文、藏文、傣文、壮文、越南文等都要求从左至右横行书写与排列。它们的输出方式与英文、俄文的要求一致，方法简便易行。

2. 从右至左横行输出方式

维吾尔文、哈萨克文、柯尔克孜文的书写习惯是从右向左横行排列。计算机存贮与处理

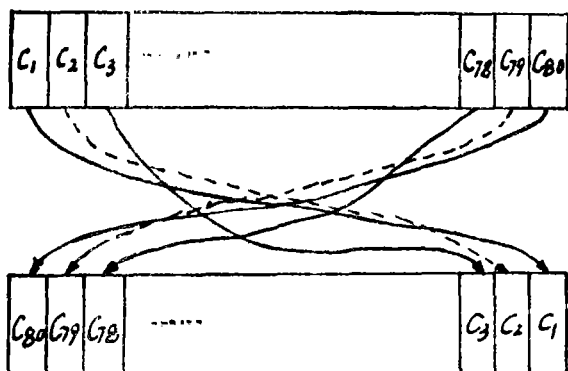


图3 输出缓冲区的映射示意图

的方法虽然是统一的，但显示与打印时必须与传统习惯一致，它的实现方法如下：

计算机内部文件的字符是依次顺序存放的。需要输出(显示或打印)时，先将数据流依次填入缓冲区Buffer1(假设输出列为80个字节)，然后通过一个简单的函数变换，将Buffer 1中的字符序列映射到输出缓冲区Buffer 2(如图3所示)。再从Buffer 2中按常规方法进行输出。这样就符合维吾尔文等少数语种的显示与打印要求。

3. 从上至下直行输出, 从左至右换行

这种输出方式是专为蒙古文、满文设计的。由于蒙古文字符有三种高度 (16×4 、 16×8 、 16×12)，因而蒙古文采用变长编码。它分别用单字节、双字节、三字节三个不同的码长编码。蒙古文字符的显示靠光标从上至下移动时控制，根据字符高度要求显示点阵字模，并相应地移动光标位置。蒙古文的打印比较复杂。根据传统的印刷习惯，蒙文字符的高度不是三种，而是十多种。打印字符点阵为 $24 \times X$ ，其中 $X = 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16, 19, 21, 22, 24$ 等。打印缓冲区两个，一个存放打印字符的机内码索引表(表中有两项，一项是上述的 X 值，另一项为点阵地址)。另一个缓冲区存放点阵字模信息。打印头的移动由 X 值控制。

五、专用模块的设计

有了统一编码和输入/输出接口的配置，综合管理多语种文献就有了基础。在常用的文献管理中，计算机编目、建立文献数据库、多语种联机文献检索、图书情报档案自动化管理等一系列项目是我们研究和应用的重点。

1. 编目模块

编目是多语种文献计算机管理的基础性工作之一，只要计算机编目工作投入正常运行，其他管理工作就有了基础。当计算机编目累积到一定规模后，建库、联机检索、编辑排版、咨询服务等就可以顺利实现。

编目软件分三个子系统^⑥实现。它们分别是中文编目子系统、蒙文编目子系统、维吾尔文编目子系统。编目模块的主要功能是：数据文件的建立、校验与维护，标准化数据文件(ISO2709格式或GB2901格式)的生成，查重、输出编目产品，最新资料报导与分析统计等。各子系统的功能相似，但实现的方式、方法各异。

2. 建库模块

建库模块是完成联机文献库生成的专用程序模块。一般借助于编目系统的副产品——文

献数据文件生成。建库模块包括数据转换、主文档的组块装入与主文档索引文件的生成、抽词、排序、归并、倒排档的装入与倒排索引文件的生成等。

多语种文献库的建立一般分语种进行。这样做将方便管理。多语种混合建库也行，但考虑到输入/输出接口问题，最好是将多语种文献分三类建立混合型文献库。

3. 联机文献检索

多语种文献联机检索是提供服务的一个重要手段。该模块可以将用户提问直接操作联机文献库。既可进行单途径的快速检索，也可以进行多途径组配检索、及在此基础上的限定检索和特定要求的编辑输出等。

多语种文献库若是分语种建库，则检索程序配有三个并列的程序段。分别处理汉字型文献库、蒙文型文献库和维吾尔文型文献库。若是将三种类型的文献混合建立一个数据库，则系统均只用一个程序段(中文状态)来处理。这时的维文字符也视同汉字处理，蒙文字符向逆时针方向旋转90°躺倒从左向右横行输出。这样，所有的多语种文献均可用一种方式处理，检索模块的规模也将相应地缩小。

4. 其它管理模块

多语种文献管理还有许多专用模块。如，编辑、排版，自动标引，流通管理等都是常用的管理模块。这些模块的处理思想与基本方法与上述三个模块相似。既可以在程序设计中用并列的三个程序段来实现，也可以用一個程序段进行综合处理。系统设计人员可以根据用户要求进行选择。

六、标准化与智能接口

为了方便多语种文献信息的数据共享，就必须逐步实现文献信息管理的标准化。其内容就是支撑环境的标准化，文献数据加工处理的标准化、建库工作的标准化等。

如果我们所建立的文献库都是按语种分别建立的，而且文献数据库的名称也用该语种的有关字符定义；那末，当系统转入某一数据库操作时，系统工作方式状态不是用常规方法靠人工干预进行切换，而是由系统自动识别数据库的语种，自动切换到所对应的系统状态。这就要求我们配上相应的智能接口，从而完成系统智能化的目标。

参 考 文 献

- ① [美] OCLC 亚太部 王行仁：OCLC 在传播中国文化中的作用。《图书馆新技术应用国际学术讨论会论文选》西安交通大学出版社1991年6月。
- ② 周 宁：东方多语种文献信息处理的进展《情报科学》第12卷第5期，1991年10月。
- ③ Unicode—A Character set standard for all language now in draft form. Supported by the Unicode Consortium which has representatives from IBM, Apple, Microsoft, Next, Xerox, RLIN. 1990. 12.
- ④ ⑤ Zhou Ning: STUDY ON MULTILINGUAL DOCUMENT DATABASE. Paper to be presented to The 3rd Beijing International Symposium on Computerized Information Management. 1991. 10.
- ⑥ 中华人民共和国国家标准 GB2312—80《信息交换用汉字编码字符集 基本集》。
- ⑦ Andrew H Wang: Feature of OCLC CJK 350 system, OCLC online Computer Library Center, Inc. Jan. 19, 1990
- ⑧ 林 蓝：东方多语种文献编目方法的初步研究。中国图书馆学会第四届图书馆自动化学术讨论会论文。1991年11月 深圳。

(本文责任编辑 江 平)