西夏文数字化的现状与未来

景永时

摘　要：历经多国学者近半个世纪的努力，西夏文的数字化逐渐得以实现并被不断完善。字库设计在技术层面上渐趋成熟，而亟待一款更为标准的西夏字体以满足学术需求。同时字符的存放区域、录入编码和系统兼容等问题也是这一工程所要面临的问题，笔者对上述问题都提出了相应的建议及解决措施，以期推进西夏文数字化的进程。

关键词：西夏文　数字化　字符　编码

一

20世纪70年代，伴随着计算机技术的发展和在文字处理方面的应用，西夏文数据化问题也被提了出来。

现在有人认为，西夏文数字化最早是从格林斯蒂德开始的，具体成果是《西夏文字的分析》[①]一书。格林斯蒂德使用的西夏字表是俄国西夏学家苏敏整理的，也是当时所能见到的收字最全者。这个字表共收录了5819 个西夏单字。[②]格林斯蒂德的编码类似于“电报码”，[③]即一字一码。但笔者认为，格林斯蒂德给西夏文所作的编码只是一种检索代码，而非是电脑字符的录入代码。正如李范文先生的四角号码检字表一样，当初只是为检索西夏字而编制，后来人们觉得它体现了西夏字的构造特点并可以用于电脑西夏字符的录入代码，这样才由一般的检索代码变为电脑录入代码。格林斯蒂德的编码也有这种可能，但它因为是一字一码，既便是用于电脑录入代码，也不能提高录入速度，故不可能被采用。

据聂鸿音教授说，20 世纪80 年代末，荷兰莱顿大学的藏缅语言学家范德利姆（George van Driem）与俄罗斯科学院圣彼得堡东方研究所的西夏语言学家克平（Ксения. Болисовна Кепинг）共同就西夏文数字化做过努力。不过，工作在两年以后就由于技术上的原因而中辍了，因为他们只能实现“左右结构”的文字录入，而一旦遇到“上中下结构”的字时就无计可施，所以最终通过这项研究进入电脑的西夏字只占全部字数的三分之二左右，而仅靠这些文字是无法进行文字和文献处理的。

宁夏社会科学院的科研人员也曾做过尝试，曾任该院科研处处长的林清研究员于1991申报国家社会科学基金项目，试图用点阵法制作西夏文字符，由于技术问题，该项目未能完成。

20世纪90年代，电脑技术在全世界范围内取得了极大的进展，其速度之快是令人始料不及的。由于各种字符制作工具软件的问世，使当年难以实现的字符制作问题变得不再神秘，所以在20世纪90年代以后，在世界上陆续出现了多种西夏文字符集。

① Eric Grinstead, *Analysis of the Tangut Script*, lund 1975.
② М. В. Софронов, *Грамматика тангутского языка*, кн. 2, Москва:《Наука》, 1968, стр. 279-403。
③ Eric Grinstead, *Analysis of the Tangut Script*, p.70。

1996 年，由中岛干起主持研制的第一个完整的西夏字库在日本东京外国语大学亚非语言文化研究所问世。[①]随后的几年，这个研究所将这套字符用于西夏文字研究，即利用电脑对西夏文字进行各种统计分析；[②]同时，也用于出版一些西夏学学者的研究著作。[③]不过中岛干起主持研制的西夏文字符集，除了本单位出版和运用于几种文献的分析研究外，并未曾得到推广。

1999 年，由宁夏大学计算中心主任马希荣教授主持国家自然科学基金项目，开发研制出《夏汉字处理及电子字典》，[④]其中就有西夏文字符集。与日本东京外国语大学的字符集相比，宁夏大学的“夏汉字处理系统”具有更强的实用性。它建立有6000 号西夏字的两种字库，并实现了西夏字的任意缩放输出，且开发有西夏字与汉字、西文混合排版、编辑的具有一般文字处理功能的窗口。但由于课题组成员对西夏语文过于陌生，又为了“赶时间、赶任务”而未能进行缜密的校核，所以其中的瑕疵似乎比人们可以容忍得要多。事实上，《夏汉字处理及电子字典》中的字符集除去继承了《夏汉字典》的几乎全部印刷错误之外，自己还新添了一些失误，例如错码、漏码、错字、漏字等。对于编码上的错漏，软件使用者还可以利用其中的“西夏字总表”来彻底查找，至多影响录入的速度而已，但对于字符本身的错漏，使用者就没有办法了。同时，由于《夏汉字处理系统》的西夏文字录入只能在本系统的窗口中实现，加上窗口功能的过于简单，使用者不得不先将西夏文录入系统所设窗口中，然后用复制粘贴方式将所需西夏字转到其他文字处理系统之中。最大的问题还在于，这个系统将西夏字符放在4个汉字字体的补字区空位码里，即在不同字体的补字区同一码位放置着不同的西夏字符，在使用中稍有忽视就会出错。有时候安装该系统的电脑在某些软件运行时，屏幕中会出现乱字符，对其他软件的正常使用造成干扰。总而言之，西夏学研究者如果需要在自己的著作中插入不太多的西夏字，那么利用这个系统还算是比较方便的。但如果需要大规模地处理西夏文献或者对西夏文字进行穷尽式的研究，则是“夏汉字处理系统”难以达到的。

21世纪初，中国台湾中央研究院历史语言学研究所制作了一套西夏字符集，同时研制了西夏方属性数据库。但由于该系统运行环境是繁体汉字WINDOWS2000以上版本，而且并未将字符集和数据库公开出版发行，人们也只见到使用该字符集出版的台湾学者的成果。[⑤]

另外还有几种西夏字库，即有原宁夏大学西夏学研究中心韩小忙教授制作的字库、日本今昔文字镜研究会制作的字库和小高裕次制作的字库。韩小忙字符集只能在“北大方正排版系统”中应用，而且只能由他个人使用。日本文字镜研究会字符集，目前所能见到的有使用该字符集出版的书籍。除此之外，有一个电脑公司也制作有一套西夏字库，据说是以《夏汉字典》（1997版）中的西夏字作为底本而制作的，我们在互联网上可见到这套字库的样品，但尚未见过该字符集的正式出版物。

2005年，笔者和贾常业研制成《基于北大方正典码之上的西夏文字录入系统》，[⑥]该系统借用汉字楷体笔画制作了西夏文字符集。由于这个字符集从一开始就是以研制供公众使用为目标，所以该套字符在准确性方面优于其他字符集，又因兼容性较好，故被广泛使用于电脑写作和书刊的排版印刷。2007年，该字符集还被国际标准组织（ISO）选定为《信息技术通用多八位编码字符集（UCS）》——西夏文编码字符集。该套字符是目前所有西夏文字符集中字形最准确、收字最多者，也为较多国内学者所采用。

---

① 中岛干起编，今井健二、高桥まり代协力《西夏文字の研究に向けて》，东京外国语大学アジア•アフリカ言语文化研究所，1996 年。

② 例如中岛干起、今井健二、高桥まり代《电脑处理西夏文字诸解对照表（稿）》，国立亚非语言文化研究所，1998 年。又《西夏文献〈文海〉反切法解析》、《电脑处理西夏文字字素分析》，国立亚非语言文化研究所，2000 年。

③ 李范文、中岛干起《电脑处理西夏文杂字研究》，国立亚非语言文化研究所，1997 年；史金波、中岛干起《电脑处理西夏文《文海宝韵》研究》，国立亚非语言文化研究所，2000 年。

④ 马希荣《夏汉字处理及电子字典》（光盘版），清华大学出版社 1999 年。

⑤ 应用样品见龚煌城《西夏语文研究论文集》，中央研究院语言学研究所（筹备处）《语言暨语言学》专刊丙种之二（上），2002 年；林英津《西夏语译〈真实名经〉释文研究》，《语言暨语言学》专刊甲种之八，2006 年。

⑥ 景永时、贾常业《基于方正典码之上的西夏文录入系统》，香港社会科学出版社 2005 年。

二

目前，我们所知道的较完整的西夏文字符集共有八九套之多，有科研机构开发研制的，也有电脑公司研制的，还有个人或学术团队开发研制的，以及纯个人使用的，但还没有一款学者们公认的标准字符集，存在问题有二：一是字形准确度问题，二是字体风格问题。

从多年来西夏电脑字库的设计实践中不难看到一个趋势，即字库设计在技术层面上的问题已经逐渐退居次要地位，而在学术层面上的问题则越来越凸现了出来。换句话说，西夏字形的准确性已经成为衡量字库是否成功的首要指标。西夏字笔画繁复，有些字的区别只在毫厘之间，不经过较长时间的系统学习很难领悟，我们不能指望依靠一个现成的西夏字符表就能解决所有的问题。退一步说，即使有了一个准确率极高的字符表，我们在电脑摹写过程中的“笔势”也往往会产生鲁鱼亥豕之差，而这往往是电脑技术人员再细心也难以觉察到的。

根据目前的西夏文字研究水平，要求设计出一个“百分之百正确”的字符集是不现实的，这是因为我们始终没有一套判定何为正字的科学标准，只能由西夏学家凭个人的经验去“死校”。西夏在近二百年的历史上并没有进行过文字规范工作，自然也没有“正字”之类的文献保存下来，现在的学者们确定西夏“正字”的文献依据主要是西夏字典《文海》中的字形说解，然而如上所述，这些字形说解并不一定反映了造字的初衷，何况保存至今的《文海》刻本自身也还有待校勘。[①]

毫无疑问，经验在校订西夏文字时固然重要，但我们更需要找到一种为大家所公认的科学方法以避免没完没了的争议。设想一下，由同一个人把同一个字连续写十遍会是什么结果？用极端精确的标准衡量，应该是十个样子，只不过谁也不会认为字被写错了，这是因为在文字使用者头脑中事先已经形成了对这个字的判断标准，这个标准不是一个“点”，而是一个“面”。只要在这个“面”允许的范围之内，出现些许差异绝对不会影响人们对文字的识别。我们说的这个“面”可大可小，而且因字而异，关键在于会不会使人误会成另一个字。以汉字为例，当我们写“三”字的时候，第一横长还是第二横长是无所谓的，而当我们写“末”字的时候，第一横长还是第二横长就有所谓了，因为写不对就会使人误会成“未”字。我们可以把汉字中平行两横的相对长度称为“最小区别”。如果能根据不同字之间的比勘来确定西夏字的“最小区别”都有那些类型以及每种类型的出现条件，那么我们的字形校勘就有了一个坚实的理论基础。就目前西夏语言文字研究水平和已发表的成果，以及公布的文献而言，是可以进行这样的工作了，但这项工作必须由具有较高研究能力的人担当。

现在世界上已有的多种西夏文字库，就字体风格而言，还没有一款是大家公认的“标准”西夏体。所以在解决字形问题的同时，还要考虑的是如何体现西夏书写风格。

现有的西夏字库使用了与汉字“宋体”、“楷体”、“仿宋体”和“毛笔楷体”相类似的几种字体。使用“宋体”的有日本东京外国语大学、日本今昔文字镜研究会和小高裕次的字体，是借用日文用宋体汉字笔画制作出来的；使用“仿宋体”的有韩小忙的字库和宁夏大学计算中心的《夏汉字处理系统》字体之一种，前者是借用电脑仿宋体汉字笔画制作的，后者是手写仿宋体。使用“楷体”的有景永时、贾常业制作的字库，是借用电脑汉字楷体笔画制作的；使用“毛笔楷体”的有台湾中央研究院语言学研究所的字库和宁夏大学计算中心的《夏汉字处理系统》字体的另一种，前者是通过扫描《同音研究》中的毛笔书写字体制作的，后者是通过扫描今人毛笔书写字体制作的。以上各种体的字库，虽然都运用于电脑写作和出版，出版了不少学者的成果，但毕竟都不是来自西夏原始文献，使人看起来总觉得没体现出西夏字体特点。理想的西夏字库应该每一个字符都应该直接来自西夏原始文献，而不是今人对某种电脑字体的改造或书写。现在的电脑字体制作技术完全仿真某一种笔体已不成问题，如汉字就有舒体、启功体等，但首要的条件是必须有个人书写的全部文字做底稿。就目前发现的西夏原始文献而言，还没有一种文献包含了所有的西夏文字符，无论是手写还是刻印的文献，字形又存在差异，有的甚至存在很大差异，所以，完完全全照搬西夏人原始文献中的字体制作新字库是不大可能的。如果

① 关于《文海》原文的讹字，参看史金波等《文海研究》第 360~361 页。

非要去做，制作出的字库肯定字体风格不统一，也就如同在一个汉字字体中有几个字是黑体、有几个字是楷体或是其他体一样。试想一下，我们在电脑上互相掺杂着黑体、楷体、宋体、隶书写成的文字，而又不属于刻意设计的美术作品，打印出来将是怎样一种情景？因此，所谓西夏风格也只能是今天人们相对的认可，而并非来自西夏人自己。西夏文字体要达到笔画准确又使人看起来具有西夏字体特点，唯有的办法是从诸多文献中选取一种字体作为标准，摹写出一套字样并制作成电脑字符。实际上已有人关注此方面的问题，大家似乎已有共识，认为最适合作为字形依据的是《同音》（或译《音同》）字典。这部字典在西夏时期先后修订过多次，今存字迹工整的刻本多种，是全部西夏文献中收字最多者，其中的丁种本和乙种本字形最为接近，可作为样本，少量残缺和讹字可以参照该种字体进行补充。

电脑字符的制作好后，存放区域、录入码编制和与某个文字处理系统的兼容等，也是数字化所面临的问题。目前，因西夏文还没有国际标准组织（ISO）认定的通用多八位编码字符集，所以各家所制作的字符存放区域各自为阵、互不兼容，文件格式也五花八门。如日本学者制作的西夏文字符有3种，一种是由日本文字镜研究会制作的字库，用了两个字库存放，即Mojikyo M202、Mojikyo M203，文件名分别为Mojikm66.TTF 和Mojikm67.TTF；另一种是小高裕次的字库，字体名TTEditFont，文件名为XXXtan.TTF。台湾字库字体名为EUDC，文件名为TANGUT.TTE。宁夏大学的字库不是独立的，而是利用了汉字的补字区，即将所有西夏文字符，分成4部分分别放在汉字仿宋、黑体、宋体和幼圆四种字体的补字区。笔者主持研制的字体名为“西夏字体”，文件名为“XXZT.TTF”，用一个字库将所有西夏字符（包括基本字与偏旁部首）存放在一起。韩小忙的字符，用的也是补字区码位，即将一套字符分成8个部分分别与8种不同形体汉字字体链接后使用。

将西夏字符放用两种以上字库存放，在使用时必须特别小心，因为录入不同的西夏字要在不同汉字字体之间切换或加注排版语言，一旦注意不到，尽管是同一字体，就会造成“张冠李戴”的错误。笔者和台湾历史语言研究院的字符集，虽然放在同一区域，本身不会重码，但却与其他字符重码，所以也会使西夏文显示成汉字或其他字符的问题。造成这种情况的原因很简单，主要是因各科研机构、电脑公司和个人在掌握到一定的计算机技术后，各自为阵，单独行动，加上不同国家或地区的文字系统以及可资利用的操作系统不同，所以制作出来的字符集，除了形体上存在差别外，每一种字符存放的区位也不同。目前，由中、美、英三国学者联合提出的《西夏文国际标准编码提案》正在接受国际标准组织的审议，这个提案通过之日，将是西夏文字符有了国际标准组织认可的编码，也就是说每一个西夏文字具有正式的国际“身份证”号码。到了那时，不管是谁制作的西夏文字符，都必须遵循国际组织认可的编码，至于字体的区别，就如同汉字有宋体、楷体、仿宋体、隶书体一样，人们可以根据自己的爱好和需求去选择使用。

三

任何一种文字制作好了字符，只是完成数字化处理的第一步，要让字符在各种应用软件系统中顺利应用，也即在电脑屏幕上显示和打印输出，还需要给字符编制录入代码并通过特定的方法调出（录入）。当然，不用再编码也可以直接利用，即直接采用区位码（内码）录入。这种输入法需要将每一个字符的区位码记住或通过检索专门的代码手册获得。在现代电脑技术十分发达的时代，这无异于在公路上开飞机。

西夏学专家李范文先生也曾就电脑录入计算机作过构想，正如他将古人所创造的四角号码检字法引用于西夏文检索一样，他想以五笔输入法用于西夏文录入。实际上李先生所编制的四角号码检字代码作为电脑录入代码，是最为快捷和科学的，使用四角号码法既易于学习掌握，又不会使学习者大费脑筋去记忆繁多的字根。宁夏大学马希荣等、台湾中央研究院历史语言研究所和笔者等，都是采用了这一方法进行编程或编码的。采用这种方法除科学合理外，还有一个好处是使用过《夏汉字典》四角号码查字法的人很快就能学会和熟练使用。至于日本的字体是用区位码逐字录入，制作出一个字符

电子文本，使用时是将所需字符复制、粘帖而实现。韩小忙字体的录入就更麻烦一些，不仅需要用内码录入，而且在录入时必须与字体注释语言相配合。综合各种录入法，应该说研制或借用某种成熟的录入法为平台，是比较好的做法，这也是经过实践证明了的。

实际上人们也在发挥现代电脑技术进行编码，一种是拆字法，即类似于汉字的五笔型和郑码之类，这种方法是根据西夏字的结构特点，将字符按各种部件拆分，然后给每一个或每类部件在键盘上指定一个按键为代码，使用时在键盘上键入代码就可以实现字符的录入。这种也是目前汉字所采用的方法之一，但西夏字与汉字不同，汉字可以按规律编制字、词和短句代码，用词语或短句代码录入，速度自然就加快。所以虽然汉字无论是五笔型还是郑码，只要花些功夫将拆字和代码编码规律掌握，以后用起来就方便快速了。西夏文字则不同，因为目前还只是单字使用而没有达到词组录入，更谈不上短句，因此编码只能以单字为主进行编码。台湾的中央研究院历史语言研究所、马希荣和笔者，都是以李范文的四角号码法为基础进行编码的，不同的是各自所采用的平台有别。台湾是利用 WINDOWS 的输入法生成器，而且运行环境 WINDOWS 又是繁体字版本，大陆学者是无法使用；马希荣是在自己开发的夏汉字处理界面中设置了录入窗口，也是采用了四角号码方法编制录入码。用此系统录入西夏文字，可以实现西夏文与汉文等其他文字的混排，但由于该文字处理系统的功能十分有限，无法满足一般要求，因此使用时须先将西夏文录入该系统的窗口，再通过复制方式复制到需要的文字处理系统中，其不成熟性显而易见。笔者也利用四角号码方法对西夏文字进行编码，初次是借用方正典码录入平台实现西夏文字的录入，同时也解决了录入法与各种文字处理系统和排版软件的兼容，但由于方正典码不能与 WINXP 兼容，所以在录入法出版发行后基本上没有应用就已落后于电脑的更新换代。后来，我们又和深圳的士强软件开发公司合作，以他们的万能五笔输入法为平台，在经过改进后，顺利地实现了西夏文字的录入问题。目前，有部分学者开始使用 WINDOWS VISTA 或 WINDOWS7 操作系统，万能五笔系统不能与之兼容，为适应新的操作环境，我们制作了新的录入系统，可以满足学者们的需求。

西夏文字的字形方整，笔画繁冗，李范文先生借用了前人使用于汉字的四角号码检索方法给西夏字编码，这种方法充分体现了西夏字构成特点，是一种简捷高效的好方法，众人使用起来也很方便。这种方法同样可以用于电脑编码，这样做的最大好处是，人们不用花费很大精力就可以掌握。实践证明，采用四角号码法是西夏文字电脑录入的最好编码方法，它既科学又方便，就目前而言，还没有其他可以取而代之者。这也是今后研制西夏文字输入法值得重视和考虑的，除非有更加科学合理的编码外，一般还是要在继承前人成果的基础上结合使用者的具体情况来创新，否则即使自认为科学的也不一定得到认可。

附记：本文是在聂鸿音先生所提供的资料基础上写成。

（作者通讯地址：北方民族大学西夏研究所　银川 750021）