李宇明．语言规划论．长春：东北师范大学出版社，2005：246-257．
原载《中文信息学报》，2003（2）：1-6．

搭建中华字符集大平台*

李 宇 明

内容提要 为使中华文献有一个可进行文字加工的永久性本面目保存本，为满足数字化图书馆、博物馆、档案馆的建设要求，为促进用于知识发掘数据库的建设，为保证中华文化信息在国际互联网上的无障碍交际，必须尽快构建中华字符集。本文主要讨论中华字符集的内容及需要解决的技术问题。

关键词 中华字符集 文献保存 数字化 知识发掘 互联网

Building the Big Platform of China Character Set

LI Yu-ming

Abstract This paper mainly discusses the content of China Character Set and its technique problems．The necessity and urgency of building China Character Set includes：1．In order to get an eternal text which saves the literatures of China in original face and also achieves word-processing； 2．Satisfy the needs of constructing digital library，museum and archives；3. Promote the building of databases which serve knowledge processing；4. Achieve non-obstacle communication of China culture on Internet.

Key words China Character Set；literature saving；digitalizing；knowledge process；Internet

一、中国汉字处理的简单回顾

计算机语言处理的内容非常广泛，汉字处理是其中的重要内容之一。20 世纪 50 年代，中国就有先行者开始研究俄汉机器自动翻译问题，并采用当时的电报码或四角号码充当汉字编码。[①]70 年代，汉字的计算机处理问题开始受到重视，键盘编码、汉字的点阵描述和输出

* 本文曾在第二届肯特岗国际汉语语言学圆桌会议（2002 年 11 月 26～29 日·新加坡）上宣读，并得到厉兵、张书岩、陈双新先生的指教。此次发表又作了一些修改。

① 见傅永和（1999，P51）。

等，成为当时的“必战之役”。[①]80 年代中期，国家站在时代前沿，正式把语言文字的信息处理纳入国家语言文字工作的重要内容。在 1986 年 1 月召开的全国语言文字工作会议上就提出：“汉语汉字的信息处理是一门新兴的边缘学科，有广阔的前景，加强这方面的研究，对经济、文化、科学技术的发展具有长远的意义。因此，当前语言文字工作的任务必须包括这项内容。”[②]

我国计算机汉字处理，近 30 年来获得了重大成就。[③]陆续开发的汉字键盘输入方法，解决了汉字进入计算机的难题；汉字输出实现了多字体[④]、多字号；汉字字库的制作也由点阵字库、矢量字库逐步过渡到曲线轮廓字库；汉字自动识别技术达到国际先进水平，并有商品投入市场；ISO/IEC10646 的CJK字符集，[⑤]由早期的 6763 个通用汉字逐步扩充，扩充集A和扩充集B已经完成，现在正讨论扩充集C_1。待扩充集C_1完成后，中日韩编码汉字已经 7 万有零。[⑥]这些成就，促进了中国计算机的普及，为发展计算机应用技术和信息化创造了基本条件。

但在计算机汉字处理方面，还存在一些需要解决的问题。例如：

1.“万码奔腾”的局面，令业内人士深感担忧。随着计算机语言处理由“字词”向语言理解的深度进展，当前提高键盘输入的质量，已不仅仅取决编码的本身，而主要取决于与语言理解有关的各种资源的建设，如深度开发的各种语料库，各种电子词典，语法典的编纂等。用这些资源支持键盘输入，才能使键盘输入出现跨越式的发展。

2．怎样进一步提高汉字识别质量，特别是脱机手写体识别的突破。

3．语音与文字的自动转换，包括语音输入的文字显示和文字的语音输出，应用前景广阔，但是当前还存在着较多的技术难题，需要重点攻关。汉字与汉语拼音、国际音标等的自动转换、标注，在技术上虽然难度不大，但并没有引起软件制作者的重视。

4．汉字简繁体的自动转换还没有完全实现自动化，特别是由简体向繁体的转换，尚需较多的人工干预。

5．各组异体字之间的关联亟待实现。

6．宋体、仿宋体、楷体、黑体这四种常用印刷字体的字形标准应扩展到更大范围，而且还应建立其他字体的规范或规范原则，应建立掌上电脑等小型计算机上使用的低点阵字形的省减规范。此外，还应研究手写体及书法艺术的计算机表现问题。

除了以上问题之外，汉字处理方面还需要考虑的就是建立中华字符集的问题。本文讨论中华字符集的内容、建立中华字符集的意义、需要解决的技术问题等等，希望引起大家对这一关系到中华文化在信息时代发展的宏大事业的关注和支持。

二、中华字符集及其意义

20 世纪是人类科技大踏步前进的世纪，计算机和国际互联网，应该是 20 世纪人类最重要的发明之一，特别是国际互联网，给人类构造了一个全新的虚拟空间，展示了信息时代的光辉前景。但是，计算机和国际互联网的出现，也在全世界范围内形成了巨大的“数字鸿沟”，带来了各种新的国际性的社会问题。现代国家在制定语言文字规划之时，不能不考虑计算机的语言文字处理问题，不能不考虑网络上的语言文字交际问题，不能不考虑怎么样通过语言

① 参见石云程（1986）。
② 刘导生（1986，P12）。
③ 详见中华人民共和国教育部（2002，P254～259），许嘉璐（1999，P117～130、P202～237），冯志伟（1999，P173～150），傅永和（1999）。
④ 计算机汉字字体，现在已多达 120 余种。
⑤ C：中国；J：日本；K：韩国。
⑥ 现在的提交稿为 70205 个汉字。

文字的信息化来缩小数字鸿沟的问题。

互联网不仅是技术的，而且更是文化的。克服“数字鸿沟”，不仅需要技术，而且也更需要文化。文化是人类文明的积淀，是人类生活的底蕴。文化的传承与开发，是人类生存与发展的重要方面。以计算机和互联网为基本技术支撑的数字化信息时代，文化的传承与开发更加重要。中华文化是由多民族文化构成的源远流长的灿烂文化，在世界文化发展史上具有不可代替的重要性。促进中华文化在信息时代发挥更为重大的作用，是我们的历史责任。

承载中华文化的文字与符号的总和，称为“中华字符集”。中华字符集大致包含如下九个方面的内容：

1．汉字发生隶变之后的记录主流文化的今汉字，包括简体字、繁体字、传承字和异体字等。[①]

2．方俗汉字，包括汉语方言用字、古今民间使用的俗字以及碑别字、错讹字等。

3．古汉字，包括甲骨文、金文、战国文字、简帛玺印字、小篆及汉字隶变之前的其他文字。

4．汉字系的古今少数民族文字，如古壮文、布依文、侗文、毛南文、仫佬文、老苗文、瑶文、白文、哈尼文、西夏文、女真文、契丹大小字等。[②]

5．古今少数民族非汉字系的表意文字，如纳西的东巴文、水文、尔苏沙巴文等。[③]

6．古今少数民族的表音文字，如蒙文、藏文、维吾尔文、哈萨克文、朝鲜文、彝文、傣文（数种）、满文和柏格里苗文、八思巴文、佉卢文、回鹘文、察合台文、突厥文、于阗文、焉耆－龟兹文、粟特文等。

7．记录中国古今语言（普通话，方言，古代汉语，民族语）所需的国际音标；以及为中国语言所设计的各种拼音符号等。[④]

8．文字（主要是汉字系文字）的部首、笔画、偏旁及其他构件。

9．具有文化意义的其他各种符号，如西安半坡出土的陶器上的符号、64 卦图符、道教符号、古代数学史和音乐史等涉及的符号。

建立中华字符集，在信息化时代，对于中华文化的保存和发展具有极为重要的意义。第一，以中华字符集作平台，中华古今文献在电子世界中，就可以有一个可进行文字加工处理的本面目的永久保存本。如果没有中华字符集，中华文献（特别是古文献）在电子空间的保存，只能采取图像方式，不能进行文字加工；或者将文献转写为现在字符集（比如扩充C_1完成后的 7 万余字的范围内）中的字，但是原来的文字面貌失去了，相应地也就失去了许多语言文字信息和文化信息。

第二，为数字化图书馆、博物馆、档案馆的建设提供基础。20 世纪 90 年代以后，美、英、法、德、日和俄罗斯等近 20 个国家和地区相继投入巨资，开展数字图书馆研究。据报道，美国投入到数字图书馆研究的经费已经超过 8 亿美元；俄罗斯在经济尚未全面恢复的情况下，政府计划从 1999～2004 年，每年出资 2 亿卢布支持数字图书馆研究；日本不仅投入了 15 亿日元开发日文文献数据库，而且还以建设国会图书馆分馆--关西电子图书馆为契机，投入 4 亿美元，拟将其建设成为亚洲文献中心。我国随着电子政务、电子商务和远程教育的发展，数字化图书馆、博物馆、档案馆的建设已经，并制定了《中国数字图书馆工程建设一期规划（2000～2005 年）》。[⑤]但是要看到，汉字的特点决定了没有中华字符集，中国的高水平数字化图书馆、博物馆、档案馆的建设，是不可能实现的。

---

① 也应包括 1977 年开始试用、1978 年停止试用、1986 年正式宣布废止的《第二次汉字简化方案（草案）》中的的 248 个“二简字”。

② 女书、傈僳文也可收进到此类字符中。这类字符大约有 25000 个。

③ 这类字符大约有 17000 个。

④ 包括汉语拼音、注音字母以及清末以来主要的拼音文字或拼音方案的符号形体。

⑤ 资料来自d-library．com．cn。

第三，为用于知识发掘的数据库建设提供坚实的平台。智能计算机的发展，依赖于各种可进行知识发掘的数据库，通过数据库中知识的发掘来增强计算机的智能。要使计算机“精通”中华文化，必须建设庞大的可供计算机“学习”的中华文化数据库。科学研究，特别是人文科学研究，数据库逐渐会成为主要的研究手段，通过文献数据库检索文献信息，通过素材数据库统计数据、发掘知识、总结规律、验证结果。中华字符集使得用于知识开掘的中华文化数据库的建设成为可能。

三、中华字符集的实现

中华字符集，是自计算机对文字进行处理以来人类所提出的最为庞大的字符集，在实现中必然会遇到许多技术难题，必须认真研究，妥善解决。

第一，字符种选择。这里所谓的字符种，包括需要进入中华字符集的字种和各种符号。中华字符集的字符种，涉及到古今汉字学、古今民族文字学、语音学、方言学、民俗学、文献学及其他众多学科，必须广泛动员各相关学科的学者，遵照“有见必收”的原则，全方位地拉网式地进行搜集。在字符种搜集过程中，要充分利用已有的各种字典和相关研究成果，利用已有的古文献库和专业语料库、文献库，同时对散布在民间的俗字和一些少数民族的文字，还需要进行田野访查，争取最大限度地搜集字符种。

在广泛搜集的基础上，还需组织专家对这些字符种进行甄别、查重（chóng）、筛选，最后择定。

第二，形体规范。对选定的字符种，要根据不同文字系统和符号的特点进行形体规范，给出典型的或标准的字形或符号形体。被选定的字符种中，有许多没有印刷体形式，如一些民间俗字、碑别字、错讹字；如一些敦煌变文俗字、地下出土的竹帛字，一些古代的少数民族文字等等。这就需要将手写字体“印刷体化”。对于甲骨文、金文、篆文等古汉字，不仅要给出符合这些形体规范的字形，而且可能还需要隶定楷化，并在隶定楷化前后的两种形体之间建立关联。

第三，造字。根据字符种的形体规范，造出适合计算机字库使用的形体，并输入计算机。如果条件允许，还应逐渐造出不同字体的字。鉴于造字量大且技术复杂，应当设计出便捷的计算机造字软件，借助计算机完成造字任务。

第四，存储。中华字符集具有数量大、非匀质的特点。中华字符集的非匀质性表现在：1．它是多种文字系统和符号的集合，在符号性质上具有非匀质性；2．字符的使用范围和使用频率上有巨大差异，有的是现代语言生活所必用，有的只在特殊的领域中使用，在使用上具有非匀质性。

因中华字符集字符种数量巨大，现有的音序、形序等排序方法，都无法适应。以何为顺序将字符种排列存储起来，是需要研究解决的。因中华字符集的非匀质的特点，必须考虑将全字符集分为若干子集，以满足不同领域、不同人群的使用需要。子集的划分可考虑三大原则：

1．民族原则。某一民族的文字或某些形体来源有关系的民族文字尽量结成一个子集。如藏文应当结成一个子集；维吾尔文、哈萨克文、柯尔克孜文因文字来源关系应当结成一个子集；汉字同汉字系的民族文字应当结成一个子集，等等。

2．古今原则。古今文字的用途往往有较大差别，比如隶变之前的古汉字，主要供专业人员使用，一般人很少涉及这些文种；隶变之后的今汉字，有许多只在古籍中使用，现代交际中不使用或很少使用。所以，在依照民族原则划分的子集中，往往还需要以古今再划分子集，当然，像西夏文、女真文、契丹大小字等今天已经不具交际活力的文字，不需以古今原则划分子集。

3．字频原则。字符集应当考虑依照不同的使用频率再划分子集。据华东师大中国文字研究与应用中心研究，西周金文，字频在 10 次以上有 612 字，仅占总字数的 19％，但是这些字的总字频却达 65792，占西周金文字频总量 69289 的 92％。西周金文中，字频在 100 以上的常用字，已识字的比例为 99．2％，已识字比未识字的使用率高。这说明古文字需要考虑字频问题。今天仍然使用的今文字，特别是非拼音文字，有些字符种已经退出现代交际，只具有备用性质，使用频率很低，应建立备用子集。现代交际中还在使用的字符种，如现代语言生活中使用的汉字，还应当再以字频所形成的级次划分子集。

这些具有并列关系、不同层次包容关系和在不同层次上形成交叉关系的各种子集，便形成中华字符集错综复杂的内部结构。根据这种内部结构来处理中华字符集的储存问题，有利于字符集的应用。

第五，检索与输入。运用传统的检索、输入方法难以适应中华字符集。传统的检索、输入方法主要有语音方法和形体方法。运用语音方法进行检索和输入难以奏效，因为：

a）字符集中同音字太多；

b）字符集的语音系统不统一；

c）有许多字符有形无音。

运用形体方法进行检索和输入也困难重重，因为：

a）字符集形体系统差异显著；

b）构件切分不具统一性；

c）部首难以起到统领作用。

检索和输入是使字符集发挥作用的最为重要的两项操作，必须集中精力研究准确、快捷的检索方法和方便的输入方法。

四、中华字符集与国际标准

互联网是人类信息交换的最先进的手段，它所构造的虚拟世界，将成为人类新的十分重要的交际空间。互联网的交际活动，有各种即时的互动交际和网络出版等等。在互联网已经将整个世界勾连起来的今天，建立中华字符集必须考虑互联网的交际问题，甚至说首先应考虑到互联网的交际问题。

如果不考虑互联网交际的话，中华字符集的内码可以有多种，即可以允许每个公司的产品有一套自己的编码系统，或是国家间、地区间使用不同的编码系统。这样的技术路线导致的必然结果是，涉及中华字符集的电子产品，必须携带一个庞大的专用字库；用户使用这样的电子产品时，必须同时装载专用字库，或是通过编程的方式建立某种特定的字库间的映射关系。计算机存储空间的急剧增大和运算速度的快速发展，允许采用这样的技术路线，但是其弊端也十分明显：

1．增加了产品的开发成本；

2．增加了用户下载字库或在多个编码系统之间来回转换的麻烦；

3．网络交换的信息不稳定，使用户常受乱码干扰之苦。

为了克服网络交际中的乱码、空码等问题，保证信息传输的稳定性，实现网络的无障碍交际，为了减去自带字库、多个编码系统来回转换的麻烦，降低开发成本，增加使用的便利，中华字符集的内码必须走国际化的技术路线，实行国际统一编码。

国际统一编码，会使中华字符集具有两大特性：1．中华字符集的每个字符种的代码是唯一的；2．全世界只使用一套统一的编码系统。

ISO/IEC国际标准化组织，为世界各种文字和符号提供了 128×256×65536 个字符的足够大的空间。这些空间分为 128 组（0～127 组），每组由 256 个平面（0～255 平面）构成，

每个平面可以存放 65536 个字符。当前已经定义了第 0 组的 17 个平面（0～16 平面），其中 0～14 平面用来存放字符，15、16 两个平面留作特殊用途。第 0 组的第 0 平面是多文种平面（BMP-basic multilingual plane），当前这一平面已基本用完，[①]字符集再扩充就需要启用其他平面。ISO/IEC/JTC1/SC2/WG2/IRG（表意文字）工作组，负责中国、日本、韩国等国家和地区使用的汉字的国际编码，如前所述，即将完成的编码汉字已逾 7 万。此外，ISO/IEC10646 已经建立了藏文、维吾尔文、朝鲜文、四川彝文等的国际标准，蒙文、傣文的国际标准即将通过。这 7 万多汉字和藏、维、朝、彝等民族文字，已经能够满足日常语言生活交流和一般汉文古籍的出版，功莫大焉。[②]但是，在此基础上要完成中华字符集的国际编码，并使其成为国际标准，还有大量的工作要做。

五、余言

教育部语言文字信息管理司根据许多专家的倡议提出建立中华字符集的设想，该设想受到了语言学界、信息学界的专家学者和社会有关的部门、行业、公司的大力支持。2002 年 3 月 15 日，教育部语言文字信息管理司在北京召开了国际标准“信息技术 通用多八位编码字符集”会议，向一些部门、行业征集“集外字”。7 月 13～16 日在哈尔滨召开了“全汉字库暨国际音标问题商讨会”，研究建立中华字符集的的必要性、紧迫性、可行性及其操作思路。7 月 26 日和 9 月 2 日，在北京两次召集了少数民族语言文字学家座谈会，专门研究少数民族古今文字进入中华字符集的问题。10 月 30～31 日，邀集大陆、香港、澳门、台湾等中国四方专家和官员在福州开会，沟通想法，在进一步讨论扩充集C_1问题之外，初步讨论了中华字符集的子集及以后四方协调的有关问题，达成了共识，并成立了两岸四地中文数字化合作论坛（Chinese Digitization Forum，简称CDF）。中华字符集的建设工作已经起步，正进行字符种收集和子集设计，并已将中华字符集意向及设计方案提交有关的国际会议。

此外，2001 年 5 月，国家语言文字工作委员会再次立项研制《规范汉字表》，[③]可望 2003 年完成初稿并向社会公开征求意见。课题组在研究过程中，还分别在上海、井冈山、合肥、烟台召开了四次专题性的学术讨论会。[④]2002 年 12 月 26 日又在北京听取了有关专家的意见。《规范汉字表》包括现代汉语用字、人名地名用字、科技用字等，可以满足现代语言生活中汉字交际的基本需要，解决现代汉字生活中最基本的问题。根据使用频率和使用领域，字表内部还要分不同的级次，并要进行读音、字形的规范等。

《规范汉字表》也许可以看作中华字符集的一个最为基本的集合。[⑤]一般汉字用户，只需在计算机中使用《规范汉字表》字库即可。特殊用户可以采取在《规范汉字表》基础上再挂上他所需要的中华字符集某个或某些子集，形成“1＋X”的解决方案。

中华字符集规模宏大，难以一蹴而就，许多问题需要分期分批解决，而且需要多方的参与和支持。希望海内外人士关注这一问题，并为中华字符集的实现献策出力。

---

① 大约还有 1000 多个字符空间。

② 参见张轴材（2002）。

③ 该项目已被教育部列为 2003 年工作要点。见《中国教育报》2003 年 1 月 2 日第 2 版。

④ 2001 年 12 月 21～22 日，在上海召开了汉字规范问题学术讨论会。2002 年 5 月 16～17 日，在井冈山召开了异体字问题学术讨论会。2002 年 6 月 22～23 日，在合肥召开了简化字问题学术讨论会。2002 年 8 月 22～23 日，在烟台召开了印刷字体字形问题学术讨论会。

⑤ 中华字符集的基本子集，如果作为CJK的国际标准，还需要在《规范汉字表》的基础上加上日韩的常用汉字。

参考文献

[1]冯志伟．应用语言学综论．广州：广东教育出版社，1999
[2]傅永和．中文信息处理．广州：广东教育出版，1999
[3]胡鞍钢，周绍杰．新的全球贫富差距：日益扩大的数字鸿沟．中国社会科学，2002，（3）
[4]李宇明．信息时代需要更高水平的语言文字规范．术语标准化与信息技术，2001，（3）
[5]李宇明．信息时代的中国语言问题．语言文字应用，2003，（1）
[6]刘导生．新时期的语言文字工作．语文建设，1986，（1～2）
[7]石云程．语言信息处理的新任务．语文建设，1986，（1～2）
[8]许嘉璐．语言文字学及其应用研究．广州：广东教育出版社，1999
[9]姚亚平．中国计算语言学．南昌：江西科学技术出版社，1997
[10]张轴材．国际标准编码字符集与 CJK 的演进．信息时代语言文字规范标准建设工作会议（2002 年 9 月 23～23 日 • 武汉）上的学术报告。
[11]中华人民共和国教育部．跨世纪的中国教育．北京：高等教育出版社，2002
[12]周庆生主编．国外语言政策与语言规划进程．北京：语文出版社，2001