IDS与集外字处理方法研究

肖 禹 | 著

上海远东出版社

IDS与集外字处理方法研究

肖 禹 | 著

上海远东出版社

图书在版编目（CIP）数据

IDS 与集外字处理方法研究 / 肖禹著 . —上海：上海远东出版社，
2016
ISBN 978-7-5476-1245-3

Ⅰ . ① Ⅰ … Ⅱ . ①肖… Ⅲ . ①文字处理系统 Ⅳ . ① TP391.12

中国版本图书馆 CIP 数据核字（2016）第 319752 号

国家科技支撑计划“中国地方志数字化关键技术研究与演示平台设计”项目（2015BAK07B00）
“地方志资源调查与数字化加工规范研究”课题（2015BAK07B01）研究成果

IDS 与集外字处理方法研究
著者　肖　禹
责任编辑 / 徐忠良　装帧设计 / 李　廉

出版：上海世纪出版股份有限公司远东出版社
地址：中国上海市钦州南路 81 号
邮编：200235
网址：www.ydbook.com
发行：新华书店　上海远东出版社
　　上海世纪出版股份有限公司发行中心
制版：南京前锦排版服务有限公司
印刷：昆山亭林印刷有限责任公司
装订：昆山亭林印刷有限责任公司

开本：787 × 1092　1/16　印张：17.25　字数：345 千字
2017 年 3 月第 1 版　2017 年 3 月第 1 次印刷

ISBN 978-7-5476-1245-3/G · 791
定价：68.00 元

出版说明

国有史，地有志，家有谱，家谱、方志、正史从不同层面构成中华民族历史的记忆。

中国自古就有修志的传统。《周礼·春官》载："外史掌四方之志。"东汉郑玄注："方志，四方所识久远之事。"中国地方志作为珍贵的文献资源，其内容不仅包括各地区的疆域、气候、山川、物产等地理资料，也涵盖户口、人物、赋税、艺文等人文历史各方面的记载，是地方的百科全书，一地之全史。地方志所详细记载本地区的政治、经济、社会等发展状况，形成了独特的区域文化，具有鲜明的地方特征；地方志以记述某一段时间当地的情况为主，是一个特定时期文化积淀和历史的产物，反映出了特定时代的经济、政治、文化等方面的烙印；地方志内容广泛，系统性强，从天文地理、名胜古迹、物产资源、民族宗教、方言俗语、金石碑刻到政治经济、科学文化、典章制度、著名人物、重大事件等，分门别类按照内容的要求选择合理的记录方式；资料性是地方志所有特征中最基础的一个特征，是方志生命力之所在。

据不完全统计，汉文古籍超过 20 万种，地方志约占 5%，地方志同时具备的地域性、时代性、系统性、资料性和科学性，既包含丰富的内容信息，又适合与现代技术相结合，建立资源库、知识库和 GIS 系统，进而构建中国传统文化基础平台。以地方志为核心的中国传统文化基础平台将地方志目录、图像、文本、关联数据等不同粒度的数据与地理信息数据相结合，实现时间、空间、文献三个维度的智能检索、数据分析和图形化显示。同时，平台具有高度的容纳性与扩展性，可将各种类型的文献资源、各种格式的数字资源和各种功能的知识工具有机地整合在一起。中国国家图书馆古籍馆陈红彦馆长和肖禹等专家在地方志数字化工作实践中不断积累，研究古籍数字化中遇到的技术问题，进行理性总结。科技部科技支撑计划"中国地方志数字化关键技术研

究与演示平台设计”正是基于地方志这样的特征，希望通过地方志数字化技术、数据抽取技术、可视化技术的统合应用，为古籍数字资源建设利用做出有益的尝试。

实现现代技术与传统文献的紧密结合，打造基础平台，支持数据分析与智能检索，必须以统一的标准规范为先导，因此项目中设计了实现平台相关功能必需的理论研究、加工规范制定等内容，最终以《古籍文本数据格式比较研究》《IDS 与集外字处理方法研究》《国家图书馆藏清康熙时期纂修方志书录》《方志文献特性与数据抽取研究》《地方志数字化加工规范汇编》《地方志数字化加工规范应用指南》六部书的形式呈现。

上海远东出版社

2017 年 2 月

目　录

第一章 绪 论

上世纪 70 年代，计算机开始进入中国。作为强大的信息处理工具，计算机在处理英语等拼音文字上已经取得了巨大的成功，为了使计算机能处理汉字，一大批专家学者投入到相关的研究中，研制了大量的输入法、字符集和相关的软硬件，取得了丰硕的成果，常用汉字的处理早已不再是问题。

一、汉字编码

汉字信息处理（Chinese character information processing），用计算机对汉字表示的信息进行的操作和加工，如汉字的输入、输出、识别等；汉字输入（Chinese character input），利用汉字的形、音或相关信息通过各种方式，把汉字输入到计算机中去的过程；汉字输出（Chinese character ouput），将计算机内以数据形式表示的汉字在显示终端、印字机等设备输出的过程[①]。在汉字处理系统中，如图 1-1 所示，字库处于核心地位，而字符集是构建字库的基础。

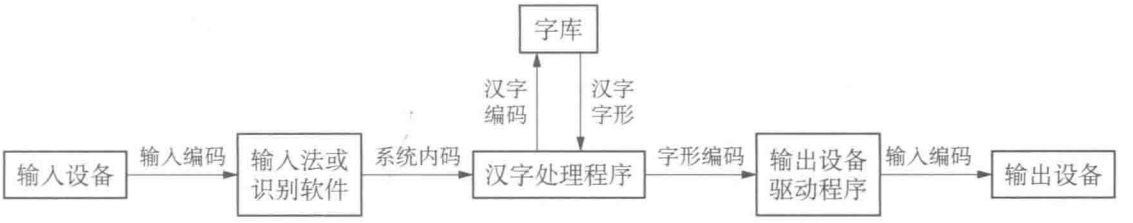

图 1-1 汉字处理系统示意图

① GB/T 12200.1-90，汉语信息处理词汇 01 部分：基本术语 [S]. 北京：中国标准出版社，1990：2.

汉字编码（Chinese character coding），按照一定的规则，对指定的汉字集内的元素编制相应的代码；汉字编码字符集（Chinese character coded character set），按一定的规则确定的包含汉字及有关基本图形字符的有序集合，并规定该集合中的字符与编码表示之间一一对应的关系；汉字编码方案（Chinese character coding schema），汉字集元素映射到其他字符集元素的一组完整规则[①]。字符集是汉字编码的主流解决方案，每一个编码对应一个汉字字形。每个字符集都有固定的汉字编码数量，如GB2312-80字符集，收录6763个简体字。若字库中只包含汉字编码和字形称为静态汉字字库，只能支持静态汉字字形显示；而包含笔画时序信息的，能够进行汉字书写模拟的中文字库称为动态汉字字库[②]。

除了字符集外，还有另一种汉字编码方案。动态组字是一种汉字在计算机等领域的编码理论及技术，是通过一定数量的字根部件（等同英文的字母，但仍为表意）动态生成汉字，并显示到计算机屏幕上，使用者可以根据需要自行组字[③]。基于动态组字的汉字信息处理系统称为无字库系统。

动态组字的研究始于20世纪80年代，略晚于字符集。1984年，周何发表了《中文字根孳乳表稿》，重新分析现代汉字，得出了869个声母及265个形母；张时钊完成了字根组字方案，每个字根固定一种结构特性，基本不需额外的结构码；黄大一在美国完成达意中文符氏语言、达意文书处理系统，以forth描述文字的部件组合进行组字，是最早的组字系统之一。1985年，张时钊完成并推广PC-1500的无字库系统。1991年，谢清俊开始带领庄德明等人用构字式系统分析所有汉字，随后定期发表结果为汉字构形数据库。1998年，CBETA（电子佛典计划）[④]项目成立，由于佛经缺字问题特别严重，叶健欣等人开始研究缺字问题的解决方案。1999年，unicode3.0发布，定义表意文字描述序列以及表意文字描述符[⑤]。2000年，朱邦复工作室发表文昌1610中文CPU，第一个内建汉字字形生成器的CPU，装置于仓颉电书[⑥]。2003年，易符科技发表“易符无限组字编辑器”包括硬件版和软件版；日本京都大学发布影系统（Kage System），可以在服务器上产生组字图片；美国文林科技发布CDL（Character Description Language，组字描述语言）。2004年，张时钊完成并公布微处理器的无字库（即动态组字）演示软件。2005年，张时钊提出元笔画概念，且确实以该理论组出全部ASCII字符及大量图符；王志攀在CBETA项目中使用易符无限组字编辑器，处理了一万六千多计算机缺字，这是动态组字第一次大规模运用。2006年，王志攀在刹那搜

① GB/T 12200.1-90，汉语信息处理词汇01部分：基本术语[S]. 北京：中国标准出版社，1990：4—5.
② 俎小娜. 基于全局仿射变换的分级动态汉字字库[D]. 华南理工大学，2008：5.
③ 动态组字[OL]. [2016-8-16]. http://zh.wikipedia.org/wiki/%E5%8B%95%E6%85%8B%E7%B5%84%E5%AD%97.
④ CBETA简介[OL]. [2016-8-16]. http://www.cbeta.org/intro/origin.htm.
⑤ 动态组字的发展历史[OL]. [2016-8-16]. http://docs.google.com/View?docid=ajh8m4f3vdcc_16n6sh86.
⑥ 苍颉电书平台简介[OL]. [2016-8-16]. http://mail.tku.edu.tw/yjlin/cbf_web/ebook_intro.htm# 蒼頡電書平台簡介.

寻工坊（原易符科技），以动态组字重编《康熙字典》电子版。2007 年，张正一发布开源的动态组字程序，使动态组字程序可以移植普及到各开源平台；刹那搜寻工坊在（台湾地区）“中央研究院”举行可携式造字引擎专利发布会[①]。

经历了几十年的不懈努力，伴随着计算机软硬件和中文信息处理技术的发展，动态组字从理论研究阶段，正逐步走入实用化阶段。纵观动态组字的发展过程，参与研究的人员和机构都很有限，形成的学术成果也不多，学界和社会关注的程度都不高，这种情况直到 2000 年以后才有所改善，其原因有三：其一，现行的汉字编码系统研究起步较早，很快就形成了较为完备的解决方案，动态组字在常用汉字的计算机处理上优势不明显；其二，现行的汉字编码系统投入应用较早，形成了较为完备的软件体系，完全能够解决常用汉字的计算机应用问题，而动态组字在常用汉字的计算机处理上实用性不强，很难推广和普及；其三，1999 年发布的 Unicode3.0 中包含了表意文字描述符，使动态组字的标准化成为可能，而近三十年古籍数字化的发展对现行的汉字编码系统提出了更高的要求，集外字处理问题日益突出，需要引入新的理论和方法[②]。

二、字符集与集外字

20 世纪 80 年代以来，使用汉字或曾经使用汉字的国家和地区相继推出含有汉字的字符集，如表 1-1 所示，这些字符集在一定程度上解决了汉字文化圈国家和地区的汉字计算机处理问题。但是这些早期的字符集收字较少，只适于常用字的处理，而且这些字符集采用不同的编码方案，造成了不同国家和地区间的数据交换、信息处理和显示的问题，而 Unicode 字符集的出现很好地解决了这一问题。

表 1-1 各国各地区字符集收字情况表[③]

国家地区	字符集	收字数	备注
中国	GB2312-80	6763 个简体字	基本集，包含 7445 个字符
中国	GB12345-90		第一辅助集，与基本集对应的繁体字
中国	GB7589-87	7237 个简体字	第二辅助集
中国	GB/T13131-91		第三辅助集，与第二辅助集对应的繁体字
中国	GB7590-87	7039 个简体字	第四辅助集
中国	GB/T13132-91		第五辅助集，与第四辅助集对应的繁体字

① 动态组字 [OL]. [2016-8-16]. http://zh.wikipedia.org/wiki/%E5%8B%95%E6%85%8B%E7%B5%84%E5%AD%97.
② 肖禹，王昭. 动态组字的发展及其在古籍数字化中的应用 [J]. 科技情报开发与经济，2013(5)：118—121.
③ 王荟，肖禹. 汉语文古籍全文文本化研究 [M]. 北京：国家图书馆出版社，2012：26—27.

续表

国家地区	字符集	收字数	备注
中国	GB13000.1-93	20902个字	GB13000.1-93等同采用ISO/IEC 10646.1:1993
中国	GBK	21003个字	包含882个符号
中国	GB18030-2000	27533个字	GB18030-2000兼容Unicode3.0
中国	GB18030-2005	70244个字	GB18030-2005兼容Unicode4.1
台湾地区	CCCII-1	4808个字	常用字集
台湾地区	CCCII-2	17032个字	备用字集，收6025个次常用字、5364个罕用字、2112个异体字以及3531个其他资讯用字
台湾地区	CCCII-3	20583个字	罕用字集，收12924个罕用字、314个次常用字及7345个其他资讯用字
台湾地区	CCCII(异体字集)	11517个字	异体字集，收异体字11517个
台湾地区	Big5	13053个字	包含441个符号
台湾地区	CNS11643-1986	13051个字	去掉了Big5中的2个重复字
台湾地区	CNS11643-1992	48027个字	
台湾地区	CNS11643-2004	54858个字	包括第1至第7字面和第15字面，共54858个字
香港特区	HKSCS-2004	4500个字	HKSCS是香港基于Big5之上扩展的字符集，包含441个符号
日本	JISX0208-1983	6353个日本汉字	包含6877个字符
日本	JISX0212-1990	5801个日本汉字	JISX0208的扩展集，包含6067个字符
日本	JISX0213-2004	11233个字符	JISX0208的扩展集
韩国	KSC5601-1987	4888个韩国汉字	包含8244个字符
韩国	KSC5657-1991	2856个韩国汉字	KSC5601的扩展集

Unicode是一个经过字符宽度整合的编码方式，它是为文字及符号所建立的国际性编码，它几乎覆盖了世界上任何一种语言的字符[①]。Unicode是一种统一的编码标准，为每个字符编码定义了唯一的编码值，能支持上百万个字符编码。Unicode提供了一个标准化的方法，使得在同一系统平台上可以使用多种语言的编码。在Unicode中定义了中日韩统一表意文字（CJK Unified Ideographs）集，收录简体汉字、繁体汉字、方块

① 苗军. Unicode/XML在电子出版物中的实现[D]. 河北工业大学，2002：3.

壮字、日本国字、韩国独有汉字、越南喃字等。目前，Unicode 的最新版本是 9.0.0[①]，日韩统一表意文字集收字 80376 个。

相对于字符集有了集外字的概念，集外字是字符集所不包含的文字，若不采用其他的技术和方法，集外字无法输入、处理和显示。集外字的数量与字符集的收字数量直接相关，若数字化对象的用字总量和文字处理规则固定，字符集收录的文字越多，集外字的数量越少。以国家图书馆数字方志项目[②]第一期（全文化明至民国间的方志 744 种，14682 卷，506485 筒子页，采用键盘手工录入方式进行全文化，使用“中易汉神 e”汉字系统，支持 CJK 基本区、扩 A 区和扩 B 区的 70195 个字符）为例，使用 CJK 基本区 16801 个字（203781248 次），CJK 扩 A 区的 2959 个字（274847 次），CJK 扩 B 区 9117 个字（732675 次）。若使用 GBK 字符集（收录 21003 个字），集外字将多出 12136 个（1007522 次），若使用 GB18030-2000 字符集（收录 27533 个字），集外字将多出 9117 个字（732675 次）。

三、古籍数字化中的集外字处理问题

经过近 30 年的发展，古籍数字化的研究与实践取得了丰硕的成果，产生了一大批有影响的古籍数字化项目。这些古籍数字化项目已经可以实现检索和浏览的功能，但是在文字处理方面还有所欠缺，尤其是在集外字处理方面。

姚俊元在《计算机辅助古籍整理研究的现状与思考》[③]中指出，现有的计算机软硬件状况不能完全适合古籍整理研究的需要，字库字数不够，输入法不适合大字库，64×64 的点阵字形不能满足古籍用字的精度；要确立一个完全适合各种整理研究工作的通用字库是不现实也是不可能的，应考虑设立多套字库；选用一套基本字库，这个字库大约包含 2 至 3 万个古籍常用汉字，对字形适当进行规范处理，常用的异体字尽可能收录，罕见的异体不予考虑；根据研究工作的对象来确定专用字库，可以依据朝代设计，如唐代大汉字库、清代大汉字库，也可以依据字体设计，如甲骨文字库、篆文字库等；基本字库确立以后，中文平台的设计还要具备扩充汉字功能，因为字库再全，总难免缺少部分冷僻字（特别是地名、人名）。

陈洪澜在《中国古籍电子化发展趋势及其问题》[④]中指出，古籍用字繁难，电脑字库需要扩展；GB2312 字符集（6763 个汉字），只适应于一般的文字处理工作，要对古

---

① Unicode® 9.0.0[OL]. [2016-8-16]. http://www.unicode.org/versions/Unicode9.0.0/.

② 国家图书馆数字方志项目始于 2002 年，先从馆藏旧方志（1949 年以前编撰或出版）中选出 6800 余种进行彩色扫描，采集图像 330 余万筒子页，编制卷目索引数据 50 余万条，之后分批进行全文化，截至 2015 年底，已完成 3100 余种 200 余万筒子页。

③ 姚俊元. 计算机辅助古籍整理研究的现状与思考[J]. 图书情报论坛，1995(3)：68—71.

④ 陈洪澜. 中国古籍电子化发展趋势及其问题[J]. 中国典籍与文化，1998（4）：121—126.

代文献中的汉字进行处理就远远不够了；只有尽快颁行能够适用于处理中国古籍又符合国家标准的大型字库，使古籍的处理工作有标准化、规范化要求，才能创造良好的古籍电子化利用环境。

宫爱东在《新世纪图书馆古籍数字化的几个问题》[①]中指出，汉字库的问题是实现古籍数字化最核心的问题，也是目前古籍要采用字符方式实现数字化的最大技术困难；据统计，古籍内通用字约4万个，常用异体字约两万个，生僻少见或自创的怪字，最多约两万个；因此，字库内有6万汉字应能满足基本需要，有8万字，一般说应该满足需要了，最多到10万字就能完全满足需要。

陈立新在《古籍数字化的进展与问题》[②]中指出，对古籍进行数据处理，首先遇到的就是汉字库及中文平台的问题；据查，《康熙字典》收字49030个，《汉语大字典》收字约为56000个，其中的异体字、避讳字、冷僻字给文字处理带来了很大的难度，而且现有的一些较为标准的中文汉字平台都缺乏完备的甲骨、金文、篆、隶等字库，也没有少数民族的文字字符；因此，应尽快开发新的汉字大字符集，建立支持古籍数据化的汉字平台，建立汉字属性字典，建立词料库等。

陈力在《中文古籍数字化的再思考》[③]中指出，汉字处理是古籍数字化工作最早遇到的问题，以前学术界关注的焦点是用繁体字客观再现古籍内容；目前业界大多采用Unicode作为文字处理的标准，Unicode已经定义了7万多汉字，不久将再扩充2万汉字；因此，古籍文本的简单转换已不是什么太大的问题了；目前最大的问题是如何处理古籍在传抄、刊刻过程中所产生的一些问题，如异形字、避讳字、通假字等等；当然，目前业界普遍采用的Unicode本身也存在许多问题。

尉迟治平在《电子古籍的异体字处理研究——以电子〈广韵〉为例》[④]中指出，纸本古籍由书家抄写，刻工雕版，不仅是异体，凡认为是字的，不管实际上有没有这个字，也不论写得对还是错，都可写成印出，甚至率意更作，增加新的异体；而电子古籍只能显示电脑字库中有的汉字，字库收纳的字种由字符集（国际或国家标准）规定，字形由字库生产厂家制作；这样就形成了两种异体字系统，我们将前者称作“刻写异体”，后者称作“数码异体”；在历史上刻写异体是一个开放系统，而且迄今仍没有进行过穷尽性的调查和完全的认定，即使是像《广韵》这样的常用典籍，我们对书中异体字的情况也缺乏清晰的认识；现时的数码异体是一个封闭的系统，虽然中日韩统一汉字（CJK）已达70195个字符，但是即使将来再加扩展，也只可能是刻写异体的一个子集。

---

① 宫爱东. 新世纪图书馆古籍数字化的几个问题 [J]. 图书馆学刊，2000（1）：18—20.
② 陈立新. 古籍数字化的进展与问题 [J]. 上海高校图书情报工作研究，2003（2）：36—38.
③ 陈力. 中文古籍数字化的再思考 [J]. 国家图书馆学刊，2006（2）：42—49.
④ 尉迟治平. 电子古籍的异体字处理研究——以电子《广韵》为例 [J]. 语言研究，2007（3）：118—122.