· 新论摘编 ·

中文古籍数字化建设

龚姪君, 刘春金

(浙江中医药大学图书馆, 浙江杭州 310053)

按照《古籍著录规则》中的定义,古籍主要是指1912年以前在中国书写或印刷的、具有中国古典装订形式的书籍。中文古籍是前人留下的宝贵文化遗产,对后人进行科学研究和文化传播都具有非常重要的意义。但随着时间的推移,这些宝贵的文化遗产变得相当脆弱,从而限制了大部分用户的使用。古籍数字化就是利用现代信息技术将古籍转化为电子媒体的形式,通过光盘、网络等介质予以保存和传播^[1]。古籍数字化是古籍整理工作提出的新要求,也是古籍整理发展的方向。本文试图就中文古籍数字化建设现状、古籍数字化建设样式、古籍数字化建设技术、古籍数字化建设存在的问题等内容进行探讨。

一、中文古籍数字化建设现状

(一) 台湾地区中文古籍数字化建设先声夺人

我国台湾地区的中文古籍数字化工作一直处于全球领先地位。从1984年开始,台湾“中央研究院”就着手进行古籍数字化工作^[2]。目前,已开发的古籍检索系统有汉籍电子文献瀚典全文检索系统(<http://www.sinica.edu.tw/~tdbproj/handy1/>)、文物图像研究室资料库检索系统(<http://saturn.ihp.sinica.edu.tw/~wenwu/search.htm>)、历史语言研究所藏内閣大库档案(<http://saturn.ihp.sinica.edu.tw/~mct/newpage1.htm>)、台湾大学中华电子佛典线上藏经阁大正藏全文检索系统(<http://www.cbeta.org/index.htm>)、台湾元智大学“网路展书读”(<http://www.cbeta.org/index.htm>)等。

汉籍电子文献主要收录了二十五史、阮刻十三经疏、《清实录》、上古汉语语料库、台湾文献丛刊、《文心雕龙》、《中华民国史事日志》等文献,基本涵盖了唐代以前绝大部分传世文献,并且大部分资料免费开放,可提供网络检索和按目录路径全文浏览古籍文献。这些古籍资料都经过加工,加上了句读和注释,方便读者理解原文内容。最新推出的3.1测试版收集的汉籍字数已突破一亿,系统中加入了地名、朝代等专词检索功能。

(二) 香港地区古籍数字化建设成果丰硕

香港的古籍数字化工作是从20世纪80年代末开始,从事这一工作的主要是香港中文大学中国文化研究所下属的“汉达古文数据库中心”(<http://www.chant.org/>),所开发的全文数据库资料包括甲骨文数据库、竹简帛书出土文献电脑数据库、金文全文计算机化数据库、先秦两汉一切传世文献计算机化数据库、魏晋南北朝传世文献数

[收稿日期] 2006-04-16

[本刊网址·在线杂志] <http://www.journals.zju.edu.cn/soc>

[基金项目] 浙江省图书馆学会2005年度立项课题(ZTX2005B-6)

[作者简介] 1. 龚姪君(1971-),女,浙江杭州人,浙江省中医药大学图书馆馆员,主要研究方向为图书情报管理; 2. 刘春金(1972-),男,浙江杭州人,浙江中医药大学图书馆馆员,主要研究方向为知识信息管理。

据库、中国传统类书数据库。目前,该数据库注册后可提供30天试用期,试用的文献量大概占总量的10%,如果要充分利用数据库,还必须下载ICS1—ICS7等多种字库。

先秦两汉一切传世文献计算机化数据库是1988年在“香港资助研究局”资助下建设的,该数据库以一固定年代之全部传世文献为对象,不论经、史、子、集和字数多寡,均悉数入库;入库文献所据版本均为旧刻善本而未经后人擅自改动者,多为《四部丛刊》本。该数据库还编制了作者索引、书目索引和年代索引,并且可以通过输入串进行检索;另外,阅读资料时,在有注释的地方会自动弹出注释,以便读者理解原文。

甲骨文数据库的建设历时五年,收录了当今海内外七种主要大型甲骨书籍,共计卜辞53834片,合并整理、校勘近一百万字甲骨卜辞,并具有多项类目功能检索。它的建成不仅方便了有志研究甲骨文学者的需要,也符合了古文字学专业学者的学术要求,具有重大的意义。该系统可以同时显示甲骨文字原字形及隶定释文,方便读者参照;另外,收录了所有甲骨文字的字形总表和提供便捷检索功能和甲骨文字出现字频数据,以便专业学者研究分析。

(三) 大陆古籍数字化建设奋起直追

从20世纪90年代开始,大陆一些图书馆进行了古籍数字化建设的尝试。1996年,上海图书馆与长江计算机集团合作,建立了“古籍影像光盘制作及检索系统”,该系统将古籍善本以图像形式扫描,全文录入,并研制了古籍善本的检索查询系统,到1999年4月,该系统已经完成古籍善本1000多种60万页的扫描和标引工作,使大陆地区的古籍数字化从计算机书目服务走向全文服务^[3]。

中国国家图书馆在2000年正式成立中国数字图书馆有限责任公司来推动图书馆的数字化建设工作,其中古籍图书馆数字化建设也是国家图书馆的一项重点工程。近年来,该馆在建设特色珍稀文献数据库方面成果卓著,建设过程中采用了人工智能检索、信息海量存储、自动标引、电子上网等现代信息技术^[4]。

除了图像形式的古籍数据库外,有些图书馆也进行了古籍书目数据库和索引数据库建设的尝试。如南京图书馆建立了四十多万条中文古籍书目数据库等^[5]。此外,中国社科院研制的《全唐诗》、《先秦魏晋南北朝诗》、《全唐文》、《诸子集成》等数据库检索系统,深圳大学的《红楼梦》数据库,北京大学的《全宋诗》等古诗研究系统,哈尔滨师范大学的《史记》全文检索系统,河南大学的“宋人笔记”检索系统等都为大陆地区的古籍数字化做出了一定贡献。

(四) 世界其他地方中文古籍数字化建设情况

中文古籍的数字化建设不仅在国内开展,在国外也有一定的研究。如美国的普林斯顿大学建设了中文古籍善本书目数据库,挪威的奥斯陆大学建设了先秦诸子百家全文检索系统,这些都丰富了中文古籍数字化建设的内涵。

二、中文古籍数字化建设的形式

中文古籍数字化建设的形式主要有古籍书目型、全文型、全图像型和图文型。古籍书目型是最早建立起来的有关古籍资料的数据库,只能说是古籍数字化的雏形,还不能算是真正意义上的数字化。它是将馆藏古籍的书名、著者、分类、索引等信息输入计算机从而形成古籍书目数据库,读者可以通过书名、著者等检索到某古籍的相关资料,该类型的代表是南京图书馆建立的含四十万条数据记录的中文古籍书目数据库。另外,北京大学和清华大学等高校图书馆也编有各自的馆藏古籍书目数据库,在一定程度上方便了读者的使用。全文型是将古籍资源全文通过手工方式录入,形成电子文本,供用户查阅。这种数据库的优点是便于阅读和检索,存储空间小;缺陷是未保持古籍原貌,文字录入难度比较大。其中台湾的汉籍全文资料库就是这种类型数据库中的佼佼者。全图像型是将古籍直接以图像格式扫描加上简单的标题和分类再存储到计算机中,这是国内采用最多的一种方式。该方式的优点是能保存古籍原貌,有助于专业学者的研究,缺点是检索不方便。例如武汉大学出版社以文渊阁本《四库全书》为底本,将全书二百余万页逐页扫描,并将“总目”手工录入,开发出“四库全书光盘版”。图文型是在古籍书页图像存储基础上,将书中具有检索意义的内容数字化转为电脑可识别的文字,并配以适当的软件工具,为读者提供快捷的

检索、统计、整理和编辑功能。该方式的优点是检索快捷,原貌再现,查证方便。这种类型结合了上两种形式的优点而克服了其不足之处,是目前古籍数字化建设的最佳形式,是现阶段建立古籍数据库的首选方式。香港的汉达古籍数据库采用的就是这种形式,它可以在视窗系统上直接显示简帛图片和对照文本。

三、中文古籍数字化建设技术

由于中文古籍具有抄本和刻本之分,体例繁多,抄本的字迹不规范,用字和版面复杂,有的甚至含有批注、线上字等,同时因年代久远,部分页面有污损、字迹模糊和重影字等情况,从而与普通图书数字化建设相比,中文古籍数字化建设工程更为复杂,难度更大。

在进行古籍全文数字化建设时面临最多的问题还是有关用字的问题,据查,仅《康熙字典》收字就达 49 030 个,其他古籍中含有的生僻字、异体字等就数不胜数了,这些字在现有电脑系统的字库中不存在。无论是大陆的 GB2312-80(《信息交换用汉字编码字符集·基本集》)还是中、日、韩三国的 ISO10646 大字符集,都远远不能达到古籍的用字量。为了能够全部浏览古籍的内容,很多古籍数据库都要求下载相应的字库,例如使用香港汉达文库时,根据数据库的类型要分别下载不同的字库到本机电脑,其中甲骨文数据库须下载 ICS3 和 ICS4 字库等。另外,北京中易信息技术公司以 7 万汉字环境、ISO10646 国际标准内码、4 字节检索为核心的技术,构造了中文网络应用平台——中易汉神 E,较好地解决了处理大汉字字符集的需要,也可以和国际互联网连接,实现全球浏览和检索^[5]。

在对古籍全文数字化时,有时也用到 OCR(光学字符识别)技术,其中 OCR 技术运用比较成熟的是“数码翰林”软件^[7],用该软件制作的“四库全书”及“四部丛刊”电子图书,文本页面保持了原书的竖排格式,增强了古籍阅读的真实感,基本实现了图文关联和阅读检索及全文检索的功能。此外,在对古籍数字化处理时,还用到一些其他图像处理软件,如 Photoshop、ACDSee 等。

四、展 望

古籍全文数字化建设无疑是一项艰巨而复杂的系统工程,古籍数字化基本上还是以专用型为主,在实用性、通用性和共享性方面还存在不足之处,建设中还有很多问题亟待信息工作者研究和探讨。信息技术的飞速发展,也给古籍数据库建设和利用展现了广阔的前景。

[参 考 文 献]

- [1] 陈阳. 古籍数字化发展概述[J]. 电子出版, 2003, (8): 2-4.
- [2] 杨虎. 港台地区古籍数字化资源述略[J]. 电子出版, 2003, (8): 8-11.
- [3] 托雅. 浅谈我国古籍数字化发展进程及发展方向[J]. 内蒙古科技与经济, 2004, (16): 139-140.
- [4] 段泽勇, 李弘毅. 古籍数字化的回顾与展望[J]. 图书馆理论与实践, 2004, 9(2): 37-39.
- [5] 余述淳. 古籍资源数字化建设的一些探讨[J]. 大学图书馆情报学刊, 2003, 21(2): 69-71.
- [6] 北京中易郑码新技术有限公司. 历史文献全文数字化——难点与解决方案[J]. 数字与缩微影像, 2005, (1): 18-21.
- [7] 包钟. 让古籍数字化成为可能——古籍数字化处理技术[J]. 数字与缩微影像, 2003, (4): 7-10.