

中医古籍文献数字化建设的实践与思考

吴桂英

(南京中医药大学图书馆 南京 210046)

〔摘要〕 结合南京中医药大学图书馆中医古籍数字化的实践,介绍中医古籍数字化过程,探讨书目著录处理、检索功能等方面存在的障碍,对中医古籍数据库建设提出建议,包括加强检索功能、制定统一标准、建立专业队伍几方面。

〔关键词〕 中医古籍文献;数字化建设;古籍数据库

Practice and Thoughts on the Digitalization Construction of Ancient Chinese Medicine Literatures WU Gui - ying, Library of Nanjing University of Traditional Chinese Medicine, Nanjing 210046, China

〔Abstract〕 Combining with the digitalization practice of ancient Chinese medicine literatures in Nanjing University of Chinese Medicine, the paper introduces the digitalization process and discusses the existing obstacles in the aspects of processing of bibliography of cataloging, retrieval functions, etc. It also puts forward the advices on database construction, including strengthening retrieval functions, formulating uniform standards and organizing a professional team.

〔Keywords〕 Ancient Chinese medicine literatures ; Digitalized construction; Ancient literature database

中医学历史悠久,具有独特的医学理论体系和丰富的实践经验,海量的中医古籍是传承发展中医药学术的主要载体。随着计算机、网络化时代的到来,传统的手工整理方式已远远不能满足快速、准确、全面获取古籍信息资源的需求。中医古籍数字化就是采用计算机技术,对中医古籍文献进行加工、处理,将古籍的有关信息转换成数字信息,存贮在计算机可读的各种介质上,制成中医古籍文献书目数据库、古籍全文数据库和古籍知识库,用以揭示中医古籍文献中蕴含的丰富的信息知识资源,从而达到保护、利用和挖掘古籍文献的目的^[1]。

1 中医古籍数字化的过程

〔修回日期〕 2009 - 12 - 30

〔作者简介〕 吴桂英,馆员,发表 6 篇论文。

1.1 中医古籍书目的选取

南京中医药大学拥有大量中医古籍图书,是 2008 年国务院公布的第一批国家珍贵古籍名录所在单位和第一批全国古籍重点保护单位,馆藏中医古籍版本鉴定准确、体例完备、著录规范。由于人力物力财力等原因,许多明清时代的中医药文献被束之高阁,造成资源极大的浪费。应充分利用计算机技术、数码技术和通讯网络等技术将这些宝贵的文献进行整理加工和发布,最大限度地揭示挖掘中医古籍的作用。根据本馆收藏的中医古籍的年代、版本价值、文献价值、保护等级等确定收录标准,对破损严重的珍善本古籍优先进行数字化处理,目前入库书籍(普本)已有 200 余册。

1.2 元数据加工

根据古籍文献的特点,中医古籍按分类、题

名、责任者、出版年代、出版者、版本项、版式、卷数、馆藏地、书籍特征、保存状况、复制状况等特征进行著录,著录规则参见 CALIS 联机合作编目手册和古籍著录规则,采用《四库法》对本馆古籍进行分类,本馆已经完成全馆 3 481 种 (27 692 册) 古籍图书的分类编目工作。

1.3 数字化处理

考虑到对古籍的再生性保护,以古籍原文图像作为主要数据来源,部分古籍实现全文本化。包括古籍的扫描、质检、图像处理、文本录入等。为实现保存和利用双重目的,扫描时采用全彩模式,300 dpi, JPEG 格式保存,按页码顺序重命名,对扫描后图像的顺序、偏斜度、完整性、清晰度进行检查,并进行纠偏、去污、裁边等处理。采用图文关联方式,将文本对应放在与之相匹配的图像之下,将书中具有检索意义的内容数字化,并辅以适当的软件工具,为读者提供检索、统计、整理和编辑功能^[2],并建立索引关系,实现图文对照。

1.4 古籍原文图像标引

采用自主开发的古籍编辑器对文中有效的检索点进行抽取和 XML 格式标引,如病症方药的名称、索引信息、各级标题等,力求实现图像的全文检索,避免标引过繁或过简,影响检索效果。对标引后数据进行审核。

1.5 数据发布

将图像数据、元数据、标引数据、文本数据合成、入库,形成中医古籍数据库,发布到图书馆网站后进行数据库测试,实现中医古籍的数字化保护与利用。

2 中医古籍数字化存在的障碍

2.1 书目著录的处理

对图书馆馆藏 3 481 种古籍图书进行书目著录,这项工作与数字化工作同步进行。古籍图书印刷格式多样、出版方法随意性大,多处标有正题名且各

处题名不尽相同^[3];另外“提要或文摘附注项”的著录是由编目人员根据一书的内容,参考有关资料编写成的该书的思想内容方面的评价,以及著者生平和本源流等有关事项的简要介绍。中医古籍 MARC 著录除注意以上一些问题外,编目人员的责任心也影响编目数据的质量。编目人员不仅要具备一定的图书馆学、版本学、目录学、文献学和古汉语知识,还要具备相关的计算机知识,并有较熟练的计算机操作能力。更重要的是编目人员要有较强的事业心和责任心,认真仔细,这样才能提供高质量的 MARC 数据。

2.2 中医古籍中生僻汉字的处理

中医古籍数字化过程中遇到的字体障碍有两个方面,第一,是对字体的认知障碍,生僻汉字的准确录入需要很深的中医文献学功底;第二,是技术障碍,Unicode 字符集虽然可以解决字库问题,但是很多软件不支持,很多生僻字仍不能正确显示和检索。用造字的方法费时费力,而且很难共享和检索,生僻汉字的处理仍是中医古籍数字化过程中的障碍^[4]。

2.3 古籍数据库检索功能的限制

本馆自建的中医古籍数据库图文关联,保留了中医药古籍原貌,但是由于人力不足,现在主要侧重于保持古籍的原貌,检索功能不强。快速、准确地检索到所需要的内容是古籍数字化应有的基本功能,也是评价其优劣的重要指标^[5]。全文本的数字化古籍可以实现任意字符的检索,但是数据准确度小、检索噪音大;图像版可以保证信息的准确无误,但在全文检索上存在一定缺陷,通过对有效检索点的标引可以在一定程度上弥补上述不足,但仍不可避免地会遗漏某些信息。由于中医古籍中病、证、方、药的同名异义、异名同义现象普遍存在,使内容的精确检索更加困难。如何建立后控词表,实现关联检索也是中医古籍数字化面临的一个难题。

2.4 版权保护

为保护电子资源的知识产权,维护本校图书馆

的声誉,也为保障广大合法用户的正当权益,将中医古籍数据库放置位于图书馆五层的中医古籍书库。现在读者对数据库只能浏览目录,全文浏览暂时还未对外公开使用,如果需要看全文,需到中医古籍库的专用机器上浏览,并设置登录密码,在浏览时读者要遵循有关的版权规定。

3 中医古籍数据库建设的建议

3.1 加强数据库的检索功能

古籍汉字识别需要高水平的专家对文字把关。但有些操作人员水平有限,不能识别古籍中的异体(形)字、俗字,影响了古籍数字化的质量,降低了数据库的检索功能。图文关联是实现中医古籍数字化的较好方式,将文本对应放在与之匹配的图像下,并建立索引关系,实现图文对照。既可弥补“图像”在检索和内容编辑上的不足,又可校对“文本”的准确性,使读者既能看到原书原貌的古籍文献,又可以快速准确地查到所需内容。古籍研究者对原始版本信息和文本的准确性要求更高,通过图文对照可以很好地解决这一问题。

3.2 重视统一标准的研究和制定

古籍整理工作相对独立,采、编、阅、藏自成体系,古籍整理保留旧有模式,采用旧的分类法,沿用传统的著录方式,馆际之间也没有国家统一标准^[6]。古籍的数字化文件的格式不统一,不利于用户使用^[7]。建立完整的中医古籍资源库需要全行业的合作。因此,有必要建立统一的标准,包括中医古籍的分类法、著录规则、软件和数据库的使用、数字化加工标准、加工和利用平台等。统一的标准是实现中医古籍数字化资源共享的基础,也是促进中医古籍数字化建设进程的有力保障^[8]。

3.3 专业队伍的建立

中医古籍数字化是古典文化与现代技术的结合,需要有精通中医古籍整理、目录学和版本学知

识的人才与掌握计算机网络知识和多媒体技术的专业人员通力协作^[9]。中医古籍数字化虽属新的古籍整理方式,但仍归于中医文献整理范畴,离不开文献整理的知识。在整个过程中,中医知识不可或缺;同时,在数据处理、数据库设计、应用平台构建等方面都需要计算机专业人员参与。中医文献专家的需求和建议能指导计算机技术人员对整个系统的设计,进而改善中医古籍数字化系统的功能。

4 结语

中医古籍数字化是对文化遗产的保护和弘扬,是一项系统工程,需要中医药科研人员付出长期的艰辛和努力。本馆中医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,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,给中医古籍数据库建设和利用提供了广阔的前景。

参考文献

- 1 杨继红. 中医古籍数字化资源建设概述 [J]. 现代情报, 2008, (5): 136-138.
- 2 符永驰, 刘正国, 李斌, 等. 中医古籍数字化研究 [J]. 中国中医药信息杂志, 2004, (6): 563-564.
- 3 李璐. 古籍全文数据库建设的技术与实践 [J]. 图书馆学研究, 2004, (11): 22-25.
- 4 陈进, 王杰贞, 胡成湘, 等. 我馆中医古籍文献数字化制作 [J]. 医学信息学杂志, 2006, (3): 200-201.
- 5 李兵. 常用中医古籍数据库评价与分析 [J]. 江西中医学院学报, 2009, (4): 86-88.
- 6 李兵, 刘国正, 符永驰, 等. 从中医古籍数据库建设看中医古籍数字化 [J]. 中国中医药信息杂志, 2009, (3): 92-93.
- 7 龚娅君. 中文古籍文献数字化建设的实践与思考 [J]. 中华医学图书情报杂志, 2009, (1): 19-20.
- 8 龚娅君. 古籍文献数字化建设的几点思考 [J]. 现代情报, 2008, (12): 95-96.
- 9 吉聪. 中医古籍数字化建设问题探讨 [J]. 长春中医药大学学报, 2004, (3): 64-65.