

· 思路与方法 ·

中医古籍数字化探讨

符永驰¹ 李 兵¹ 王建文¹ 李 斌¹ 孙海舒¹ 王 蕊¹ 刘国正²

(1. 中国中医科学院中医药信息研究所, 北京市东直门内南小街 16号, 100700; 2. 中医杂志社)

[摘 要] 回顾了古籍数字化的发展历史和研究背景, 提出中医古籍数字化具有保护、利用、整理 3个目的。基于不同目的, 从书目选取、内容加工、利用平台、研究支持及整理研究 5个方面探讨了中医古籍数字化的方法, 并提出了展望。

[关键词] 中医古籍; 古籍数字化; 古籍整理

古籍数字化最早开始于 20世纪 70- 80年代。近 30年来, 人们对古籍数字化的探索始终在不断向前推进。结合专家学者的不同认识, 可以将古籍数字化做如下定义: 从保护和利用古籍的目的出发, 应用现代信息技术将古籍中的文字或图像信息转化为能被计算机识别的数字符号, 并使之结构化, 从而实现古籍整理、存储、检索、阅读、传输的电子化, 是古籍整理和开发的一项系统工作^[1]。其中中医古籍数字化是整个古籍数字化工作中的重要部分。

1 中医古籍数字化的研究背景

回顾古籍数字化的发展, 可以将其发展过程大概分为 3个阶段。第一阶段是在 20世纪 90年代中期以前, 可以说是古籍数字化的探索、起步阶段, 这一时期主要是探讨新兴的计算机技术与古籍整理和应用的结合, 包括古籍机读目录和一些文史古籍全文索引的编制, 由于计算机运算能力的限制, 这一时期还未能涉及全文加工和数据库建设。

第二阶段是在 20世纪 90年代中期到 2002年以前, 计算机和网络技术的发展使这一时期的古籍数字化得到较快发展, 数字化的方式涉及图像化、文本化及图文结合, 应用形式也扩展到全文数据库、光盘版以及网络版古籍。古籍数字化的研究也涉及了全文检索、汉字字符集、数字化输入技术以及数据库的建设等, 可以说这一阶段古籍数字化的技术与实践都有了全面的进步, 是古籍数字化的提升、发展阶段。

第三阶段是在 2002年以后, 这一时期古籍数字化主要体现在理论上的逐渐成熟和大规模数字化项目的开展。计算机运算能力的极大提高和网络的普及使大

规模古籍数字化建设成为可能, 古籍数据库在容量、数量、功能等方面都有了极大的提高。古籍数字化理论问题的研究要晚于技术问题的研究, 这一时期关于古籍数字化的探讨更多集中在理论方面, 并逐渐转向知识库、知识发现等古籍深度利用研究方面, 古籍数字化的概念和相关理论也逐渐明晰, 是古籍数字化的进一步发展和成熟阶段。

中医古籍数字化的研究是随着整个古籍数字化的发展而逐步开展的, 其开始要晚于其他类古籍。中医古籍在各类古籍中占有很大比例, 并且中医古籍较之其他类古籍更具有实用价值。因而近年来, 中医古籍数字化的研究发展迅速, 并逐渐在各类古籍数字化研究中脱颖而出。

2 中医古籍数字化的目的

任何项目和工作的开展都应具有明确的目的, 目的不同, 则采取的方法和工作方案就会不同。中医古籍数字化作为专题、特色的古籍数字化研究, 主要有以下 3个方面目的: 一是中医古籍的保护。就是用“原图原貌”的图像实现中医古籍的永久保存和再生性保护, 并且通过数字化中医古籍的利用减少对原书的磨损而实现中医古籍的原生性保护。二是中医古籍的利用。如中医古籍阅览的电子化、网络化, 方便快捷的检索, 以及知识发现、知识挖掘等。相对于其他学科的古籍, 中医古籍在其学术价值之外, 更具有较高的实用价值。因此, 中医古籍数字化建设应该重视其实用性。三是中医古籍的整理。古籍数字化不仅仅是技术领域的问题, 也是对古籍内容的重新整理, 包括载体的转换、文字的重新考订、内容的标引等, 因此, 很多学者认为, 古籍数字化属于古籍整理的研究范畴^[2]。中医古籍数字化也同样

基金项目: 科技部科研院所社会公益研究专项 (2004D1B5J188)

通讯作者: 刘国正, liugz1919@126.com, (010) 64014411 转 3034

(C)1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3 中医古籍数字化的方法

中医古籍数字化是一项系统性工作,涉及到目的、内容、技术、标准等众多问题,其中内容是核心,技术是关键,所以在方法选择上要考虑到以上方面。在明确目的后就应选取合适的系统方法,包括书目的选取、内容的加工方式、数字古籍的编辑、数字化形式和利用方式、阅读和检索功能、应用和服务功能等方面。我们以中国中医科学院中医古籍资源库的建设为例,简述中医古籍数字化的方法和过程。

3.1 书目的选取

现存历代中医古籍数量众多,版本复杂,中医古籍数字化建设要在一定人力、财力条件下根据目的选取合适的书目进行。中医古籍书目选取应着重其学术、文献价值,兼顾中医古籍保护与利用的双重目的,并需要有中医文献专家把关,可以从品种和版本两方面体现。

品种的选择应更注重其内容和文献价值,尽量选取在中医学发展过程中具有重要影响力的经典名著、代表著作,或属于某一疾病、某一问题的独有文献,或载有独特的诊治疾病和养生保健经验,具有广泛实用价值的古籍。版本的选择更应注重其保存和学术价值,尽量选取珍本、善本、完本、精校本古籍,孤本或刻印较少的罕见版本以及损坏严重的古籍也应优先选取,以便对其进行数字化保护。

3.2 内容的加工

古籍内容的数字化转换有图像、文本及图文结合 3 种方式,图像形式便于古籍的保护和阅读,文本形式便于整理和检索,图文结合可以弥补“图像”在检索和内容编辑上的不足,又可以校对“文本”的准确性,使读者既能看到原书原貌的古籍文献信息,又可以快速准确地查询到所需的内容,是中医古籍数字化的较好方式^[3]。

图像数据的加工应以尽可能保存古籍的完整信息为原则。图像的采集应优先选取扫描的方式。图像的存储可以分为“典藏版”和“阅读版”,前者分辨率宜高,作为典藏保存,后者可降低分辨率,便于利用和节省空间,用于机读;古籍文本化的加工,有手工录入和采用专门的 OCR 软件识别,辅以人工校对的方式。据统计,古籍的 OCR 识别准确率在 93.2% 左右,其余部分则需人工校对,选用合适的校对方法可以提高校对的质量。

3.3 数字中医古籍的编辑和数据库构建

经过加工处理的古籍图像和文本化的数据需专门的编辑软件处理才能入库。古籍图像的编辑处理是实现古籍内容检索的重要一环,我们在中国中医科学院

中医古籍资源库的建设中,为此专门开发了古籍编辑器来处理这些数据。古籍编辑器的功能包括:① 编辑古籍的基本信息,包括书名、作者、馆藏地、书籍版本、保管措施、书籍特征、复制状况、内容提要等;② 对古籍图片元数据进行标注,古籍图片中的主要知识点就是图片的置标信息,如图片中反映的病名、证名、方、药等。为此我们制定了标注的基本原则及标准。经过标注加工处理,实现了对图片内容的检索。这一方法较全文检索大大地提高了检索的精度。

网络和数据库为数字化中医古籍的利用提供了良好的平台,根据中医古籍的文献组织结构和用户需求设计合理的数据库结构,构建全文数据库和图片数据库,为数字古籍提供存储、检索、管理、利用的平台。同时为了更好地利用中医古籍数字化的成果,古籍数据库必须具有网络支持,以实现超链接的阅读环境、数据的远程传输利用、数据更新和软件升级等。中国中医科学院中医古籍资源库采用 Oracle9i 数据库,Oracle Text 数据搜索,可实现古籍全书浏览,也可以进行简单检索、组合检索和全文检索。

3.4 数字化中医古籍的支持服务功能

数字中医古籍的一大优势就是能够在海量的资源中快速找到所需信息。因此,如何提高检索功能对于数字中医古籍的利用至关重要。通过适度标引和多途径的检索方式,可以有效地提高查全率和查准率。数字化的中医古籍除应具有阅读、检索等基本功能外,还应具备支持服务功能,包括语义关联、辅助工具、参考文献及词频统计等。其中,语义关联将对中医古籍的利用提供重大帮助。由于古汉语与现代语言之间的差异,特别是专业词汇间语义的差别,中医古籍的利用存在一定障碍。要突破这一障碍,就要建立起中医古籍语言系统。语言系统是通过语义关联实现语义间的“互译”,但中医古籍的语言系统目前还在研究中。目前中国中医科学院中医古籍资源库采用了后控制词表的方法,就是在检索后台对中医古籍的检索结果进行语义控制,探讨实现中医古籍与现代语言之间的交流。后控制词表应用于数字中医古籍利用平台可以实现对中医古籍的更好利用,将会大大增强数字中医古籍的知识服务功能。

3.5 数字化中医古籍的整理

中医古籍数字化属于中医古籍整理的研究范畴。数字化后的中医古籍突破了文献的物理形态,深入到了它所包含的信息单元,利用计算机算法对这些信息

单元进行重组,实现古籍内容的结构化,可以向读者和研究人员提供针对性更强、内容更丰富的服务。中医古籍的知识挖掘还需要古籍的校勘、编纂、词切分、数据挖掘和知识库等技术的支持。

中医药专家和学者也一直对中医古籍的数字化整理进行探索。数字化中医古籍的整理必须要中医专家的参与,要忠于原著,尽量避免由于主观认识差异而产生的原文信息改动。

4 中医古籍数字化展望

中医古籍数字化是一项长期的系统工程,其建设应该兼顾保护和利用的双重目的,走标准化、规范化的道路,并需要中医文献与计算机专业人员的通力合作,才能构建出理想的中医古籍数字化成果。同时,中医古籍数字化应该在以下几个方面有所发展

第一,每种中医古籍是一个知识整体,包括典籍图像、文本、概念和元数据。现代文献中相关著作、学术论文、工具书、多媒体资料等,应当对每个独立的知识整体中的知识进行关联、整合,构建中医古籍的知识库,变中医古籍的资源服务为知识服务。

第二,中医古籍数字化资源的深度整理开发。中医古籍数字化应该突破文献载体本身的转变,对文献的

内容进行知识挖掘,进一步发现中医古籍内在的隐含知识。使数字化后的中医古籍成为中医药研究人员的知识宝库和高效的研究工具,使传统的中医古籍增值。

第三,中医古籍智能化处理技术的研究和应用。中医古籍数字化整理应深入到其内容本身,除了完成古籍文献的录入、存储、检索等基本功能外,还应该进行自动化、智能化处理方向的研究,进行中医古籍自动校勘、自动编纂、自动切分词、自动标点等方面的研究,开展中医药古文献计量学的研究。有了以上这些研究的支持,将会极大地提高中医古籍数字化的整理水平和学术研究价值。

参考文献

[1]毛建军.古籍数字化的概念与内涵[J].图书馆理论与实践, 2007(4): 82- 84.
[2]史睿.数字化条件下古籍整理的基本问题(论纲)[J/O L]. [2010- 07- 07]. http://demo.nlc.gov.cn/publish/wjls/html/8_07.htm.
[3]符永驰,刘国正.中医古籍保护与利用研究[J].中国中医药信息杂志, 2006, 13(4): 99- 101.
(收稿日期: 2010- 08- 01;修回日期: 2010- 09- 06)
[编辑: 李春梅]

(上接第 1124页)

NEI网络稳态机制紊乱引起并加重血管病变。这不仅证实了中医学气血相关的理论内涵,也有助于揭示情绪刺激、过劳等社会心理因素在血管病变发生中的作用。结果显示,通心络胶囊与辛伐他汀在改善血管内皮功能障碍局部病理损伤方面效果相当,在调节整体NEI网络失衡方面,通心络胶囊明显优于辛伐他汀,显示出通心络胶囊不仅修复局部血管病变同时具有调节NEI网络紊乱的优势。中医学把疾病看作是机体全身性功能失调的局部病理改变,治疗上把调整全身性功能状态包括调整气血阴阳平衡与局部病变的治疗有机结合。本研究结果显示,络气郁滞、虚滞证候因素可通过影响全身性NEI网络稳态失衡引起并加重血管局部病变,为上述理论提供了实验依据。

参考文献

[1] Black PH, Garbutt LD. Stress, inflammation and cardiovascular disease[J]. J Psychosom Res, 2002, 52(1): 1- 23.
[2]贾振华,高怀林,谷春华,等.基于 3469例血管病变患者证候分布规律的脉络病变共性病机探讨[J].中医杂志, 2009, 50(10): 920- 924.

[3]熊振芳,朱清静.柴胡疏肝散对慢性束缚应激性肝郁证大鼠的影响[J].中国中西医结合消化杂志, 2004, 12(4): 220- 221.
[4]程志清,姚立,龚文波,等. Wistar大鼠心气虚证模型的建立与评价[J].中国医药学报, 2003, 18(11): 654- 658.
[5]温进坤,韩梅,杜玮南,等.一种快速建立大鼠动脉粥样硬化模型的实验方法[J].中国老年学杂志, 2001, 21(1): 50- 52.
[6]梁晓芳.环境因素对冠心病心脏事件的影响[J].心血管病学进展, 2008, 29(2): 217- 220.
[7]修琳,刘广益,李作孝.应激对神经- 内分泌- 免疫调节网络的影响[J].医学综述, 2003, 9(11): 686- 688.
[8]王中琳.中医心身疾病研究态势评析[J].中医药学刊, 2003, 21(10): 1662, 1704- 1705.
[9]吴以岭.中医络病学说与三维立体网络系统[J].中医杂志, 2003, 44(6): 407- 409.
[10]吴以岭.气络- NEI网络相关性探析[J].中医杂志, 2005, 46(10): 723- 726.

(收稿日期: 2010- 11- 08;修回日期: 2010- 11- 14)
[编辑: 行 之]