【引文格式】张旭,赵彦辉,刘树春.本草古籍数字化及嵌入学术资源平台的探索与实践[J].中国中医药图书情报杂志,2017,41(6):5-9. DOI: 10.3969/j.issn.2095-5707.2017.06.002

·中医药信息研究·

本草古籍数字化及嵌入学术资源平台的探索与实践

张旭[1]，赵彦辉[2]，刘树春[2*]

1.辽宁中医药大学药学院，辽宁 大连 116600；2.辽宁中医药大学图书馆，辽宁 沈阳 110032

摘要：本文回顾了国内古籍数字化的现状，比较了不同类型数字化的特点，讨论了中医药专业古籍数字化建设及本草类古籍在内容结构上的特殊性，分析对比了国内常见的中医药专业古籍数据库的功能特色。以《植物名实图考》为例，开展了本草古籍数字化服务模式的新尝试。提出将数字化古籍进行知识点切割和文字识别，经过编号、命名、标引，形成独立的知识单元，嵌入到基于 J2EE 的 SSH 框架的东北地产药用植物学术资源平台，通过语义知识点与平台相关联，扩展检索路径和利用方式，形成专题性知识服务系统。在丰富平台服务内容的同时，扩展古籍的利用途径和探索古籍数字化的新模式，有助于对古籍进行深入挖掘和利用。

关键词：本草古籍；古籍数字化；资源整合；知识嵌入；知识服务

中图分类号：G250.7　　**文献标识码：**A　　**文章编号：**2095-5707(2017)06-0005-05

Exploration and Practice of Digitization of Ancient Books about Chinese Materia Medica and Embedding Academic Resources Platform

ZHANG Xu[1], ZHAO Yan-hui[2], LIU Shu-chun[2*]

(1. College of Pharmacy, Liaoning University of Traditional Chinese Medicine, Dalian Liaoning 116600, China; 2. Library of Liaoning University of Traditional Chinese Medicine, Shenyang Liaoning 110032, China)

Abstract: This article reviewed the present condition of the digitization of ancient books in China, compared the characteristics of different types of digitization, discussed the particularity of digitization construction of professional TCM ancient books and books of Chinese materia medica in contents and organization, and compared the functional features of the common TCM professional ancient book databases in China. Taking the *Zhi Wu Ming Shi Tu Kao* as an example, this study conducted a new try for digitization service mode of Chinese materia medica books. It proposed semantic knowledge point cutting and character recognition for the digitized books, form an independent knowledge unit through numbering, naming and indexing, and to embed to the Platform of Northeast Local Medicinal Plant Academic Resources based on SSH framework of the Java. The expansion of the use of ancient books and exploration of the new mode of digitization of ancient books can be realized at the same time with enriching platform services, which can help deep excavation and use of ancient books.

Key words: ancient books about Chinese materia medica; digitization of ancient books; resource integration; knowledge embedding; knowledge service

---

基金项目：辽宁省教育厅优质资源共建共享专项；辽宁省高等学校图书情报工作委员会特色资源库建设专项（L2016018）

第一作者：张旭，2014 级中药学专业本科在读。E-mail：1378950687@qq.com

*通讯作者：刘树春，研究馆员，研究方向为中医药信息挖掘。E-mail：scliu45@sina.com

古籍是指以纸为载体抄写或未采用现代印刷技术印制的书籍，而且这些书籍往往经过百年甚至千年的保存和利用，已经非常脆弱。为了对其实施保护，同时方便开发和合理利用，最有效的方法是进行数字化处理，实现古籍整理、存储、检索、阅读及传输的电子化。虽然古籍数字化相关研究与实践探索已经有 30 余年的历史[1]，但在数字化技术、采取的数字化模式、建立的服务平台等方面还存在着参差不齐的现象，在对古籍的保护和利用方面还有诸多需要探索之处。本文以本草类古籍为例，对古籍数字化的路径、技术、方法及嵌入至学术资源服务平台的可行性进行探讨。

1 古籍数字化研究与实践现状

古籍保存对温湿度、照明、紫外线、空气净化、通风、防虫防鼠、消防安防等各种环境要求非常高，最重要的是在古籍使用过程中的人工磨损给古籍的保存和利用带来现实上的矛盾。随着计算机技术的应用普及，自 20 世纪 80 年代初开始，在我国陆续开展了对古籍的数字化研究探索，既有助于对古籍进行永久性保存及再生性保护，以减少因环境和人为等因素造成的损失，同时还可以方便对古籍的整理、存储、交流、传播与利用，促进对古籍文献信息开展有效利用和深入研究。

在古籍数字化研究与实践中，存在着不同的数字化处理方式和服务模式。除了古籍书目数字化以外，最主要的是将古籍以“文本版”或“图像版”形式数字化，以光盘或磁盘作为存储媒介提供浏览检索服务。两种方式在文字识别、全文检索、存储空间、浏览阅读等方面各有优缺点[2]。目前，无论是对单种古籍的数字化还是对批量古籍数字化后建立数据库，无论是图像库还是文本库，无论是光盘版或是网络数据库平台，基本都是以图书整体为单位提供浏览或检索阅读服务，可以称为文献型数据库。在已经建设的古籍数字化平台所采取的文本型、图像型和图文型等数字化处理类型和服务模式中，也体现出不同的特点（见表 1）。

表 1 不同类型的古籍数字化方法及特点比较

类型	数字化方法	优点	缺点
文本型	通过手工方式重新录入及人工校对，形成古籍全文的电子文本	经过标引、校勘，可全文检索，且贮存空间小，检索速度快	不能保持古籍原貌，文字录入难度大，生僻字难处理，易出现错讹，不利于古籍研究
图像型	将古籍直接以图像格式扫描并著录题名、作者、版本等题录信息	录入简单，能保持古籍原貌，避免错误，便于古籍研究	不能全文检索，贮存空间较大
图文型	扫描图像格式，并 OCR 识别，转换为文本，校对，并提供古籍的图文对照	既可实现全文检索，又可浏览原貌古籍，方便古籍研究与利用	易出现文字差错

在古籍数字化过程中，除了整体数字化并提供服务外，也有学者提出了一种基于知识元的知识表示方法。通过对中医古籍知识结构、语义解释方式以及语义关系的分析研究，建立中医古籍语料库，对古籍知识元进行解析，实现基于内容的数据库检索和知识关联[3]，从而在常规的古籍文献型数据库的基础上，进一步发展成为古籍知识库。

2 专业性古籍数据库平台建设及本草类古籍特殊性

2.1 中医药古籍数据库建设现状

经过探索与实践，在初期的注重综合性古籍图像或文本数据库建设的基础上，逐步扩展建设专业性、专题性古籍数据库。国内的中医药信息研究机构在中医古籍数字化方面也取得了可喜的成果，陆续建立了多个中医药专业性古籍文献数字化服务平台，如由中国中医科学院开发的“中医药珍善本古籍多媒体数据库”“海外回归中医古籍善本集粹”等。此外，国内专业数据库公司也相继开发建设了多个中医药古籍数据库平台，而且这些平台在收录古籍数量、录入方式、利用途径、服务模式等方面各具特色，基本实现了文字识别或录入、人工校对、全文检索、图文对照等功能（见表 2）。

2.2 本草类古籍的内容结构特点

本草古籍是中医典籍的重要组成部分，记载着中草药在疾病治疗、食疗养生、美容保健等方面的应用，凝聚着古代医家的临床实践经验。古籍数字化建设为本草古籍的保护和开发利用带来新的契机，为相关研究提供了更加丰富的素材。

与中医药其他类别的古籍相比，本草类古籍在编制结构和内容上具有结构性明显和条目化清晰等特点。例如《植物名实图考》，每个植物药均为一个完整的结构化条目，包括：植物药名、别名、功能主治、生长特点、药用方法、用法用量及注意事项等，构成了完整的知识单元。其他本草类古籍也有类似结构。一般药用植物的记载大都包含名称

（别名、俗名、代称）、分类（上中下三品、来源、自然属性、功能分类）、来源（物种、部位、生境、记载）、性味（阴阳、五行、四气、五味、归经、升降、毒性）、配伍（单行、相须、相使、相畏、相杀、相恶、相反、君、臣、佐、使）、功用（功效、副作用）、主治（主证、主病）、组方、炮制（制法、器具、炮制时间、辅料、贮藏、禁忌）、采收（时间、方式）、地域（产地、道地）、鉴定（色泽、气味、形状、质地、辨伪、质量、类药）、用法（入药方法、服用方法、服药时间、用量、注意事项）、禁忌（配伍禁忌、饮食禁忌、人群禁忌）、引用（人物、论述）等知识点。

表 2　5 种中医药古籍数据库平台及特色比较

平台名称	制作单位	收录范围	功能特色
中华医典	嘉鸿科技	中医药古籍 1156 种	重新文字录入，无原文图像，实现文字自由选择拷贝、关键字词检索，检索结果可直接导出文本
中医典海	北京爱如生数字化技术研究中心	中医古籍 1000 种	手工录入，图文逐页对照，还原式页面，快速全文检索，可编辑、下载和打印
瀚堂典藏中医药文献库	北京时代瀚堂科技有限公司	中医古籍 750 种	采用国际通用超大字元集进行加工校勘，文本精准无缺字，图文对照，高速全文检索
中医中药古籍大系	北京书同文数字化技术有限公司	中医药古籍 104 种	图文对照，文本页面保持了原书板式，差错率低，实现全库整合全文检索
域外汉籍数据库	上海睿则恩信息技术有限公司	海外汉籍子部医家类 270 种	彩色扫描图像，可按分类浏览及按书名、作者、年代、版本等题录检索

本草类古籍的这些特点有利于对知识单元的抽取并与其他相关数字资源进行整合和相互关联，以及在全文对照和构建多途径检索功能方面实现更为精准的检索。同时，也有利于对相关概念、属性、功能主治的聚类和社会网络分析。因此，在对本草类古籍的数字化研究探索中，有学者在构建图像库的基础上，进行文字识别、解析和校对处理，进一步构建数字化文本库，并实现对古籍的字词频统计和异体字汇聚显示等辅助研究支持功能，建立集加工、阅读、检索、维护、交流为一体的本草古籍数字化信息平台[4]。

3　本草古籍数字化服务模式的新尝试

3.1　数字化古籍嵌入平台的设计思想

根据过去数十年古籍数字化的经验总结及本草类古籍的编制特点，我们结合“东北地产药用植物学术资源平台建设项目”，尝试将本草类古籍数字化并嵌入平台结构中，整合平台服务与古籍知识，通过古籍内容的知识点与平台相关联，形成专题性知识服务系统，从原来的古籍文献服务向古籍知识服务的转化，有助于对古籍文献的知识挖掘与利用。本研究以在历代本草著作中收载植物数量最多的清代古籍《植物名实图考》为例，对其数字化过程和平台嵌入方法进行探索尝试。

3.2　制定图像扫描原则并实施

根据扫描设备状况及古籍数字化平台的需要，制定详细的古籍图像扫描规则，以及图片编号、文件夹命名、工作量计算、任务分工等方法细则。并根据选择的书目和版本，有计划地进行古籍图像扫描和系统编号。

3.3　图像处理及知识单元抽取

为了便于数字化平台对古籍的识别和应用，根据平台的要求，对扫描的图像进行色彩转换、去噪、倾斜度校正等处理。根据《植物名实图考》内容编制结构和知识点进行图片切割、文字识别和人工校对，并分别进行编号、命名、标引，形成独立的图像和文本格式的知识单元，上传服务器。

3.4　嵌入学术资源平台

平台建设的总思路是参照已有的中医古籍数字化建设成果并结合东北地产药用植物学术资源平台现已开发运用的状况，基于 J2EE（Java2 平台企业版）的 SSH 框架（struts＋spring＋hibernate 的集成框架）予以实施。平台设计对古籍内容提供图像和文本格式两种显示界面。将本草古籍的知识单元内容通过超文本链接嵌入到平台药用植物的相应条目中，并借助平台的多种检索途径实现对本草类古籍知识的灵活利用。

4　学术资源平台的框架与模块功能设计

4.1　平台的界面设计

在东北地产药用植物学术资源平台系统框架的基础上，对平台系统和子系统进行重新设计和扩

充，增加古籍全文图片和文字对照浏览页面；增加药用植物参考文献出处，通过超链接与古籍知识单元图像相关联；增加后台文献著录、全文提交和语义标注等管理页面。

SSH 框架属于轻量级应用型框架，在实际应用中注重软件设计的可复用性和系统的可扩展性，应用广泛，从逻辑层面上分为用户界面层、业务处理层和数据存储层。用户界面层分为前台用户界面和后台管理员界面，是进入学术资源平台的窗口。前台用户界面包括检索服务和类目导航，提供系统登录、密码修改、系统退出等。后台管理员界面包括药用植物增删改查、文献题录管理、全文语义标注及用户管理等页面。业务处理层是数字化系统框架体现核心价值的部分，处于用户界面层和数据存储层之间，可起到数据交换承上启下的作用[5]。根据用户界面层发出的请求，在数据存储层获取相关数据传送给用户界面层。数据存储层中储存了整理后的所有数据资料，在保证安全性和完整性的前提下实现对数据库的维护和管理。

4.2 平台的模块设计

根据学术资源平台的功能需求，将系统分为药用植物管理、古籍书目管理、古籍知识元管理、用户及系统管理等 4 个模块。其中古籍书目管理和古籍知识元管理 2 个模块最为核心，内含文献著录信息、古籍原文图像和平台原有的按科属分类的药用植物资料。对系统进行模块设计，不仅使古籍数字化加工更为高效、方便，还使用户可以在任意时间和地点通过网络访问系统平台，实现真正意义上的资源共享。

4.3 平台的功能设计

东北地产药用植物学术资源平台原有的设计功能为提供按科、属分类的药用植物浏览方式。在此基础上，结合本草古籍的内容对其功能进行调整和扩充，在平台的主界面提供按现代科属分类、按药用植物名或拉丁名浏览及利用关键词等多途径的全文检索功能。

现代科属分类浏览功能即原有的检索方式，可以在菜单中根据植物的类型、科属种进行浏览，查找所需要的植物，进而找到该植物的鉴别特征、入药部位等文字信息及图片信息。

本草古籍检索功能则分为 2 种途径。一是在本草古籍元数据录入时，将古籍中所论述的植物的属性和性状进行标引，以实现在菜单中根据植物的性状和属性在本草古籍原文中找到对应的相关描述；二是将古籍中的植物按科分类整理并进行标引，可以实现利用植物所属的科在菜单中进行搜索。

5 古籍数字化并嵌入学术资源平台带来的启示

数字化古籍嵌入学术资源平台实现学术资源与图书文献知识单元的整合，最关键的步骤是元数据的录入。元数据是古籍数字化的基础，是数据共享的主要接口。从目前本草古籍数字化的实践来看，其应用范围窄、规模相对较小的原因在于本草古籍知识的元数据标准不统一。国际标准化组织 2014 年 6 月发布了《中医药学语言系统语义网络框架》（ISO/TS 17938）和《中医药文献元数据》（ISO/TS 17948）国际标准，为本草古籍数字化内容的语义标引奠定了基础。

在元数据录入过程中要根据本草古籍的编制特点，尽可能涵盖古籍的完整信息。一是版本信息。很多古籍会因重刻、重印或被后世校注、点校等原因，造成出版社、出版时间、编著者等发生变化的问题。因此要仔细考证并标明版本类型、年代、版式特征及其出版、编著信息（字、号、朝代、生卒、籍贯）等。二是本草古籍的分类信息。同种古籍在不同的文献收藏单位也难以实现统一的归属类目。刘培生等[6]研制的《中医古籍分类表》在古籍分类中可以作为统一分类参考。三是本草古籍定级信息。对古籍所属的朝代、版刻形式、内容、存世价值等珍贵程度进行鉴定及等级评定。

同时，数字化古籍嵌入学术平台还要确保平台系统的安全性、数据的完整性以及平台操作的兼容性，确保用户在使用过程中安全、方便、高效。

6 小结

本草古籍数字化不仅是载体类型的改变，更重要的是古籍利用方式和利用深度的改变，对古籍的开发与利用有很大的促进作用。将数字化本草古籍与药用植物学术资源平台相整合，为进一步开发本草古籍的学术价值开辟了空间，具有可行性。但是，由于受到本草知识表示、存储，及软件、硬件环境、信息技术手段等多因素限制，使中医药相关知识达到全面一致的理解和共享还存在着一定的局限性，还需要更进一步的研究和探讨。

参考文献

[1] 龚娅君,刘春金.中文古籍数字化建设[J].浙江大学学报(人文社会科学版),2006(4):174-176.

[2] 吉聪.中医古籍数字化建设问题探讨[J].长春中医学院学报,2004,

20(3):64-65.

[3] 柳长华. 基于知识元的中医古籍计算机知识表示方法//中国中医科学院, 世界中医药学会联合会. 第三届国际传统医药大会文集[C]. 中国中医科学院, 世界中医药学会联合会, 2004:47.

[4] 裴丽, 曹霞, 张宏伟. 本草古籍数字化信息平台现状与实践[J]. 中医药学报, 2013, 41(4):30-33.

[5] 曹霞, 常存库, 裴丽. 中医古籍数字化建设及其平台设计和实现[J]. 中华医学图书情报杂志, 2016, 25(3):45-47, 53.

[6] 刘培生, 张伟娜, 李鸿涛, 等. 《中医古籍分类表》的研制及应用[J]. 中国中医药图书情报杂志, 2017, 41(2):52-54.

（收稿日期：2017-08-04）

（修回日期：2017-09-18；编辑：魏民）