国家中医临床研究基地文献信息库系统研究与设计

徐 荣

（中国中医科学院广安门医院 北京 100053）

〔摘要〕 以中国中医科学院广安门医院为例，介绍国家中医临床基地文献信息库建设原则，设计中医药特色文献系统，阐述系统架构、功能、建设步骤以及围绕基地 3 个重点病种建设的 7 个文献信息库主要内容。

〔关键词〕 国家中医临床研究基地；文献信息库；系统设计

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2019.11.018

Study and Design of Bibliographic Database System of National Clinical Research Base of Traditional Chinese Medicine *XU Rong, Guang'anmen Hospital, China Academy of Chinese Medical Sciences, Beijing 100053, China*

〔**Abstract**〕 Taking Guang'anmen Hospital of China Academy of Chinese Medical Sciences as an example, the paper introduces the building principle of the bibliographic database of National Clinical Research Base of Traditional Chinese Medicine (TCM), designs the bibliographic system with a distinctive feature of TCM, and elaborates on the system architecture, functions, building procedures and the main contents of the seven bibliographic databases built centering on three major disease species in the research base.

〔**Keywords**〕 national clinical research base of TCM; bibliographic database; system design

1 引言

国家中医临床研究基地是 2008 年 12 月国家发改委和国家中医药管理局共同启动实施的建设项目[1]，旨在通过基地业务建设工作，系统构建中医临床研究、协作攻关、成果转化推广平台，培养领军人才，全面提高自主创新能力，提升中医药防病治病能力，促进中医药事业的发展。文献信息库作为中医临床研究基地基础平台建设的一项重要内容，对搜集临床研究基地重点病种古今中外诊疗信息及研究资料、挖掘和整理中医诊疗经验、实现对中医诊疗经验的传承和创新有着非常重要的作用。中国中医科学院广安门医院是国家中医临床研究基地之一，围绕本基地重点病种肺癌、糖尿病、冠状动脉粥样动脉硬化性心脏病进行文献信息库系统研究和设计。

2 建设原则

2.1 规范实用性

在文献信息库建设过程中要以国际、国家、学科领域标准规范为基础，着眼于信息库的服务对象、内容组织、技术架构等多个方面，形成相对完整的规范体系[2]，建设易操作、易管理、易维护和易扩展的各类文献信息库。

〔收稿日期〕 2019－04－01

〔作者简介〕 徐荣，硕士，馆员，发表论文 10 篇。

2.2 安全可靠性

文献信息库的建设要有强大的安全保障体系来保证系统中数据存储和传输的安全。要选用高可靠性设备和技术支持数据资源的冗余、备份、容灾、恢复等功能[3]。同时还要建立一整套安全管理制度，从管理和技术上确保系统及其资源的安全访问与监控。

2.3 特色继承性

文献信息库的建设要依托广安门医院数字图书馆平台，全方位展示肺癌、糖尿病、冠状动脉粥样动脉硬化性心脏病这 3 个重点病种的临床及科研成果，根据不同病种研究需求，全面收集古今中外的诊疗信息及研究资料，使文献信息建设集成化、动态化、知识化，满足用户个性化的信息需求，实现对信息库内容的实时更新和拓展。

3 系统设计

3.1 平台架构

文献信息库服务系统平台采用大量的元数据作为数据源，本地底层数据通过管理层进行管理，应用层对其进行各种应用的分布式架构。在数据底层定期更新，在管理层管理各种元数据及各级机构、学者、科研成果的对照关系，在服务层为用户展示重点病种的全方位、多层次的信息数据[4]，提供检索和指标评价分析服务。文献信息库平台框架，见图 1。

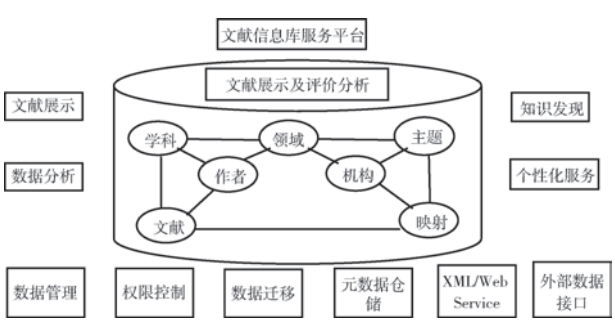

图 1 文献信息库平台框架

3.2 功能模块

3.2.1 信息采集 分为两个子系统，即互联网信息采集和本地文献数据库采集，可根据用户指定的数据采集范围进行模板定制开发，信息采集功能架构，见图 2。信息采集系统支持对各种数据库及网页内容的解析和抓取，包括各种附件和音视频内容。系统具有高效的数据去重处理机制和多种对网络采集屏蔽技术的反制措施，采集内容的噪音去除和正文自动抽取准确率高，能帮助用户有效利用网络资源和降低功耗。

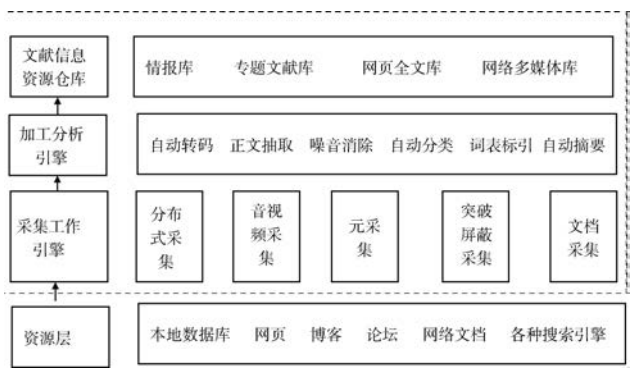

图 2 信息采集功能架构

3.2.2 分布式全文检索 分布式全文检索系统（SolrCloud）作为搜索引擎的重要组成部分，为用户提供平台各种核心资源的检索服务，系统部署，见图 3。

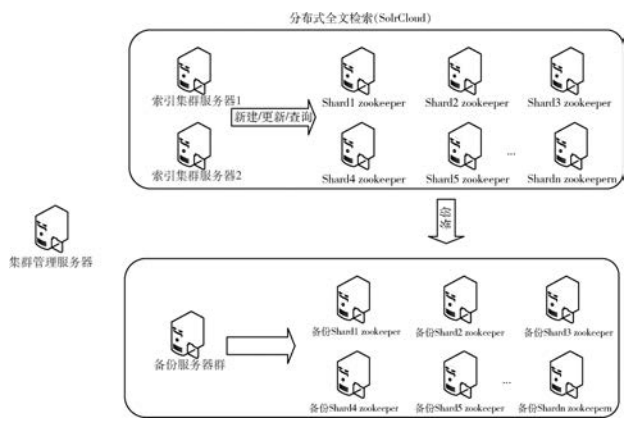

图 3 分布式检索系统部署

使用 Zookeeper 作为集群的配置信息中心，实现集中式配置信息，将 SolrCloud 的相关配置文件上

传 Zookeeper，多机器共用。实现自动容错，SolrCloud 对索引分片并对每个分片创建多个备份。每个备份都可以对外提供服务，1 个备份出现错误也不会影响索引服务；实现近实时搜索，立即推送式的备份可以在秒内检索到新加入索引。此外，SolrCloud 在用户查询时可以实现自动负载均衡，SolrCloud 索引的多个备份可以分布在多台机器上，均衡查询压力[5]。如果查询压力大，可以通过扩展机器，增加备份来减缓。

3.2.3 数字化加工 主要是为实现已有内容资源的结构化拆分解析，为内容资源的产品化重组奠定数据基础。数字化加工系统可以进行文本、图片、表格拆分并对经过加工的数据进行清洗和规范化存储。（1）文本拆分。系统可对文档进行细化到段落层级的拆分加工并将拆分的结构进行结构化存储。处理组件首先将读入的待处理文档进行载入，依据挂载的待解析内容模块依次将每个资源项解析出需要的元数据，将这些数据传递给存储组件进行后续处理。存储组件按照元数据类型将不同资源存储至预定义的目录结构及数据库中。文档拆分整理完毕后，各碎片可作为元数据项供其他程序使用。（2）图片拆分。对文档中的图片进行单独提取，拆分后的每张图片及其附属文件均存放于独立文件夹下，每张图片均保存原图（原分辨率导出）、低分辨率图（根据用户在页面中输入的数值导出）、预览图（72dpi 导出）及描述文件。（3）表格拆分。对文档中的结构化表格进行单独提取，拆分后的每个独立表格均存放于独立文件夹下，每个表格保存为一个对应的独立目录，目录下存放表格对应的描述文件，描述文件中存储表格碎片的位置、内容信息。

3.2.4 元数据管理及资源仓储 文献信息库建设采用国际通用标准都柏林核心元素集（Doblin Core Element Set，DC），依照中国高等教育文献保障体系《特色库项目本地系统技术规范》以及国家中医药管理局制定的《中医药文献数据库数据来源规范》、《中医药文献数据库数据资源加工指导规范》[6]等相关标准进行元数据处理。系统平台能够进行元数据类型、映射以及索引管理，其功能架构，见图 4。系统提供元数据定义与编辑维护功能，包括设置元数据的规范名称（中英文）、数据库数据类型、solr 字段类型、字段描述、是否多值等；可对数据源与信息库字段进行一一映射，完成从数据源到知识库的数据导入操作；能够直接与检索逻辑相关联，灵活设置元数据是否可索引、可查询展示及权重和排序策略配置，管理与维护情报分析所需要的各级分类，实现不同类型资源整合以及统一检索。基于元数据存储的数字资源仓储系统支持元数据存储、添加、修改、删除、整合以及数据的导入和导出。数字资源仓储系统能够为不同特色的文献信息库构建不同结构规范的元数据仓储库，通过不同适配器从不同数据源中提取数字资源的元数据信息，将元数据信息通过生成的结构保存，提高文献信息库建设效率。

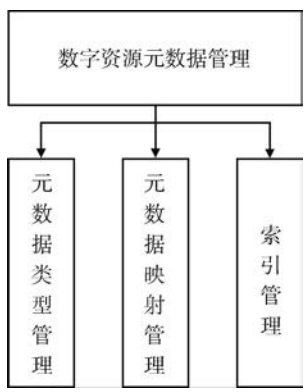

图 4 数字资源元数据管理功能架构

3.3 数据库内容设计

3.3.1 信息来源及信息组织方式 文献信息库资源来自于期刊论文、会议论文、学位论文、报纸、图书、专利、报告、标准、网页及论坛等，支持按年度、成果类型、收录类型、学科、基金、数据来源等多个方面筛选查看内容。突破以往以文献为单位进行信息组织的模式，在一些特色数据库如诊疗方剂、特色制剂、中医古籍文献库等的建设过程中对中医药文献题录数据、引文数据以及中医古籍内容全部进行关联和深度标注，全方位构建元数据之间的关系网络。确定对应的特色数据库中存在的所有数据类型。在数据资源的

基础上确定数据库的专家、特色方剂、特色疗法、科研成果以及报道和研讨活动等信息，这些信息将以元数据异构共享的形式组成特色数据库的共享资源。

3.3.2 建设知识本体形式化的特色数据库 在文献信息库的构建过程中，基于本体的模型能够有效地明确特色数据库内容和形式、标准与规范等，针对各种数据资源类型的相互关系和外部数据源的连接标准，形成元数据与知识本体形式化的特色数据库。本体构建主要包括创建、管理和服务 3 个方面[7]。在本体创建方面，要实现从结构化词表中自动获取概念和概念层级结构以及从中医药相关专业教材和文献中学习概念的属性及其属性值，提取领域知识概念之间的关系；在本体管理方面，要实现多人在线的辅助校对和版本管理，文献信息库的知识关系更多地蕴含在深层次的领域知识当中，要获取深层次的领域知识关系，应充分运用图书情报学领域知识，同时也需要各个临床重点学科专家对所建立的知识关系进行及时校对，形成各个临床专家和本体学习功能的良性互动；在本体服务方面，要实现本体知识的搜索和展示，主要包括知识导航、知识检索和概念关联的可视化展示、概念属性的展示。在建设过程中首先需要根据不同类别的中医药特色文献信息库元数据体系构建形式本体模型，有效表达各种数据资源和类型的形式化。将各类专题文献信息数据库的系统性、学术性以及深度广度等信息作为数据库的构建内容，通过对某些特色领域中元概念的构建以确保最终的数字化表现形式能够满足不同类别数据库之间元数据共享的需求。将需要建设的文献信息库中的元概念和元关系描述出来，构造具有元数据与知识本体形式化的特色数据库模型。在数据库建设过程中应与各个重点病种以及计算机学科专家互相配合，构建出适合不同临床学科的知识本体形态。根据国家中医药管理局科技司对国家中医临床研究基地业务建设方案的要求，项目建设的 7 个文献信息库主要内容包括：（1）专家。重点收集各个科室代表专家基本信息、临床经验、学术思想、研究成果、发表的论文论著和媒体报道等信息。（2）优势病种。针对各个学科单病种的中医疗法、诊疗方案、诊疗技术、疗效评价方法等信息资源进行整合，全面搜集与该病种有关的论文、论著、视频及课件等信息。（3）诊疗方剂。收集古籍及现代文献中的古今中药方剂，全面介绍方剂信息，提供有关方剂药味组成等统计信息，详细介绍每一方剂的不同名称、处方来源、药物组成、功效、主治、用药禁忌、药理作用、制备方法等信息。（4）特色制剂。对学科在中医理论及临床实践中研制出的特色制剂进行介绍，包括药物的合理组方、功能主治、用法用量及不良反应采集等。（5）特色疗法。针对各个学科在长期临床经验中形成的大量特色诊疗方法进行归纳总结，收集疾病的概述、诊疗要点、辨证要点、治疗规范、疗效评定标准、临床分期等信息。（6）中医古籍。对中医经典古籍进行数字化加工，对其中的医经、医理、诊断、针灸推拿、本草、方书、临证各科、养生、医案医论医话、医史等信息进行分类整理、标引入库。（7）科研成果。重点介绍科研成果的项目完成人、完成单位、研究内容、研究结果和研究意义，该数据库应充分展示不同学科的研究成果，揭示该学科在国内乃至国际上所处的研究水平及地位。

4 建设步骤

4.1 调研与初步实施

通过文献阅读以及实况调研确定文献信息库建设的技术方案以及基本框架，形成总体建设方案，将 3 个重点病种作为试点，进行文献信息库建设工作。确定数据采集加工、质量控制以及著录标引规范。与 3 个重点病种建立密切联系，了解学科需求，确定文献信息库建设的核心技术及基本框架。邀请相关专家对信息库建设方案进行论证评估，根据评估意见修改建设方案，开展信息库建设工作。

4.2 分析与建设

根据前期制定的文献信息库建设方案进行文献的收集、著录、整理和入库工作，搭建出中医学科

特色化文献信息库的框架与模式。针对不同类型文献信息库的具体要求确定数据库结构，对参与文献信息库建设的人员进行技术培训，开展文献信息库内容的收集整理以及数字化加工工作。首先确立文献搜集的范围和检索策略，分别交由各临床科室或研究室的人员进行文献资料的搜集整理工作；其次通过信息采集系统对各种数据库及网页内容的解析和抓取，结合词表、自动识别技术，对采集数据内出现的内容实体进行自动识别和抽取并进行存储；最后通过数字化加工技术对已有内容资源的结构化拆分解析并将拆分的结构进行结构化存储，为不同文献类型数据库的建设奠定数据基础。将检索到的文献逐条分析，按类别进行标引、著录，导入到相应的文献信息库中，形成文献信息库的整体模型。

4.3 运行

将试点科室的文献信息库建设方案和成果逐步推广到其他临床科室及研究室，逐步建立系统、完善、全面反映基地临床及科研成果的文献信息系统应用平台。

5 结语

临床科研平台文献信息库基于广安门医院数字图书馆的平台，建立集综合检索、开放获取、学术分析、个性化服务于一体的中医药特色文献信息库及服务系统，为中医临床研究基地建设提供强有力的信息保障。

参考文献

1 马红敏，邓文萍，孙静. 国家中医临床研究基地标准信息管理系统研究与设计［J］. 中国数字医学，2014，9（10）：34－36.

2 卢传坚，陈淑慧，蔡桦杨. 国家中医临床研究基地科研创新平台设计初探——基于广东基地的实证研究［J］. 中国卫生事业管理，2016，33（5）：360－362.

3 赵爽. 医院信息系统安全分析与管理［J］. 医学信息学杂志，2018，39（11）：32－34.

4 刘红丽. “互联网＋”时代医学高校数字图书馆知识发现系统研究［J］. 医学信息学杂志，2017，38（5）：11－15.

5 艾金勇. 藏学文献特色数据库建设实践［J］. 智能计算机与应用，2016，6（4）：45－47.

6 于琦，崔蒙，李园白. 中医药文献数据库建设规范研究［J］. 世界科学技术—中医药现代化，2014，16（11）：2304－2307.

7 王静. 基于元数据异构共享的艺术院校图书馆特色数据库建设研究［J］. 图书馆学刊，2018，40（6）：53－57.

（上接第 80 页）

3 Quintana RM, Haley SR, Levick A, etal. The Persona Party: using personas to design for learning at scale［C］. New York: 2017 ACM SIGCHI Conference on Human Factors in Computing Systems , 2017.

4 Amato G, Straccia U. User Profile Modeling and Applications to Digital Libraries［C］. Berlin: 3rd European Conference on Research and Advanced Technology for Digital Libraries, 1999.

5 Mao J, Lu K, Li G, et al. Profiling Users with Tag Networks in Diffusion－based Personalized Recommendation［J］. Journal of Information Science, 2016, 42（5）: 711－722.

6 王庆，赵发珍. 基于“用户画像”的图书馆资源推荐模式设计与分析［J］. 现代情报，2018，38（3）：105－109，137.

7 许鹏程，毕强，张晗，等. 数据驱动下数字图书馆用户画像模型构建［J］. 图书情报工作，2019，63（3）：30－37.

8 陆尧，杨代庆. 区域图书馆联盟文献传递用户行为分析［J］. 图书馆论坛，2019，39（5）：88－94，126.

9 百度百科. 会议文献［EB/OL］.［2019－03－01］. https://baike.baidu.com/item/%E4%BC%9A%E8%AE%AE%E6%96%87%E7%8C%AE/10605294?fr=aladdin.

10 百度百科. 学位论文［EB/OL］.［2019－03－01］. https://baike.baidu.com/item/%E5%AD%A6%E4%BD%8D%E8%AE%BA%E6%96%87/4678889?fr=aladdin.