• 医学信息资源管理与利用 •

高等医学院校图书馆专题知识库系统设计与实现

王 霞 汤 琳 于千策 木 楠 曹洪欣

（海军军医大学 上海 200433）

〔摘要〕 分析专题知识库平台系统需求，详细阐述高等医学院校图书馆专题知识库系统设计方法，包括数据源及数据处理、系统功能结构及检索接口设计、系统开发及关键技术实现、专题库平台数据制作流程等。

〔关键词〕 医学院校图书馆；专题知识库；系统设计；系统开发

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2023.02.015

Design and Implementation of the Thematic Knowledge Database System in Libraries of Medical Universities and Colleges

WANG Xia, TANG Lin, YU Qiance, MU Nan, CAO Hongxin, Naval Medical University, Shanghai 200433, China

〔**Abstract**〕 The paper analyzes the requirements of the platform system of the thematic knowledge base, and elaborates the design methods of the thematic knowledge base system of libraries of medical universities and colleges, including the data source and data processing, the system function structure and the design of retrieval interface, the system development and the realization of key technologies, the data production process of the thematic database platform, etc.

〔**Keywords**〕 libraries of medical universities and colleges; thematic knowledge base; system design; system development

1 引言

随着科学技术的迅猛发展以及信息技术的广泛应用，各类信息资源呈爆炸式增长。生物医学文献量每年都在大幅递增，海量而丰富的生物医学信息资源选择、利用难度随之提升。高等医学院校图书馆，有责任及时向本校以及各附属医院的教学科研人员提供医学信息的来源情况、文献分布规律，并使其能够通过简便的检索途径及时准确地获取所需要的文献资料[1]。利用计算机、网络、信息组织与处理最新技术，将零散、多元、丰富、异构、多源的生物医学专题信息资源进行序化、组织并统一揭示和发布，以满足特定用户实时、专业、共享的信息需求，已成为高等医学院校图书馆创新读者服务的有效途径[2]。因此设计与开发能够解决多源异构、批量数据输入与输出、多途径检索与多维结果揭示，集数据采集、组织、管理与检索输出等功能于一体的专题知识库平台系统尤为必要。

〔修回日期〕 2022－09－21

〔作者简介〕 王霞，副教授，发表论文 30 余篇；通信作者：曹洪欣，教授，发表论文 60 余篇。

2 专题知识库平台系统需求分析

2.1 总体需求

高等医学院校作为高等院校中一个独特的体系，其专题库建设也具有特殊性。需要紧密结合本

校教学科研实际，摸索出一套符合自身特点的专题库建设的规律和方法，搭建一个种类丰富的科研文献资源一站式服务平台。专题知识库覆盖生物医学专题领域的国内外期刊、专著、专利、标准、指南、会议、学位论文、网络以及其他相关文献，需要建立一个跨资源类型、跨学科、跨主题的文献资源统一标引、检索、揭示系统[3]。

2.2 功能需求

2.2.1 功能组件 专题知识库平台系统的功能组件主要分为数据解析、用户检索以及平台管理相关功能模块。其中数据解析主要提供所采集数据的解析入库以及后续的数据索引制作、全文导入等功能；用户检索主要提供文献题录信息检索和全文数据获取等功能；平台管理主要提供系统相关数据的增删改查等功能[4]。

2.2.2 平台系统架构 平台采用灵活可变、可扩展性强的数据架构，实现的功能如下：灵活多样的资源导航功能，包括快速检索以及高级逻辑组合检索功能；多维度聚类统计和筛选功能（资源类型、来源出处、学科分类、关键词、评价指标等）；灵活多样的用户认证体系，可实现帐号或 IP 授权方式的验证和登录；灵活便捷的多种格式数据导入/导出功能，支持 PDF 在线阅读、音视频在线播放与浏览等操作；详尽的访问日志记录和统计分析功能，全接口化设计，界面与业务功能独立，通过 Json/XML 进行内外部系统的数据对接调用；系统资源模块独立，资源自定义组合、自定义创建主题导航及批量数据导入和编辑；个性化文献索取、个人中心及资源分享[5]，见图 1。

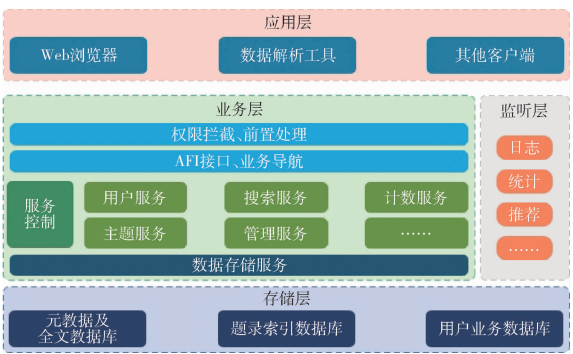

图 1 专题知识库平台系统架构

3 数据源及数据处理

3.1 概述

3.1.1 数据源 专题知识库的数据来源主要有以下几种：一是文献资料，包括政府研究部门、企业、组织、智库、学术机构所发表的文件资料、学术论文等；二是新闻媒体相关报道，包括报纸、杂志、电视、广播等综合新闻内容；三是社交媒体等网络资源，包括博客、微博、抖音、微信、推特、脸书等。对于通过不同渠道收集的半结构化和非结构化等多源异构数据，应按专题进行过滤、去重、筛选分类，并对分类的数据进行甄别，确保来源、内容可靠可信[6]。本系统支持各类异构资源的整合与检索，采用都柏林核心元数据集（Dublin core element set，DC）进行统一著录，DC 元数据集具有简易性、类型通用性、国际通用性、可扩展性等特点。

3.1.2 文献数据解析工具 采用 JavaFX 技术，界面和控制逻辑分离，方便程序代码和界面代码协作开发；有更丰富的组件，方便快速开发。另外，文献数据处理时不采用传统的期次等组织形式，平台数据管理人员可以根据实际需求有针对性地采集并将文献数据组织到对应专题下。

3.2 数据采集

本系统数据解析模块目前支持 TXT、RIS 和 Excel 3 种格式，以人工采集的方式搜集获取专题数据。如采集来源无法下载或者下载格式非平台支持的格式则需要人工处理，即按照资源类型对应的数据模板编辑数据。系统初始化资源类型有：期刊（文摘、全文）、专利、标准、临床试验。其他资源类型可由系统管理员在管理平台自定义创建。

3.3 数据导入

3.3.1 文献数据导入 涉及数据解析、异构数据整合、数据合并去重等处理过程。在数据解析过程中，设计了 TXT、RIS 和 Excel 3 种数据格式的解析器，分别负责读取并解析出不同的数据字段；由于

来源数据的多样性和异构性等问题，针对数据字段不同但涵义相同的字段需要进行合并导入，如期刊中的作者字段、专利中的发明人、标准中的制定人等；对于数据标识问题，以文献数字对象唯一标识符（digital object unique identifier，DOI）、专利号、标准号、试验注册号作为数据唯一性判断，如采集数据无对应数据标识，则利用资源类型和标题字段作为唯一性判断。

3.3.2 全文数据导入 涉及数据处理、全文索引等处理过程。设计 3 种不同上传方式：一是以全文文件名称进行标题匹配；二是以人工编辑的辅助文件标引全文文件；三是在管理平台系统中，依托文献列表进行单篇上传。采取多种上传方式可方便数据维护，提高工作效率。

3.4 索引制作

文献资源检索系统是本系统提供信息服务的核心部分，只有能够提供多维度、关键字段的组合检索，才能满足用户的实际需求。检索系统围绕“标题”“作者”“来源”“类型”“专题”“年代”等核心字段进行设计。在文献数据导入后，系统会按照每次导入行为自动生成批次号；全文索引则以批次号作为制作依据，一次性将制作批次对应的文献数据导入到全文索引库中[7]。

4 系统功能结构及检索接口设计

4.1 系统功能结构设计

在系统设计过程中，要与专题知识库的使用对象进行沟通，并充分理解其提出的专业化和个性化需求，在此基础上对各功能模块进行合理设计，这是关系到平台可用性和易用性的关键环节。在本系统中，主要从业务需求、系统平台、数据流转、用户角色[8]4 个方面对各功能流程进行规划和设计，见图 2。

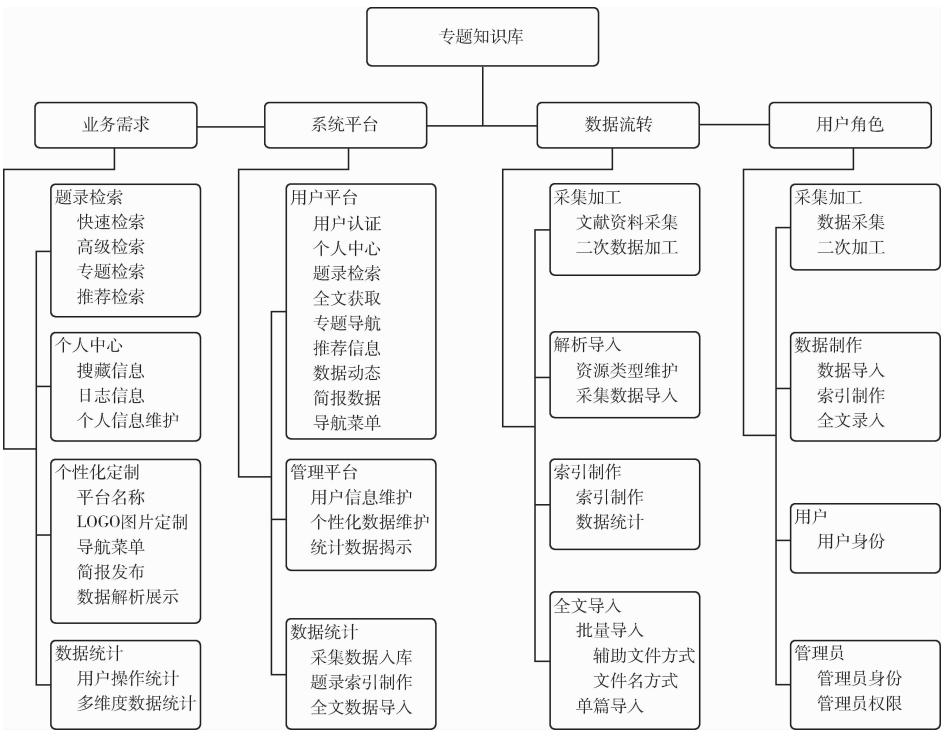

图 2 专题知识库系统功能结构

4.2 检索接口设计

4.2.1 检索信息输入、输出 文献检索接口依照检索服务提供的超文本传输协议（hyper text transfer protocol，HTTP）接口进行封装，将用户检索输入转换成对应的参数提交并返回检索结果。Solr 检索条件为机器语言，需要向用户展示自然语言表达式；在提交检索接口之前，需要进行自然语言表达式和 solr 检索表达式的转换操作，见表 1、表 2。

表 1 检索信息输入

序号	输入项	类型	参数名称	是否可空	说明
1	检索表达式	文本	exp	N	自然语言表达式
2	分页页码	数字	page	N	–
3	分页间距	数字	size	N	–
4	排序规则	数字	sort	N	排序规则：时间倒序（默认排序）、时间正序、相关度排序
5	限定条件	文本	filterJson	N	json 结构数据

表 2 检索信息输出

序号	输出字段	参数名称	说明
1	检索表达式	exp	solr 表达式
2	检索命中页数	total	–
3	检索命中记录数	count	–
4	当前页码	page	–
5	聚合统计	groups	类型、年份、数据库、来源、关键词、语种等统计信息
6	检索记录	records	文献信息

4.2.2 文献检索接口关联的数据对象和方法类 其中在 SearchCompileUtils 类中，根据正则表达式解析用户输入的检索表达式，构造 solr 全文检索服务规定的检索表达式，有助于多条件检索的实现。其中，文献检索实体对象类 SearchLiterarature Vo 的属性中存储了页面传入的检索条件；文献检索接口服务类 SolrSearchService 构建了 3 种检索方法：快速检索、高级检索、专题检索。过滤器类 SearchIntelligenceFilterReqDto 对检索条件进行过滤，并将过滤后的检索条件传入高级检索类 SearchIntelligenceReqDto，再传入高级检索方法中完成检索。在多条件检索类 SearchCompileUtils 中，根据正则表达式解析用户输入的检索表达式，构造 solr 全文检索服务规定的检索表达式，实现多条件检索。在传输响应类 HttpUtils 中对客户端检索参数、检索时间、检索过期时间等进行存储。

5 系统开发及关键技术实现

5.1 系统开发及运行环境

5.1.1 开发应用技术 采用 J2EE（Java2 platform enterprise edition）相关技术进行开发，J2EE 定义了丰富的技术标准、开发工具和应用程序接口（application programming interface，API）为开发企业级应用提供技术支持，这些技术涵盖数据库访问、分布式通信和安全等。大型企业级 Web 应用系统的开发通常要求良好的便于协作开发和扩展升级的软件架构，传统的开发模式不能很好地满足这些需求。因此本系统采用目前较流行且稳定的开源框架 SpringBoot，该框架提供了一种开发 J2EE 企业级 Web 应用的解决方案。

5.1.2 运行环境 将 Apache Tomcat 作为应用服务器，后端数据库采用 SQLServer 2008 作为数据库管理系统（database management system，DBMS），全文搜索服务器采用 solr[9]。操作系统：Windows 2008 及其以上版本系统或者 Linux 内核 2.6 及其以上版本系统。软件环境：关系型数据库 SQLServer 2008，Web 应用服务器 Apache Tomcat 8，Java 运行环境 JDK1.8。硬件环境：处理器主频 2.5GHz 以上，物理内存 16GB 以上，系统盘空间 40GB 以上，数据存储盘空间 200GB 以上。

5.2 关键技术及实现

5.2.1 文献数据推荐实现 平台系统每周以用户查询、下载、阅读等多维度的日志信息作为依据，统计当前访问量最多的文献数据，推荐到用户端首页。以最新发布数据作为补充，保证推荐资料的完整性。例如当前统计后的推荐结果只有 2 条文献数据，系统将在最新录入的数据中选取 4 条文献数据作为补充推荐；若无统计结果则选取 6 条文献数据作为补充推荐。

5.2.2 多条件检索实现 用户可以通过直接点击专题选项浏览对应的文献数据或在高级检索页面输入框中输入多个检索条件组合查询所需文献数据。查询语句在本系统中存在自然语言和 solr 机器语言的区分。

6 专题库平台数据制作流程

6.1 概述

6.1.1 专题库组成 本专题库主要由前台与后台两部分页面以及配套的数据制作工具组成。前台主要为用户提供浏览、检索及下载文献的功能；后台及数据制作工具为管理员提供各类用户及数据管理功能。数据管理是本系统技术实现中的主要难点和关键点。

6.1.2 数据制作 由数据导入、索引制作和全文上传 3 个功能组成。其中数据导入是将采集到的题录信息导入数据库中；索引制作是将数据库中的题录信息导入 solr 索引中，供用户检索；全文上传是批量上传全文数据，从而为用户提供全文下载。

6.2 数据制作流程

6.2.1 数据导入 管理员应提前准备好数据，例如从“Elsevier Science Direct”下载 RIS 格式的题录数据后，即可打开工具点击“数据导入”，会出现主题选择窗口，可选择相应的主题名称，以便将本次导入数据存储到该主题下。此处显示的主题可由管理员在后台页面自由设定，例如可设定各种正在进行研究的病毒、细菌名称等。下一步则需要选择待导入数据对应的资源类型和来源，本例中下载的“Elsevier Science Direct”数据需要选择“全文”。点击“选择文件”，在弹出窗口中找到待导入的数据文件进行导入即可。

6.2.2 索引制作 数据导入完成后，用户端仍然无法检索到导入的数据，需要将数据制作到 solr 索引中用户才可以检索到题录信息。点击工具中的“索引制作”，只需找到之前状态为“未制作”的相应批次信息进行导入即可。

6.2.3 全文上传 索引制作完成后，用户端可以正常检索到题录信息，但还无法下载到对应的全文数据，需要将题录信息对应的全文数据上传之后，用户才可以正常下载全文数据。全文上传有两种录入方式，一是以“辅助文件”的方式批量上传，二是以“文件名称”匹配的方式批量上传。其中“文件名称匹配”复选框选中则代表以“文件名称”匹配的方式批量上传，取消选中则代表以“辅助文件”的方式批量上传。

对于系统中尚无全文的数据，系统会在后台界面罗列在数据表格中以便用户进行全文上传，既可单篇上传也可通过导出模板进行批量上传。全文上传成功之后，则该文献整个数据制作过程结束，用户可以在前台页面进行检索、浏览并下载。

7 结语

建立专题知识库是深层次开发利用信息资源的有效手段，医学院校图书馆在结合本校需求、优势和特点，对信息资源进行深度开发的基础上，建设有自身学科特色的专题知识库，才能实现优势互补以及最大程度的信息资源共享。文献数量大、质量高、组织标引科学的专业生物医学专题库必将为该领域的科研活动提供高效的信息支撑。部分科研人员习惯于从 PubMed、中国知网等综合性数据库中搜索所需信息，不仅会遇到有效信息少、检索噪音多等问题，还缺少更有参考价值的专利、标准、法律、指令、手册等文献类型[10]。在搜集上述类型文献数据方面图书馆具有更多的途径和优势。因此，

（下转第 93 页）

6 吴丹. 大学生健康信息素养现状及影响因素研究［D］. 太原：山西大学，2019.
7 罗艺. 大学生信息素养及其教育支持研究［D］. 上海：华东师范大学，2021.
8 尹娜，谢丽. 大数据背景下医护人员信息获取途径及影响因素分析［J］. 中国数字医学，2019，14（8）：39－41.
9 邵思蜜，文漪. 从专业检索的角度谈医生信息素养的提升［J］. 医药与保健，2014（12）：141－142.
10 崔月婷，柴培钰，时小莹，等. 大数据背景下医务工作者医学信息素养与情报服务需求［J］. 中国卫生产业，2020，17（34）：14－18.
11 梁瑞晨，刘延淑，胡敏，等. 四川省三级甲等医院手术室护士信息素养水平调查［J］. 护理学杂志，2021，36（22）：35－38.
12 占艳，晏峻峰，刘青萍，等. 新医科背景下中医药类硕士研究生对信息技能的需求调查与分析［J］. 医学信息学杂志，2021，42（5）：86－89.
13 郑微，李金伟，陈晶. 临床医生信息素养现状调查及影响因素分析［J］. 中华医学图书情报杂志，2021，30（2）：48－52.
14 曹莉，徐翠，曹向阳. 思维导图联合多媒体 PPT 课件在急诊科护士创伤急救知识和技能培训中的应用［J］. 护理实践与研究，2021，18（17）：2660－2663.
15 EVGENIY V T，NATALIA N M，VALERII E S. Information age trends：logical－semantic modelling data visualisation in the educational space［EB/OL］.［2022－06－15］. https：//www.europeanproceedings.com/files/data/article/59/2234/article_59_2234_pdf_100.pdf.
16 FOURIE I. Personal information management（PIM），reference management and mind maps：the way to creative librarians［J］. Library hi tech，2011，29（4）：764－771.
17 韩永青. 高校图书馆学科知识服务可视化研究——学科思维导图绘制［J］. 情报科学，2011，29（8）：1262－1267.
18 杨雪萍. 思维导图在馆际互借与文献传递服务中的应用——以北京师范大学图书馆 CASHL 服务为例［J］. 大学图书馆学报，2015，33（2）：66－71.
19 LIU W. Knowledge map：a creative visual path to library guides and resources［J］. Electronic library，2020，38（5）：943－962.
20 彭迪，常红. 思维导图方法提升高校图书馆员职业能力探析［J］. 图书馆工作与研究，2020（5）：101－105.

（上接第 88 页）

由图书馆依据学校实际需求搭建的“一站式服务”专题库平台能够为医学专业人员提供全面、权威、最新的生物医学文献信息，有助于科研人员节约时间、提高科研效率。后续还将在专题知识库检索结果的多维度可视化分析与揭示方面进一步深入研究和完善，并努力拓展本系统的服务范围，为医学从业人员及时获得各种专业性更强、更加精准的信息服务提供助力。

参考文献

1 冀振武. 试论如何提高高等医学院校图书馆的情报职能［J］. 中国卫生经济，2000，19（3）：57.
2 张蕾，刘薇薇，赵志强. 高等医学院校专题特色数据库的建设与开发［J］. 医学信息学杂志，2008（1）：39－42.
3 张胜全，董佳. 高校图书馆专题特色库资源的深度开发［J］. 现代情报，2005（2）：33－34，95.
4 李蓓，兰小筠. 医学专题数据库的开发［J］. 中华医学图书情报杂志，2005，14（1）：57－59.
5 文忠，阮华，刘媛筠. 专题知识库服务系统构建方案研究［J］. 现代情报，2010，30（2）：104－108.
6 周荣. 论高校图书馆为学科建设服务的措施和模式［J］. 上海高校图书情报工作研究，2006，62（2）：18－20.
7 李宇，张明昊. 特色信息资源揭示发布系统的设计与实现——以东北大学冶金专题文献库为例［J］. 图书馆工作与研究，2019（1）：65－71.
8 李宇. OA 期刊学科服务检索系统的创建与实现［J］. 情报科学，2012（3）：387－390.
9 孙利芳，金焱. IT 项目管理在专题数据库建设中的应用［J］. 农业图书情报学刊，2010，22（12）：170－171，182.
10 李健捷，徐艳芳，朱建平. 专题知识库建设研究［J］. 内蒙古科技与经济，2020，456（14）：76－77.