

国外医药类学科知识库的构建特征与启示*

周阳, 王润海

(中国药科大学图书馆, 南京 210009)

摘要: 调研并总结PMC、TMC、German Medical Science等15个全球知名医药类学科知识库的构建特征, 从资源、支撑和功能层面分别阐述。从中得到一些启示, 包括特有资源数字化、元数据质量控制、异构资源组织标引与检索、子系统建设、学科服务内容、科研数据管理、手机平台建设等方面, 以期为国内医药类学科知识库构建提供参考。

关键词: 医药类; 学科知识库; 知识库构建

中图分类号: G25

DOI: 10.3772/j.issn.1673-2286.2016.4.010

1 引言

随着国内外开放存取 (Open Access) 运动的发展, 学科知识库的建设逐步深入。学术界就“学科知识库”并未有一个官方统一的定义, 笔者综合各种说法总结如下: 学科知识库是以学科分类为前提, 挖掘整合学术资源, 从学科角度开放存取, 为学科知识服务提供资源保障, 有利于某一学科的累积、发展和传承。学科知识库有别于机构知识库和专题资源库: 机构知识库是基于科研机构, 而学科知识库是基于某个或几个学科; 专题资源库更集中于对专题的研究, 例如武汉大学图书馆的“长江资源特色库”和厦门大学图书馆的“东南海疆研究数据库”等, 而学科知识库存储的是某一个或者某几个学科的所有内容, 范围更广。截至2016年4月5日, Open DOAR共收录3047个机构知识库, 其中学科知识库293个, 约占总数的10%, 已具有一定学术影响力。本文主要通过调研全球15个知名医药类学科知识库, 为我国医药类学科知识库构建提供参考。

2 调研对象与结果分析

2.1 样本选择

在Open DOAR收录的医药类学科知识库中, 欧美

所占比例居多, 发展较为完善, 参考价值较高。美国国立卫生院、美国国家生物技术信息中心和国家医学图书馆拥有丰富的医药类学科资源, 因此一直致力于医药类学科知识库的建设, 如U.S. National Library of Medicine Digital Collections、PubChem和PMC的建设就主要依托上述机构。美国的许多高校和科学院如康奈尔大学、德克萨斯大学、布什尔医学科学院等都建设有医药类学科知识库。欧洲也是如此, 德国科学医学学会、德国研究基金会 (DFG)、Department of Health (Ireland)、Public Health Agency (Ireland)、Safefood and Public Health England等都是参与建设学科知识库的机构。因此笔者选取的调研样本, 7个来自美国, 4个来自欧洲; 为使样本更具全球性, 笔者还选取了1个澳大利亚和2个日本的学科知识库以及世界卫生组织的医药类学科知识库。具体国家分布及资源数量见表1, 数据来源于各学科知识库网站。

2.2 调查结果分析

调查结果从资源层面、支撑层面和功能层面来分析。

2.2.1 资源层面

(1) 资源数量。由表1可见, 医药类学科知识库资

* 本研究得到2016年度中国药科大学中央高校基本科研业务费专项资金 (人文社科类-培育项目) “药学类学科资源库构建策略和建设途径研究” (编号: 2016RPY016) 资助。

表1 学科知识库地理分布及资源数量统计表

序号	学科资源库名称	国家/地区	资源数量
1	Digital Collections	美国	17 129
2	PubMed Central	美国	380万
3	Ballarat Health Services	澳大利亚	822
4	DigitalCommons@The Texas Medical Center	美国	1 744
5	German Medical Science	德国	313
6	Institute of Public Health in Ireland	爱尔兰	3 247
7	World Health Organization	联合国	161 602
8	viXra	美国	13 829
9	American College of Healthcare Sciences	美国	512
10	PubChem	美国	2.44亿
11	National Documentation Centre on Drug Use	爱尔兰	14 318
12	Belgian Health Care Knowledge Centre	比利时	279
13	Bushehr University of Medical Sciences Repository	美国	44 782
14	Shiga University of Medical Science Repository	日本	2 906
15	Tokyo Ariake University of Medical and Health Sciences Academic Repository	日本	30

源数量相差较大, 最少的30个, 最多的高达2.44亿。资源数量与建设规模和建设时间成正相关。

(2) 语种分布。这些学科知识库中, 涉及语种最多的是美国国家医学图书馆的免费生物医学书籍和视频在线资源, 共包含103种语言。总的来说, 英语是最普遍的语种; 除此之外, 德语、法语、俄语、中文、阿拉伯语、日语

也较为多见。

(3) 回溯年代。其最久的是Bushehr University of Medical Sciences Repository, 回溯至公元15年。回溯至19世纪的比较多, 例如PMC回溯至1800年, Ballarat Health Services回溯至1875年, Shiga University of Medical Science Repository回溯至1834年。

(4) 资源类型。具体见表2 (学科知识库名称以表1中的序号代替), 资源类型至少20种, 且各有侧重。资源类型数目超过5个的有期刊论文、多媒体资源、数据集、学位论文、会议论文、图书、专业报告、手稿、图片、教学资料。表中的“其他类型”主要包括新闻公告、行政文件、研究项目、宣传材料、讲座展览、实验等。

(5) 学科主题。学科主题划分主要有以下几种情况:

①以Shiga University of Medical Science Repository BIWAKO为代表, 存储的都是生物科学、生物化学与分子生物学、药理学、滋贺医科大学部心血管和呼吸系统的药物、睡眠医学、护理学、药学等完全围绕医药类的学科资源; ②以viXra为代表, 存储包括生物学、化学、人类学、数学、物理学、计算机科学等生物医学相关又包含其他学科的学科资源; ③以PubChem为代表, 包含生物、药物、化学和毒理学等专深度较高、供专业研究人员使用的学科资源。

2.2.2 支撑层面

(1) 系统软件。学科知识库构建系统最常用的是Eprints、Fedora、GreenStone和DSpace。Eprint由英

表2 学科知识库的资源类型

学科资源库 资源类型	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
期刊论文		√	√	√	√	√	√	√	√		√		√	√	√
多媒体资源	√		√					√	√		√	√	√		√
参考书目							√		√		√				
预印本								√							
数据集		√		√		√		√		√			√	√	
学位论文			√	√					√		√		√	√	√
会议论文			√	√	√	√	√	√	√		√		√	√	√
专利								√		√			√		
图书或章节	√	√	√			√	√	√	√		√		√		
软件		√				√		√		√					

续表

学科资源库 资源类型	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
专业报告	√	√	√		√	√	√	√			√	√	√	√	√
历史档案	√	√	√												√
手稿	√	√	√		√	√		√					√		√
地图/图片	√	√	√			√		√		√			√		√
幻灯片			√							√		√			
信件			√												
教学资料				√				√	√	√	√		√		√
手册					√										
综述			√					√			√				
其他类型				√		√	√		√	√	√	√	√		√

国南安普顿大学研发,最新版本是v3.3.15。EPrints 3 提供各种开源平台、定制和托管解决方案, Bushehr University of Medical Sciences Repository使用的就是EPrints 3.3.12。Digital Collection则是使用Fedora Commons Repository Software提供一个基于XML框架的资源库平台, Fedora还包括Solr/Lucene索引应用, 更有助于元数据搜索; 图像转换使用Djatoka JPE2000 图像服务器, 该软件由美国Los Alamos National Laboratory设计; 视频播放则使用NLM Office of Computer and Communications Systems开发的视频播放器, 该播放器采用结构化的标题文件转换成文本搜索, 根据访问需求以视频时间线为轴搜索图像结果。German Medical Science使用DIMDI开发的GMS永久性自由可访问文献平台, 每个文档都会有一个唯一标识符(URN-Uniform Resource Name and DOI-Digital Object Identifier)以确保文章的永久性链接。Belgian Health Care Knowledge Centre默认Firefox Brower浏览器, 某些功能实现通过Javascript、HTML代码被提交给W3C验证工具, 所有PDF文件都需要Adobe Reader 7以上版本。PebChem由于还有很多化学结构式需要检索, 因此资源库还提供结构搜索、结构聚类、基于网络的3D浏览、Pc3D查看应用程序等工具。

(2) 存取政策。对于资源使用者, PMC使用NLM的规定: “公共领域的信息可以自由地传播和复制, 但是任何的后续使用必须经过NLM的许可和确

认。在使用任何涉及个人、公司、组织的作品, 例如文件、插图、照片等, 这些受到美国或他国版权法保护的东西, 都必须经过著作权人的书面许可。历史档案文件的版权有可能无法确定, 因此资源库会给予用户恰当的使用权限。” German Medical Science也有类似的使用许可Creative Commons Attribution 4.0 International License。Belgian Health Care Knowledge Centre (KCE) 使用“by/nc/nd” Creative Commons License: “学术成果不得用于商业及生产衍生产品, 在引用和分享时, 也必须提供原文出处链接, 对于作者文档的任何处理都必须有作者授权。” 而Institute of Public Health in Ireland则将存取政策细致化, 元数据、内容、提交、学科、版权保护、版本使用等都有详细的政策规定。对于版权保护Vixra将为提交者提供一个时间戳, 用于证明作者学术成果发表的时间, 在遭遇学术侵权的时候提供佐证。

对于资源上传者, PMC规定: “美国NIH公共获取政策规定凡是受到NIH资助的学者, 在学术成果发表后的12个月内必须提交一份电子版去PMC供开放获取”, 同样使用PMC存取政策的还有TMC等。

2.2.3 功能层面

功能层面主要体现的是学科服务功能, 核心服务包括资源上传、检索、浏览、下载、推送和知识关联。一些学科知识库利用表单提问的方式与用户交

流,如“leave a message”和“Feedback”功能,或在facebook、twitter、YouTube、论坛等社交网站开设账号,供用户们联络讨论,还有的配有用户点击下载情况的统计图表和镜像服务等。

拓展服务一般都需要用户注册自己的账号来完成,主要有个性化E-mail订阅和RSS订阅服务;保存检索历史;个人知识名片,包括基本信息、学术简历、研究项目、知识产出、个人主页等;绘制知识地图;书目和参考文献管理工具;开放存取期刊数字出版。

3 医药类学科知识库构建特征

由以上调查结果,笔者归纳总结出一些医药类学科知识库的构建特征,也从资源层面、支撑层面和功能层面来阐述。

3.1 资源层面

3.1.1 资源数量

资源数量越多,学科知识库的价值就越大,但构建学科知识库不能只追求数量,更要兼顾质量。由表1可见这些学科知识库的资源数量,各个量级都有,它们都能够被Open DOAR收录,发挥其学科价值。因此,在构建医药类学科知识库时,不需在意初始资源的数量,日积月累是必经过程,收集、整合高质量的学术成果,才是构建学科知识库的初衷。

3.1.2 语种和回溯年代设置

为便于国际交流,学科知识库的语种可尽量多设置,使页面可转换各语种阅读。但调研中发现,有的页面语种转换时翻译并不精准,容易误导用户。科研人员上传数据资源时,也应当有多个语言版本可选择,并且文档、视频等都应带有翻译功能,这样才更有利于资源内容的广泛传播。学科资源的回溯年代越久远,越有利于明晰学科发展脉络,对于学科发展历史的研究大有裨益。

3.1.3 资源类型与归档呈缴

大部分学科知识库构建都是先收集常见的学术资源,满足基本需要;发展到一定阶段之后,再逐步根据自

身条件和学科特色进行建设,例如PubChem里收录的多是各种化合物、生物测定、生物活性等专业性极强的特色资源。资源收集过程中,归档呈缴是有效方法之一,如机构内科研人员的公开发表资源,在学科知识库建设初期,先由工作人员主动搜集;随着学科知识库宣传推广的深入,逐步实现由专人收集和自行提交审核相结合的方式存缴;对于未公开发表的学科资源,学科知识库以任务分派的形式,派出专人负责具体搜集工作,主动与科研工作者联系,鼓励其将学术成果纳入学科知识库^[1]。

3.1.4 知识库构建方向

医药类学科知识库的构建方向主要有三种。第一种是几乎都和医药类有关,但并不专深,带有一定科普性,专业人员和普通用户都可以使用。第二种是包含医学、药学、化学、数学、物理学、生命科学、计算机科学等多学科的学科知识库,适合进行学科交叉研究。学科交叉是广为应用的科研方法,以药物化学为例,科研合作强化了药物化学与药学、分子生物学、生物化学、有机化学等学科之间的知识流动,不仅丰富了学科知识体系,还能激发和启迪学者的开拓性思维,促进科学创新^[2]。第三种是医药类较为专深、供科研人员使用的学术资源,这部分资源使用范围窄,但科研价值大,是作为学科尖端研究的坚实基础。可见,医药类学科知识库应根据实际情况和自身需求选择不同的构建方向。

3.2 支撑层面

3.2.1 以学科馆员为主体的人员支撑

学科馆员是学科知识库构建中的重要一环。学科馆员1950年在美国首先试行,医药类学科知识库的学科馆员既需要有图书情报专业的技能,又需要拥有医药学领域的扎实知识,才能在医药领域开展具有针对性的情报服务工作,例如引导用户上传、检索学科资源、举办学科知识库使用培训、制定或者参与制定学科知识库的存取政策、学科知识库学术成果的质量控制等。学科馆员的服务模式大致可以分为“学科馆员—信息用户”服务模式、“学科馆员—图情教授”协作模式、“学科分馆—学科馆员”组合模式、“学科馆员—学术秘书”兼职模式等^[3]。

3.2.2 解决知识产权问题的细则支撑

学科服务不可避免地存在知识产权问题。在学科知识库构建阶段,需要考虑使用软件是否取得版权,信息源版权是否明晰,是否与科研人员签订版权协议,是否收录科研人员已发表论文的后印本,是否采用技术保护措施防止用户过量下载及限制用户访问权限,转载、摘编、深度链接是否侵犯著者的复制权和传播权等。

在学科服务阶段,需要考虑是否保护了科研人员个人隐私,是否获得检索授权,文献传递数量和形式是否符合著作权和数据库合同规定,是否造成知识产权的流失和外溢,独创性部分是否受法律保护等。

另外,学科馆员对知识产权风险的认知水平、道德素养和工作能力是否达到要求,用户是否把获取的学科信息另作他用或向大量不特定公众传播等,这些细化的知识产权问题急需学科知识库构建者制定相关细则并加强监管。

3.3 功能层面

学科服务功能可向细致和人性化方向发展。例如学科知识推荐,首先应建立用户兴趣库,信息来源一是用户注册时填写的兴趣,二是系统通过用户历史记录等行为分析挖掘捕获的用户偏好;其次根据获取的用户兴趣对信息过滤,把用户感兴趣的信息推送给用户;最后根据用户的反馈对其推送信息进行改进和完善。信息推送就是把适当的学科知识在适当的时间传递给适当的人^[4]。

4 启示与建议

学科知识库最主要的关注点就是建设与使用。在这两个关注点上,依然从资源层面、支撑层面和功能层面进行论述。

4.1 资源层面

4.1.1 数字化特有医药类资源,提升资源可见性

医药学科有很多图片资源如动植物标本图、解剖切面图、珍稀动物药、植物药、矿物药的图片资料,还有很多善本医药类图书、刊物、手稿、档案,这些特有的专业资源是大学和科研机构的重要资产。但这些类型的资

源本体并不易保存,还有一些是记录在穿孔卡片、磁带、软盘、CD、DVD等容易出错的介质上。如果这些资源被数字化,存储进学科知识库,更有利于知识的共享传承。

目前关于数字化特有资源所需的时间、人力和工具,一些案例可供参考: The Walters Art Museum处理较小的书籍每天平均可扫描275张,较大的书籍每天约150张,使用Omar成像装置;明尼苏达大学使用i2S CopiBook HD和两个荧光灯,每天可完成1 536份扫描;加州大学伯克利分校Bancroft图书馆使用两个NextScan Eclipse扫描仪,每天可扫描4 800张图像;美国国会图书馆使用的设备包括数码相机、投射照明、安全材料的台面、薄膜平带、封闭的空间和空气过滤器,由三名工作人员每天完成500张图像数字化^[5]。

4.1.2 提高元数据加工质量,确保资源可用性

电子资源元数据质量控制,可借鉴马里兰大学学院公园分校的例子,他们使用的是OCLC的WorldShare管理器,允许用户使用MARC21编辑器或文本视图编辑器在WorldCat中创建新条目并扩充现有条目,同时针对实物和电子资料高效地进行一次一条记录的元数据管理作业,还可以利用该管理器设置或删除资源库的书目记录并导出,搜索美国国会图书馆规范记录,创建并打印书脊标签,通过IP认证还可以链接至RDA工具包。

除了WorldShare以外,Connexion也是元数据管理器,它有基于Web的浏览器界面,也有Windows桌面应用程序,可以进行元数据编目,也可以检索WorldCat中是否存在记录可直接导出。Connexion提供90多个索引,涵盖阿拉伯文、中文、日文、韩文、泰文等15种语言,并支持使用衍生、数值型和短语选项在这些索引中灵活检索。下拉菜单会列出可用索引,点击书目记录中的检索标题,可将标题链接至相关规范记录,这样规范记录一旦有所更改,书目记录就会自动更新。

4.2 支撑层面

4.2.1 异构资源的组织、标引和检索

要解决异构资源的组织、标引和检索,将搜索引擎和学科知识库相结合不失为一个好方法。日本的高性能全文搜索引擎Hyper Estraier具备数据抓取、数据解析、创建索引和查询各种功能。学科知识库中的资源多

是pdf、doc、ppt、html、jpg、zip等格式,该搜索引擎可解析各种文档,采用GNU宽通用公共许可证作为开源协议,可在极短时间内为大量文档创建索引并得到检索位置,搜索核心采用N-gram方式,不依靠词汇也能进行高精度检索。Hyper Estraier的点对点架构,利用节点服务器,提供各索引地址,用户可以通过任一节点进行查询,还可以将结果按照某个权重排序返回给用户。功能上该搜索引擎支持短语搜索、正则表达式搜索、属性搜索和类似搜索等,还提供Java、Ruby、Perl语言的上层API,可根据需要进行扩展,自带网络爬虫,可自动抓取网络最新资源并追踪更新^[6]。

4.2.2 四大子系统有机整合

无论哪一种开源软件,学科知识库的框架支撑都需要有四个子系统构成^[7]:

(1) 数据采集和提交子系统:提供数据采集途径,包括自主提交、批量导入和自动抓取。自主提交是为科研人员提交科研成果提供接口;批量导入是为管理人员按照设定格式将数据批量导入系统数据库;自动抓取是从网络发现并收割元数据,对数据进行收集、选择、清洗、构建、集成、格式化,形成本地化收藏。

(2) 数据存储子系统:完成元数据仓储功能,对采集而来的各类型数字资源长期保存和管理。

(3) 资源发布与利用子系统:以存储的数字化资源为基础提供学科服务。

(4) 系统管理子系统:为整个学科知识库运行和管理提供基础性的辅助功能,如用户权限管理、元数据管理、分类词表等。

4.2.3 发挥学科专家优势,提供专业化科研支撑

为科研工作者提供立项查新、实时咨询、数据库利用指导、资料搜索、论文撰写和投稿指南、高被引和热点论文分析、SCI/EI/ESI/CSSCI/CSCD目录更新等服务,特别是帮助科研人员查找难以搜索的科研资料。生物医药类学科专业性强,例如有毒中草药、药用水生生物、药用藻类研究这些资料必须到非常专业的图书馆或科研机构查找,学科专家可为科研人员提供专门的检索服务,或者通过馆际互借、电子文档共享等渠道提供服务。虚拟咨询管理系统目前QuestionPoint较好,它集成了实时咨询、电子邮件、咨询知识库、报告和分析

等工具,具有全天候合作咨询机制,可提供实时帮助。

对学科立项发文做统计分析,为各级项目审批提供参考也是学科服务的好途径。中国药科大学已初步建成医药类学科发展支撑平台,该平台的生物医药产业数据库,内含中药方剂研究报告、新品种分析报告等专业报告,同时也有本校的学科评价、学科SCI论文分析报告、学科评估报告、非潜力学科论文发文及发文期刊分析报告、ESI学科发文期刊分析等学科专业分析报告。有较强的学科支撑能力,学科知识库的建设才能获取更多支持。

4.3 功能层面

4.3.1 科研数据管理功能

学科知识库可以给学者们提供科研数据的保存和保密服务。由于科研数据可能还不适合公开共享,因此数据的存储和提取都应当有较为严格的管控。科研数据也是机构的宝贵资产,很多国家都开始设专项基金来确保科研数据的保存和管理,例如加拿大联邦基金正和加拿大研究图书馆协会(Canadian Association of Research Libraries, CARL)筹建覆盖所有境内科研机构的科研数据存储设施网络^[8]。

4.3.2 多元资源访问功能

学科知识库一般通过Blog、RSS、Wiki、Ajax、IM、E-mail等途径使用户参与学科资源建设,与用户交流。随着手机阅读、平板阅读的逐步发展,移动终端阅读软件的研发也迫在眉睫。将电脑和手机平台构建相结合,开发微博、微信、二维码标识功能,利用手机对资源进行浏览、下载和上传,实现学科知识库的多途径访问。2013年PMC推出了PubReader软件,有利于读者用平板等小屏幕设备阅读PMC文章,同时PubReader也同样适用于台式机和笔记本电脑,这正为我们的研究指明了方向。

参考文献

- [1] 侯艳,王晓,卢玉红,等.学科特色视域下中医药院校机构知识库的构建[J].中华医学图书情报杂志,2016,25(1):61-63,69.
- [2] 徐晓艺,杨立英.科研合作视角下的学科知识流动分析方法研究:以药物化学学科为例[J].图书情报工作,2014,58(19):83-91.

- [3] 严栋.高校图书馆复合型学科化服务模式研究[J].图书馆学刊,2014(12):62-64. (应用版),2010(2):28-31.
- [4] 刘畅,孙越.基于本体的学科知识管理系统框架研究[J].信阳师范学院学报(自然科学版),2010,23(4):618-620.
- [5] Erawy R. Rapid Capture: Faster Throughput in Digitization of Special Collections [R/OL]. [2016-04-05]. <http://www.oclc.org/research/publications/library/2007/2007-02.pdf>.
- [6] 潘鹏程.Estrailer H在学科知识库系统中的应用与实现[J].图书馆学研究
- [7] 熊惠霖.基于DSpace的图书情报领域学科库整合平台的实现[J].情报探索,2014(12):55-59.
- [8] Erway R, Rinehart A .. If You Build It, Will They Fund? Making Research Data Management Sustainable [R/OL]. [2016-04-05]. <http://www.oclc.org/content/dam/research/publications/2016/oclcresearch-making-research-data-management-sustainable-2016.pdf>.

作者简介

周阳,女,1985年生,硕士,中国药科大学图书馆馆员,研究方向:开放存取、机构知识库、参考咨询,E-mail:zhouyang1119@sohu.com。
王润海,男,1981年生,中国药科大学图书馆馆员,研究方向:参考咨询。

The Construction Characteristics and Enlightenment of Foreign Medicine and Pharmacy Discipline Repositories

ZHOU Yang,WANG RunHai
(Library of China Pharmaceutical University, Nanjing 210009, China)

Abstract: This paper investigates and summarizes the construction characteristics of 15 world famous medicine and pharmacy repositories such as PMC, TMC, German Medical Science and so on, respectively from the resource level, support level and functional level. The author sums up some suggestions from these discipline repositories research, such as unique resources digitization, metadata quality control, indexing and retrieval of heterogeneous resources organization, subsystem construction, content of subject service, scientific data management, mobile phone platform construction, etc., in order to provide reference for the construction of domestic discipline repositories.

Keywords: Medicine and Pharmacy; Discipline Repository; Repository Construction

(收稿日期: 2016-04-05)

■ 书 讯 ■

《汉语主题词表》(工程技术卷)

《汉语主题词表》自1980年问世以后,经1991年进行自然科学版修订,在我国图书情报界发挥了应有的作用,曾经获得了国家科学技术进步二等奖。为了适应网络环境下知识组织与数据处理的需要,2009年由中国科学技术信息研究所主持,并联合全国图书情报界相关机构,完成《汉语主题词表(工程技术卷)》的重新编制工作。

全书共收录优选词19.6万条,非优选词16.4万条,等同率0.84。在体系结构、词汇术语、词间关系等方面进行改进创新。为了方便工程技术领域不同专业用户使用,《汉语主题词表》(工程技术卷)按专业分13个分册出版,同时建立《汉语主题词表》服务系统,提供在线概念检索和辅助标引服务,通过可视化技术展示各类概念关系,是图书馆、档案馆、出版社、期刊杂志社、文献信息中心等专业工作者及科研、教育及工程技术领域人员必备的参考书。

《汉语主题词表(工程技术卷)》已于2014年由科学技术文献出版社出版,全书2 300余万字,总定价3 880元,可分册购买。