

机构知识库建设实践研究^{*}

王序文

李军莲

(中国医学科学院医学信息研究所 北京 100020)

(1 中国医学科学院医学信息研究所 北京 100020

2 中国科学院文献情报中心 北京 100190

3 中国科学院大学 北京 100149)

黄利辉 王 蕾 夏光辉 吴英杰

(中国医学科学院医学信息研究所 北京 100020)

〔摘要〕 对中国医学科学院/北京协和医学院机构知识库建设理念及阶段性成果进行介绍,通过开展机构知识库建设实践,完成具有医学领域特色的机构成果库以及满足基本功能需求的机构知识门户搭建,就相关支撑机制及未来工作进行讨论。

〔关键词〕 机构知识库;医学成果汇聚;机构知识管理;CAMS-IR

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2018.07.011

Study on Building Practice of Institutional Repository WANG Xu-wen, 1Institute of Medical Information, CAMS & PUMC, Beijing 100020, China; LI Jun-lian, 1Institute of Medical Information, CAMS & PUMC, Beijing 100020, China; 2National Science Library, Chinese Academy of Sciences, Beijing 100190, China; 3University of Chinese Academy of Sciences, Beijing 100149, China; HUANG Li-hui, WANG Lei, XIA Guang-hui, WU Ying-jie, Institute of Medical Information/Medical Library, CAMS & PUMC, Beijing 100020, China

〔Abstract〕 The paper introduces the building philosophy and periodical results of the Institutional Repository (IR) of Chinese Academy of Medical Sciences/Peking Union Medical College. Through developing building practice of institutional repository, the architecture of the institutional achievement database with features of the medical domain and the building of institutional knowledge portal meeting basic functional requirements is completed, and discusses related supporting mechanism and future work.

〔Keywords〕 Institutional Repository (IR); Medical academic achievement convergence; Institutional knowledge management; CAMS-IR

〔修回日期〕 2018-03-18

〔作者简介〕 王序文,博士,助理研究员,发表论文10余篇;通讯作者:李军莲,硕士,副研究馆员。

〔基金项目〕 中国医学科学院医学与健康科技创新工程项目“生物医学科技信息支撑平台”(项目编号:2016-I2M-2-005)。

1 引言

机构知识库(Institutional Repositories, IR)是对特定范围内的知识资源进行搜集、组织、数字存储、管理并将其中绝大部分资源对网络用户免费共

享的知识库,是机构管理科研成果、传播学术知识、支持全社会创新的重要工具^[1]。自 2002 年 DSpace@MIT 系统问世,全球掀起机构知识库的建设热潮,世界范围内的机构知识库总量呈现逐年增长趋势,截至 2017 年 11 月 30 日开放获取知识库名录(The Direct of Open Access Repositories, DOAR)^[2]收录的机构知识库达到 3 448 个;而由英国南安普顿大学创办的开放获取知识库注册系统(Registry of Open Access Repositories, ROAR)^[3]收录机构知识库 4 585 个。目前国内外已先后成立多个机构知识库联盟,其目标是构建链接全球开放获取知识库的知识基础设施。我国机构知识库建设虽然起步较晚,但是发展迅速,国内重要机构知识库联盟包括中国科学院机构知识库网格(CAS IR GRID)^[4-6]和高校机构知识库联盟(CALIS)^[7-8]。2016 年 9 月 21 日在第四届中国机构知识库学术研讨会上成立的中国高校机构知识库联盟(CHAIR Road Map)对我国机构知识库联盟的发展具有重要意义。

当前关于机构知识库的理论研究主要集中在机构知识库的基本概念、与开放获取的关系、对传统学术交流的影响等方面;关于机构知识库的实践研究主要体现在不同学术机构根据自身需求和目标,构建本机构的知识库,实现机构自有成果收集、保存、管理和利用并向成员提供服务^[9]。随着信息技术的迅速发展及用户需求的不断提高,机构知识库不再单一作为机构保存和展示学术成果的平台,而是围绕用户的多样化需求提供知识服务的资源平台 and 交流平台,其发展趋势包括功能丰富化、组织联盟化、存储云端化、服务知识化等^[10-11]。

2 建设理念

作为具有百年建校史的国家级医学研究中心,中国医学科学院/北京协和医学院实行院校合一的管理体制(简称院校),设有 18 个研究所(以及 5 个分所)、7 所临床医院、6 所学院及 1 个研究生院。面对构成复杂的机构体系,为有效汇聚、展示、共享院校学术成果和知识资产,搭建为各所院成员服务的科研成果交流展示平台,院校秉持可持

续发展的理念,开展机构知识库建设实践,其关键在于设计适合院校、所院两级的机构特点且面向医学领域的架构模式。目前院校机构知识库(Institutional Repository of Chinese Academy of Medical Sciences & Peking Union Medical College, CAMS-IR)的建设任务已纳入中国医学科学院医学与健康科技创新工程项目,得到院校的长期稳定支持。其建设目标是:开发建立院校、所院两级的知识资产汇聚管理平台,汇聚、管理、共享、利用院校科研知识资产,全面展示院校学术成果;提供面向学科团队、所院机构、院校整体不同层面的学术成果展示交流、共享和保存的学术成果知识库,提升院校级、所院级、学科团队以及科学家个人的学术可见度和学术影响力。CAMS-IR 的发展规划如下:采用院校、所院两级的一体化架构模式,分阶段对医科院及其分支机构所产出的学术成果进行汇聚,由中国医学科学院医学信息研究所集中管理,下属各分支机构具有本单位成果库的独立展示页面及其成果管理权限。第 1 个任务阶段,完成满足基本功能需求的 CAMS-IR 机构知识库建设,实现院校机构库的从无到有。在 CAMS-IR 平台上线且稳定运行一段时间后,逐步推广至二级共建单位进行个性化定制、展示及示范应用,根据各单位特色数据及需求反馈,不断拓展服务和功能;分阶段在京内主要所院进行应用推广并逐步实现在整个院校的全面应用。

3 建设实践

3.1 概述

机构知识库建设的重点包括成果内容建设、门户平台建设以及相关政策的制定。CAMS-IR 的建设结合医科院自身机构特点及医学领域特色,设计统一规划、集中建设、整体揭示、分散管理的总体建设方案,围绕机构成果的内容规划建设、平台设计开发以及相关支撑机制等层面逐步开展研究和实践,主要内容包括:(1)建立院校、所院两级的机构成果库。对院校产出的多种类型科研知识资产进行采集、汇聚、加工、保存、管理及组织。(2)搭

建医科院机构知识门户。提供科学家个人、学术团队、学科发展、院所机构等不同层面的信息集成展示、知识交流、开放共享及知识服务。

3.2 CAMS – IR 成果库构建

3.2.1 概述 主要包括机构资源规划、成果内容处理以及院校学者库建设等 3 个方面的工作，其中实现全面、准确的成果典藏是构建成果库的基本原则，因此对成果数据收集回溯和质量控制是 CAMS – IR 内容建设的重点。

3.2.2 机构成果资源规划 在建设初期针对院校的机构成果资源情况进行详细调研、评估和规划，包括典藏资源的覆盖范围、资源类型、资源描述的规范等。CAMS – IR 所覆盖的知识内容主要指院校成员及其合作者所创作的知识内容以及机构出版

物。典藏资源类型涵盖公开成果和内部成果，其中公开成果包括正式出版的作者署名机构为中国医学科学院及其分支机构的期刊论文、会议论文、学位论文、文集、著作、专利、报纸等；内部成果包括教学相关的教学大纲、讲义、课件、视频音频、讲座报告工作文档等，以及科研相关的预印本、科研笔记、科研实验数据、研究报告等科研过程成果。预计在任务执行的前 3 年内实现包括期刊论文、会议论文、学术论文、专著、专利、报纸等类型资源的全面典藏，预计 5 年内支持实现其他数据类型的成果典藏。所有呈缴的内容拥有相对完整、规范的元数据，以便对资源进行详细标引。

3.2.3 成果内容处理 成果库构建任务的重中之重，涉及成果搜集、成果数据处理、成果质量控制等多个环节，CAMS – IR 成果内容处理流程，见图 1。

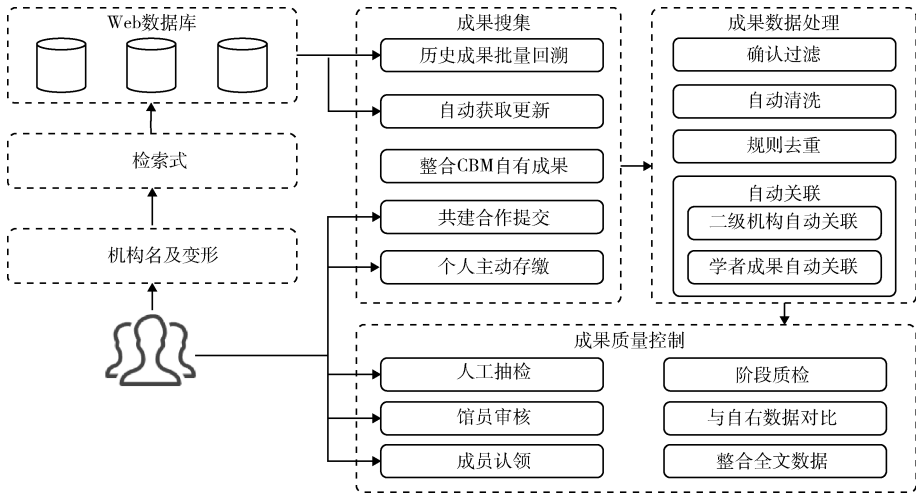


图 1 CAMS – IR 成果内容处理流程

(1) 成果数据搜集。搜集方式包括历史学术成果的批量回溯、定期自动获取及更新、与自有生物医学文献数据库（CBM）^[12] 题录数据的有效整合、通过共建单位合作提交、成员个人主动存缴等。为全面回溯院校历史成果，在初步搜集阶段根据机构名构造多个检索词对院校建校以来公开发表的 5 类学术成果，包括期刊论文、会议论文、学位论文、专利、报纸等类型的成果元数据进行全面回溯。由于医科院机构体系复杂，各分支机构名称简称、别称、变形、组合较多，部分医院存在多个国家级重点实验室，而作者发文时常存在机构署名不

全等情况，因此需要不断补充中英文机构名列表，及时调整检索策略。此外需要对不同来源的相同成果数据进行去重，保留质量较好的数据，同时根据确认关键词对模糊查询结果进行过滤，将确认为本机构的成果保留在机构成果库中，被过滤掉的数据则另外存储。截至 2017 年 11 月 27 日 CAMS – IR 共回溯 1918 – 2017 年发表的原始数据量 339 738 条，其中判断为本机构成果数量326 897条，经去重合并后入库成果数据为 215 382 条，具体成果类型及数量，见表 1。表 2 展示 CAMS – IR 已搜集成果的收录情况，其中 PubMed 索引及中华医学会（Chinese

Medical Association, CMA) 期刊收录是医学领域的重要收录索引。目前 CAMS – IR 典藏成果的 WOS 引文频次达到 469 399 次, 自动获取全文 3 736 篇。可以看出 CAMS – IR 已基本完成对医科院公开成果的覆盖, 实现对院校历史知识资产的归纳整理。

(2) 成果数据处理。对已搜集成果进行数据处理, 开展成果确认过滤、自动清洗、规则去重、成果自动关联等处理工作。其中的难点在于如何针对确认为医科院成果的数据实现成果产权的自动关联。成果关联包括“成果 – 所院”关联和“成果 – 学者”关联两个方面。其中“成果 – 所院”关联是指根据作者署名机构、二级机构名及其变形将成果数据自动分配到二级机构。关联方法包括根据成果机构署名中的二级单位直接关联, 以及根据作者确认关系对成果进行关联。目前已关联到二级所院的成果量为 174 026 条, 约占成果总量的 79%。“成果 – 学者”关联包括确认作者成果和疑似作者成果, 如果成果中作者署名的二级机构与院校学者二级机构(包含机构变形)相同, 且作者唯一, 则判定为确认的成果 – 学者关系; 对于成果中未署名二级机构

或者二级机构不一致的作者以及同一机构同名作者的成果, 都作为疑似作者成果, 需要进一步人工审核及成果认领。(3) 成果质量控制。涵盖成果回溯及处理过程中的阶段质检、与自有数据的自动对比分析、人工抽检、馆员批量审核、成员认领成果以及对自有高质量全文数据的自动整合获取。为掌握已搜集成果的数据质量, 分别开展中英文期刊成果的数据检验。其中针对中文期刊成果, 将 CAMS – IR 中的中文期刊成果总量、关联到二级机构的成果数量与自建的中国生物医学文献数据库 (Chinese BioMedical Literature Database, CBM) 收录的院校成果数据进行数量及质量对比分析, 经统计 CAMS – IR 中文期刊类型成果的搜集准确率可达 95% 以上。针对英文期刊成果的数据检验, 主要考察 CAMS – IR 英文期刊成果数量与 WOS、PubMed 等重要数据库检索结果的数量对比, 经统计没有显著差异, 基本能够覆盖公开发表的英文成果数据。针对医学领域知识的特殊性, 引入相关背景学科的馆员支持, 通过宣传吸引机构成员参与个人成果认领, 更加有效保障成果质量。

表 1 CAMS – IR 回溯成果类型及数量

成果类型	期刊论文	会议论文	学位论文	报纸论文	专利	合计
成果数量 (篇)	166 729	24 953	20 069	589	3 042	215 382

表 2 CAMS – IR 成果在 8 种收录索引中的收录情况

收录情况	SCI (E)	CPCI – S	EI	PubMed	中国科学引文 数据库 (CSCD)	CMA 系列	中文核心期刊 要目总览	中国科技核心 期刊
成果数量 (篇)	3 956	2 543	1 455	26 276	50 101	28 921	52 939	57 628

3.2.4 院校学者库 为院校每位成员建立专属的学者 ID 及学术名片, 支持关联 ORCID 等学术身份认证, 对学者的成果、履历、照片、回忆录、访问记、文集等进行收录。目前 CAMS – IR 的成员用户为 12 000 余人, 其中有公开发表成果的学者为 9 809 人, 包括 46 位中国工程院及中国科学院院士。每位学者的主页都汇集学者本人的基本信息、公开成果、学术历程、合作者、发布期刊及项目经历等信息, 全面呈现院校学者的学术积累和文化风貌。

3.3 CAMS – IR 门户平台建设

3.3.1 建设思路 建设完备且用户体验良好的机构知识门户平台有利于展现机构的学术成果和文化价值, 同时吸引机构成员积极参与机构知识库的建设和完善。CAMS – IR 按照统一整合揭示、分级个性展示的思路, 开展机构知识门户系统的设计和开发。CAMS – IR 系统主要功能架构, 见图 2。在全面整合汇聚成果内容的基础上, 围绕公开成果、内

部成果及医学特色成果,实现院校、所院、学者等不同层面的成果展现;同时面向包括医教研人员在内的普通用户、学科馆员以及科研管理部门等多类

型目标用户,提供会员服务、馆员服务以及科研服务等多样化的功能和服务。

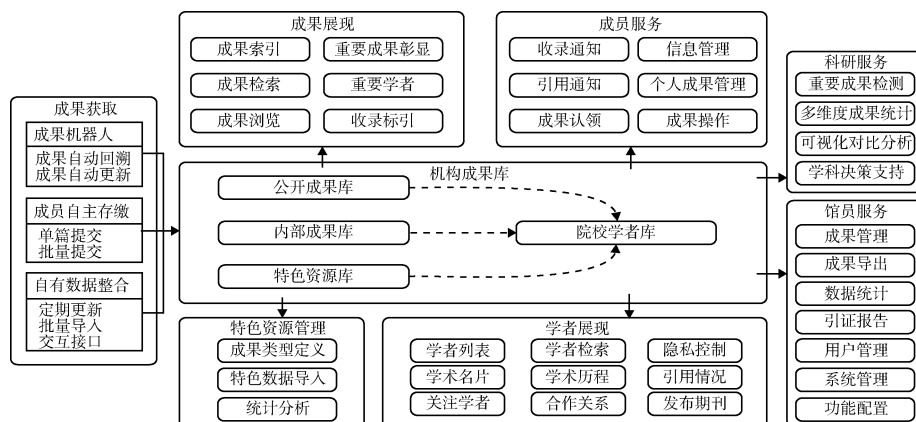


图2 CAMS-IR 系统主要功能架构

3.3.2 建设效果 CAMS-IR 门户系统,见图3。页面设计以“协和绿”为主色调,搭配象征白衣天使的白色,代表医科院人传承智慧,守护健康,为生命带来希望。首页采用巨幅滚动图片突出彰显院校的重要学术成果,重要成果通常遴选自重要医学学术期刊(如《柳叶刀》、《英国医学杂志》、《科学》、《自然》等)、高被引学术成果及高热点学术成果等。同时针对典藏成果中的关键数据统计、各所院成果统计、收录统计、成果类型、最新成果及热点成果等重点内容,分别从院校层面进行统一揭示。在首页重要位置对院校获得两院院士殊荣的46位院士照片进行滚动呈现,充分展示众多医学大家精诚奉献的风范。在二级页面主要以分支机构为单位,实现各所院科研产出的汇集与展示。支持从作者、内容类型、发表日期、收录类型、是否有全文、语种、项目等角度对各所院学术成果进行整合及分面展示,帮助用户快速定位目标资源。未来将进一步针对不同所院的示范应用需求,开展特色资源管理、个性化功能定制以及页面展示,从而突出彰显各所院的学术特色和研究方向。除在院校层面的统一揭示、在所院层面的分级展示设计外,CAMS-IR 也针对不同类型的用户提供多样化的知识服务。对于普通成员用户而言,CAMS-IR 主要

满足其获取成果信息、自我成果管理以及与医科院同行交流的需求,如用户可根据不同的维度进行快速成果检索及分类浏览,为喜欢的成果点赞;可在登录后提交、管理、认领并操作个人成果,通过学术名片关注其他学者,也可获得免费的论文收录、引用通知等服务。对于各所院的学科馆员而言,CAMS-IR 主要满足其对自有成果、用户及系统功能的管理需求,如馆员登录系统后通过馆员通道可对后台系统管理、功能配置、机构字典、重要期刊、权限设置、成员用户维护、成果内容、质量控制、数据统计等方面进行全面管理。对于院校及各所院的科研管理部门而言,CAMS-IR 支持通过接口方式为各个科研管理系统提供数据服务,为其自有成果的自动汇聚、持续更新、统计及趋势分析等需求提供支持,从而快速了解整个机构的科研状况。如可从不同机构层面开展重要成果检测;根据成果类型、所院、学科、年度、访问趋势等多样化的维度对机构科研产出进行统计分析;以可视化的形式,从作品类型分布、研究单元分布、论文收录类型分布、发文量排行、引用排行、关键词共现、作者合作网络等角度进行展示;在全面掌握自身成果产出及趋势的基础上,相关部门可以进一步开展针对本机构的学科发展规划和决策。



图3 CAMS-IR 门户系统

3.4 支撑机制

CAMS-IR 平台建成后,相关政策、制度的制定对保障其可持续发展有着重要作用。首先,开放获取政策是机构知识库的核心价值体现,院校在中科院、国家自然科学基金委员会等机构的经验基础上^[13],根据医科院的具体情况制定适用于 CAMS-IR 的开放获取、传播及权益保护政策,建立相应的管理运行制度、宣传推介科研激励政策,吸引机构成员参与机构库建设的同时,充分保证知识内容被可靠收集,知识产权得到妥善保护,知识管理能够规范进行,消除对竞争性使用和不合理使用的顾虑,有效规避管理和法律风险,从而保障院校机构知识库的可持续运行。

4 结语

CAMS-IR 承担收集、整理、保存、检索、利用院校各类学术成果的职责,同时也是补充和丰富医科院生物医学科技支撑平台资源整合的重要来源。随着开放获取理念的不断深化以及用户群体的逐渐增多,为实现 CAMS-IR 的可持续发展,需要更加注重成果内容建设,改善服务内容和服务质量,满足管理者和成员用户的多样化应用需求,从

而吸引用户共同参与建设完善。未来 CAMS-IR 的建设推广工作主要包括以下 3 个方面:(1) 继续完善成果内容。查缺补漏,持续自动更新,全面覆盖公开发表成果;与自有 CBM 数据进行有效整合,为成果内容建设提供质量保障,提高全文数据汇聚效率;通过多种途径宣传鼓励原发成果自存缴;联合共建单位汇聚内部成果及医学学科领域特色成果;结合系统自动处理、成员认领、馆员审核等方式,提高成果质量及产权关联准确性。(2) 可持续发展规划。建立全面、准确、及时、标准、规范的可持续内容建设机制;在硬件、软件、网络、内容处理等方面提供有力技术保障;探索与医学科学研究数据关联的可行性;从政策及业务上与各所院科研管理系统进行相互支持,实现科研知识产出的保存与重用;建立自存缴奖励机制及科研评价制度等。(3) 为用户提供多样化服务。在提升用户体验、积极宣传推广的基础上,根据共建院所需求进行个性化定制与示范应用;提供多维度深层次共享利用、学科建设数据支撑、科研评价分析等多样化知识服务。

参考文献

- 1 张晓林. 机构知识库的政策、功能和支撑机制分析[J]. 图书情报工作, 2008, 52 (1): 23-27.

(下转第 93 页)

学伦理必修课, 加强其医学伦理观教育; 其次学校应在医学生中开展网络道德与网络遵纪守法专题培训, 普及知识产权与网络安全知识, 深化并不断提高学生的信息意识和信息道德水平; 最后医学生自身应增强网络自律意识, 坚守道德底线, 规范自我网络言行, 不断增强医德医风及道德判断能力, 提高自我约束和自我保护能力, 提升个人修养, 养成道德自律。

5 结语

在互联网时代, 医学信息素养对于医学生而言是不可或缺的重要能力及素质。对于医学生信息素养教育应积极完善教育内容、创新教育方法、融入养成教育, 从而使医学生成为具有批判性思维能力和创新能力的信息素养人才, 成为符合社会发展需求的创新型医学人才。

参考文献

1 孙学钢. 高职医学生信息素养教育现状与对策研究

[J]. 科技情报开发与经济, 2015, 25 (6): 140.

- 2 初万江. 医学视听教材的现状和设计制作趋势分析 [J]. 中国医学教育技术, 2007, 21 (4): 304-306.
- 3 崔蒙. 中医药信息学概论 [M]. 北京: 科学出版社, 2016.
- 4 张意志, 徐小明, 陈华, 等. 高校附属医院医学生信息素质教育的探索与思考 [J]. 中国信息技术教育, 2014, 14 (16): 3145-3147.
- 5 邢春国, 廖云霞, 顾铮, 等. 高等院校医学生信息道德认知调查与分析 [J]. 中国中医药图书情报杂志, 2016, 40 (3): 23-26.
- 6 陈闻. 高中信息技术课堂教学中提高学生道德信息素养策略 [J]. 中小学电教, 2015, (10): 3.
- 7 张云秋, 李玉玲, 沈涌, 等. 我国医学生信息素养教育发展战略研究 [J]. 医学信息学杂志, 2010, 31 (12): 2-5.
- 8 黄芳, 吴晓海. “5+3+X” 临床医学人才培养模式下的医学信息素养教育 [J]. 中华医学图书情报杂志, 2015, 24 (12): 1-4.
- 9 邢春国. 在医学生中开展信息素养教育的必要性 [J]. 医学信息学杂志, 2010, 31 (4): 78-80.

(上接第 59 页)

- 2 The Directory of Open Access Repositories. Open DOAR [EB/OL]. [2017-11-27]. <http://www.openoar.org/>.
- 3 Registry of Open Access Repositories. ROAR [EB/OL]. [2017-11-27]. <http://roar.eprints.org/>.
- 4 中国科学院机构知识库网格. CAS IR GRID [EB/OL]. [2017-11-27]. <http://www.irgrid.ac.cn/>.
- 5 徐红玉, 李爱国. 中国科学院系统与高等学校机构知识库建设比较研究 [J]. 图书情报工作, 2014, 58 (12): 78-83.
- 6 张冬荣, 祝忠明, 李麟, 等. 中国科学院机构知识库建设推广与服务 [J]. 图书情报工作, 2013, 57 (1): 20-25.
- 7 CALIS 机构知识库 [EB/OL]. [2017-11-27]. <http://ir.lib.szu.edu.cn:8080>.
- 8 韦成府, 聂华, 崔海媛. 多馆协作开发的机构知识库建设——以 CALIS 机构知识库项目为例 [J]. 大学图书馆学报, 2014, (3): 69-73.
- 9 陈和. 机构知识库发展趋势探析 [J]. 图书情报工作, 2012, 56 (21): 62-66.
- 10 朱立禄, 宋世俊, 王琳. 国内外机构知识库建设现状及建议 [J]. 现代情报, 2017, 37 (3): 109-115.
- 11 张雪蕾, 魏青山, 尹飞. 构建服务扩展型机构知识库的实践与探索——以西安交通大学为例 [J]. 情报理论与实践, 2017, 40 (7): 93-98.
- 12 中国生物医学文献服务系统. SinoMed [EB/OL]. [2017-11-27]. <http://www.sinomed.ac.cn/>.
- 13 徐速, 张新鹤. 我国机构知识库开放获取政策体系与内容研究 [J]. 图书情报知识, 2017, (6): 42-49, 97.