基于开放获取理念的我国高校机构知识库建设探究

陈依玲 吕扬建

(同济大学电子与信息工程学院 上海 201804)

摘 要 通过对18所高校机构知识库进行调查,分析了资源收录数量、资源类型、浏览方式、语种分布、软件平台及征缴政策等现状,针对资金匮乏、资源不足、存在知识产权阻力、缺乏统一标准等问题,提出了相应的解决措施。

关键词 开放获取,高校机构知识库,IR,知识产权

中图法分类号 TP392 **文献标识码** A

Research on Construction of Chinese Academic Institutional Repositories Based on Open Access Consciousness

CHEN Yi-ling LV Yang-jian

(School of Electronics and Information, Tongji University, Shanghai 201804, China)

Abstract The paper took an investigation of the Institutional Repositories in Chinese universities, and analyses the status about amount of resources, types of resources, browse manners, language distribution, software platform, collection policies, and so on. There are some deficiencies existing in the academic Institutional Repositories, for example, lacking of funds and resources, existing intellectual resistance, short of uniform standards, etc. And then corresponding countermeasures were put forward.

Keywords Open access, Academic institutional repositories, IR, Intellectual property rights

1 引言

开放获取(Open Access,简称OA)是以互联网为依托的新型学术交流理念,不同的学术机构和学者对开放获取的定义有不同的看法。美国科技信息研究所(The Institute for Scientific Information,简称ISI)指出:任何经由同行评议的电子期刊,以免费的方式提供给读者或机构取用、下载、复制、打印、传播或检索文章,也即开放取阅[1]。布达佩斯在开放获取先导计划(Budapest Open Access Initiative,简称BOAI)中提到:"文献的'开放获取',意味着它可以在公共网络上被免费获取,允许任何用户对该文献的全文信息进行阅读、下载、复制、分发、打印、检索、超链接,支持爬行器收割并建立本地索引、用作软件的输入数据、用于其他任何法律允许的用途"[2]。

笔者认为,开放获取是基于学术信息共享理念和出版机制,为公众自由获取信息资源、实现资源优化配置而兴起的运动,迎合了网络时代信息交流的特点,并带动了机构知识库的兴起和发展,使其成为了开放获取的重要实现模式。

机构知识库(Institutional Repository,简称IR),是在开放获取的环境下形成的,以学术机构和团体为轴心的数字化信息及服务的集合,旨在实现知识产出的集中管理、长期保存及知识共享等[3]。学术出版和学术资源联盟组织的资深顾问Raym Crow从高校的角度出发,指出IR是收集并保存单个或数个大学、科研机构知识资源的数字化资源集合[4]。IR实现机构范围内的开放获取,确保机构产出结果的完整性和持续性,成为开放获取的"绿色通道"。

2 我国高校机构知识库发展现状

ROAR(Registry of Open Access Repositories)[5]提供全球IR发展的最新统计信息,由最早设计IR软件的英国南安普敦大学创办。我国不少高校建设了学位论文库和教师文库,但其主要以收集保存本机构知识产出为目的,并没有在ROAR、OpenDOAR(Directory of Open Access Repositories)[6]等开放仓储中注册。笔者采用网站访问和文献调研的方式,综合ROAR和OpenDOAR的统计,对我国18个高校IR的收录数量、资源类型、浏览方式、语种分布、软件平台等展开调查,调查时间为2013年8月7日至23日,调查结果如表1所列。

结果显示,我国高校IR建设还处在起步阶段,普及程度较低,数量和规模都不大,主要表现在以下几个方面:

(1)资源收藏内容和应用范围比较分散,大多以传统学术出版物和灰色文献为主,期刊论文、学术论文和会议资料占很大比例,演示文稿、音频、视频资料等较少。此外,存储内容普遍存在学科不均现象,例如福建师范大学和浙江大学的IR以图书馆学为主,北京科技大学奇迹文库以物理论文为主。

(2)大都以DSpace系统为原型,厦门大学对DSpace系统进行了本地化,台湾大学也完成了对DSpace的汉化和定制,使之可在英文、繁体中文和简体中文之间切换,其余高校基本采用软件默认设置。

(3)大部分高校IR资源数量增长幅度较小,更新缓慢。厦门大学IR[7]建库较早,资源增长相对稳定,是国内高校IR

陈依玲(1990—),女,硕士生,主要研究方向为嵌入式系统、计算机图形图像,E-mail:970694189@qq.com;吕扬建(1990—),男,硕士生,主要研究方向为图形处理、虚拟现实。

建设的典范，在西班牙的赛博计量学实验室(Cybermetrics Lab)亚洲区域开放获取 IR 排名第 15[8]；北京科技大学图书馆每年都需向校人事处和科研处提供本校教师的科研成果用于考核和评定，同时也将其导入 IR 中，保证数据的常年更新。

(4)香港地区使用的相同系统的高校 IR 其浏览方式基本相同，而大陆地区使用的相同系统的高校 IR 普遍呈现不同的浏览方式，从长远来看，不利于日后标准的统一和资源的整合。

表 1　我国高校机构知识库发展现状

机构知识库	收录数量	资源类型	浏览方式	语种分布	软件平台
厦门大学学术典藏库	56976	论文、会议资料、工作文稿、演示文稿	院系、发布日期、作者、题名、主题、个人专集、提交日期	中文	DSpace 本地化
清华大学机构库	78130	会议文件、报告、演示文稿、课件、中外文期刊论文、图片、科学数据、科研项目资料	院系、专题、日期、作者、标题、主题、类型、作者单位、个人主页	中文	其他
北京大学机构知识库	28673	期刊论文、会议论文、文章、书籍、工作文档、研究报告、计划蓝图、演讲介绍、口述记录	院系单位、作者、题名、主题、出版日期、提交日期、学者推荐、最新提交、热门浏览	中文 英文	DSpace 本地化
北京工业大学机构知识库	10346	期刊、会议论文集、杂志、报纸、报告	作者、年份、类型、院系列表	中文	其他
浙江大学机构知识库	38284	期刊论文、会议论文、学位论文、图书著作、专利文献、教学课件、荣誉奖励、新闻报道	资源、专家、团队、机构	中文	DSpace 默认设置
北京科技大学机构知识库	35804	期刊论文、新闻、综述、书籍、会议、讲义、笔记	题名、作者、资源、提交日期	中文	其他
中国农业大学知识库	33195	期刊论文、科研成果、博硕士论文	院系、学科、专家、关键词	中文	其他
福建师范大学图书馆学系机构库	2306	学术著作、期刊论文、会议论文、工作文稿、科研数据、演示文稿	题名、关键词、作者、机构库、发布时间	中文	DSpace 默认设置
台湾大学机构典藏库	157617	期刊论文、专书及章节、会议论文、学习教材、专利、未出版研究及工作报告、博硕士论文	题名、日期、作者、资料类型、社群与类别	中文 英文	DSpace 汉化、定制
台湾逢甲大学机构典藏系统	26068	期刊论文、学术论文、会议论文、学生作品、校史资料、学术报告、视频信息、数据集、古文书、校园影像	题名、作者、关键词、摘要、最新藏品、热门点阅	中文	DSpace 汉化、定制
澳门大学机构库	3148	期刊论文、会议论文、汇报总结	研究社群、出版日期、作者、标题	英文	DSpace 默认设置
香港城市大学机构知识库	7349	会议文件、期刊论文、简报、博硕士论文、学生期末作业项目论文、学生获奖项目论文、优秀本科生计划项目论文	题名、作者、主题、提交日期、研究社群	英文	DSpace 默认设置
香港大学学术库	133769	书籍、会议文件、检索工具、学术论文、图片、杂志、报纸文章、专利、学术项目、学生项目、研究生论文、本科生论文、简报、移动影像	院系、类型、作者、题名、文摘	中文 英文	DSpace 默认设置
香港理工大学机构库	8704	书籍、专利、学术论文、学位论文、期刊论文、技术报告、工作论文、会议文件、演示稿、预印本、简报、数据集、素描图、视频	标题、主题、作者、提交日期、研究社群	英文	DSpace 默认设置
香港科技大学机构库	8192	预印本、数据集、专利、技术报告、书籍章节、博硕士论文、会议文件、期刊论文、简报、工作文件、研究报告、百科文章、手册	院系、机构、标题、作者、题名、关键词、文摘、提交日期、研究社群	英文	DSpace 默认设置
香港中文大学机构知识库	72334	工作报告、学术文章、报刊、书籍、书籍章节、论文、会议报告、技术报告、专利	研究社群、作者、主题、标题、出版日期	英文	DSpace 默认设置
香港浸会大学机构知识库	8947	论文、书籍章节、专著、会议记录、乐曲、表演、会议论文、工作论文、演出及展览、艺术复制品	标题、主题、作者、提交日期、研究社群	中文 英文	Eprint
香港教育学院机构知识库	17653	专著、书籍章节、期刊论文、会议论文	研究社群、出版日期、作者、标题、主题	英文	DSpace 默认设置

3　我国高校机构知识库建设中存在的问题

3.1　缺乏国家政策、项目及资金支持

虽然已经有不断增长的报告、声明、政策来支持我国高校 IR 建设，但都只是停留在口号支持层面。大部分 IR 建设主要依赖高校本身的投资，经费来源单一、资金匮乏。国内最具权威的国家自然科学基金会、国家社会科学基金会以及教育部项目基金会也没有正式发表支持声明。相对于内陆高校，台湾地区高校 IR 建设得到了当局的战略决策支持，例如台湾大学 IR[9] 的建设经由教育部委托、政府牵头，成为了全台湾

各高等院校建置 IR 的参考。

国外高校 IR 建设普遍得到国家政策的支持，并启动大型的国家项目来推动 IR 建设和相关标准的实施。相比之下，我国高校缺乏合适的经济模式和运作模式来保证 IR 资源的长期保存与可获得性，进而引发资源增长率低、内容更新缓慢等现象，阻碍了 IR 建设的进程。

3.2 资源数量及质量缺乏保障

信息资源的数量和质量是 IR 建设的基础。调查显示，大部分高校 IR 建设存在资源不足、质量良莠不齐、更新缓慢的现象，用户利用 IR 和提交学术资源的积极性不高。在所调查的高校 IR 中，并没有通过行政方式要求提交学术成果，厦门大学也只是号召把有学术价值的数字资源提交到 IR 中，并没有采取强制性措施，这些柔性策略在一定程度上阻碍了 IR 资源的积累。

此外，IR 的开放性使信息资源的上传不受限制，内容质量得不到有效控制，开放获取模式所具有的快速与免费的优势被削弱，造成人力、物力、财力的浪费，还影响用户的学术研究以及高校的声誉。

3.3 存在知识产权或版权阻力

一方面，目前我国高校 IR 基本采用相对成熟、稳定的开源软件 DSpace 系统，个别高校将 DSpace 本地化。尽管这些开源软件允许修改代码，但我国在利用软件时往往忽略了知识产权的保护，高校 IR 的建设应避免对软件代码进行任意修改，避免用于商业活动，以免出现版权纠纷。

另一方面，开放获取和 IR 建设受到了期刊出版商和数据库商的反对，大部分出版商要求获得论文印刷版和电子版版权的授权。高校 IR 主要组成部分是预印本论文，自存储资源版权问题也集中体现在预印本上，存放在 IR 中的论文一旦被期刊录用，基于版权问题考虑，作者必须从 IR 中撤回论文。此外，由于 IR 允许用户不受限制地浏览、下载和打印论文，存在作者研究成果被剽窃、著作权受侵犯的风险，这些都制约了自存储资源提交的数量，也阻碍了 IR 发挥学术资源传播和共享的作用。

3.4 缺乏统一的知识组织标准及个性化功能服务

信息组织标准化是决定 IR 资源共享的关键问题，缺乏统一标准会导致在元数据和互操作性上存在一定的出入，造成“信息孤岛”现象。目前我国高校 IR 标准化建设存在元数据标引格式规范、文献著录标准检索功能等不一致的现象，进而引发用户自存档形成拼写、日期格式、主题描述等差异。

部分高校的 IR 系统不利于资源提供者操作，资源互操作和使用权限设置的实现存在一定的技术阻碍，影响了资源提供者提交科研成果的积极性，例如，数值内容与属性冲突、非专业化的词汇标引、前后标引不一致、格式不符合规范等。此外，由于我国高校 IR 采用的国外开源软件语种大多为英语，而网络上的翻译工具功能不齐全、翻译效果低下，且价格昂贵，使得不精通英语的用户存在一定的语言障碍。

4 我国高校机构知识库建设的策略研究

4.1 争取政策支持，增加资金来源

国家政策的支持和多方资金来源是高校 IR 发展的重要推动力。欧美国家 IR 发展迅速主要源于政府部门及基金会的支持、国家项目的推动、学术组织及机构联盟的呼吁。在亚洲，日本 IR 建设的领先地位也得益于政府的支持及国立情报学研究所的倡导，通过建立 IR 横向检索系统和内容分析系统，并设立专门机构协调各个高校和研究机构，开放专门的网站供用户进行跨部门搜索。

我国高校 IR 建设可借鉴国外经验，根据自身实际争取国家政策扶持，寻求多方资助机制。一方面，国家应制定高校 IR 资源收录、提交及获取政策，设立质量控制标准，统一元数据编码，重视内容存缴政策和版权保护政策的制定；另一方面，国家自然科学基金会、社会科学基金会和教育部项目基金会等应扶持高校 IR 的建设，并考虑将受其资助的科学研究成果纳入其中；此外，高校还应积极争取校方及本地区项目的资助，扩大资金来源。

4.2 争取自存储资源，重视质量保障

我国高校忽视对资源自存储政策的制定，学术认可和科研绩效评价也没有把 IR 资源纳入职称评审体系中，一定程度上制约了资源提交率。国外高校 IR 建设普遍存在三种资源征缴政策：强制存缴、激励存缴及建议存缴。Swan A. 认为政策存缴数量与政策强度成正比，强制存储是 IR 实施初始阶段内容收集最行之有效的方法[10]。利用强制政策把 IR 建设整合到学术交流和业务工作中，能促成科研成果的尽早提交，如美国国家卫生研究院对自存储政策做了重要的修订，将“请求”改为“要求”，“延迟保存”改为“立即保存”，大大提高了资源提交率。

高校应制定合适的资源征缴政策，主动了解各院系科研情况，加强对立项课题成果、获资助项目及课程的强制性收集；还应重点考虑与学术评价机制相结合的激励机制，将 IR 资源纳入科研考核和职称评价体系中，并根据自存储资源的阅读、点击、下载、被引等情况给作者相应的奖励。

此外，必须基于对象需求、研究价值和知识时效性，对所收集的学术资源进行严格的质量控制，通过建立合理的评审或评议制度，设置内容过期自动提示，及时更新 IR 内容，提高机构知识库的运行效率。

4.3 平衡多方利益，重视知识产权保护

国家应加强版权政策的制定，平衡信息资源提供者、出版商及 IR 之间的利益，确保用户非商业性的、不排他的使其作品开放存取的权利，对自存储资源的使用实施合理保护，通过签署相关协议书减少版权纠纷。IR 也应针对不同类型资源的版权问题做出明确规定：将版权归属于作者的教学课件、学术报告、论文预印本等资源纳入 IR 中；针对后印本的版权，则需研究出版商的版权政策，通过相互协商允许作者自我典藏；针对正式发表后的预印本，要及时与购买的电子资源建立链接，其余则可以文摘形式显示。

此外，在采用国外免费 IR 软件时，可根据实际需要进行适当的、非商业化的修改，但应在显著位置添加软件指定和代表软件所有权的徽标，链接该软件所有者的网站。在委托软件商开发 IR 软件时，须与软件商签署具有法律效力的知识产权协议，以免在使用过程中产生纠纷。

4.4 统一信息组织标准，重视服务功能扩展

高校 IR 建设应制定元数据的著录规范，尽量选择通用、简单的元数据格式，遵循统一的互操作性协议标准，如 OAIS (Open Archival Information System)参考模型、METS (Metadata Encoding and Transmission Standard)等，规定自存储资

源的格式，如 PDF 等，实现内容的有效传递和知识共享。在 IR 设计中尽量采用列表框选择数据的方法，如语言、学科、主题词、资源类型等规范化项目供资源提供者选择，尽量减少需要其录入的条目[11]，还可安排专人协助提交存储，辅助转换格式、纠正拼写错误、完成提交程序等。

此外，应重视 IR 服务功能的扩展，开发基于浏览、标记、订阅、检索、评论等满足个性化服务需求的增值功能，包括创建个人出版物列表、统计论文点击率和引用率等，如为作者提供建立和维护个人出版物目录的服务。还应着重考虑在 IR 中嵌入 RSS 订阅服务，为特定社区、特定作者及特定主题等提供多个站点实时更新的 RSS 源文件，并利用信息追踪技术和检索记忆理念，使科研人员输入的检索词为系统所记忆和匹配，进而实现相关资源的主动推送。

结束语 高校应通过多方渠道提高 IR 的知名度和认同感，使实施对象突破图书情报领域的实践者。积极借鉴国外 IR 建设经验，充分发挥图书馆辅助协调管理的主体角色，更加注重版权许可、质量控制、个性化增值服务及标准化建设等内容，促进学术交流和资源共享，扩大高校争取人才和资金的优势，提高核心竞争力。

参 考 文 献

[1] “开放获取 (OA)”推动信息共享 [EB/OL]. http://www.people.com.cn/GB/Paper464/15160/1344583.html,2013-08-08

[2] Budapest Open Access Initiative [EB/OL]. http://www.soros.org/openaccess/read.shtml,2013-08-11

[3] 孙振良. 高校机构知识库建设现状及策略研究[J]. 情报科学,2010(03):353

[4] Crow R. The Case for Institutional Repositories: A SPARC Position Paper. [EB/OL]. http://www.arl.org/sparc/IR/ir.html,2013-08-13

[5] ROAR: Registry of Open Access Repositories [EB/OL]. http://roar.eprints.org/,2013-08-10

[6] OpenDOAR: Directory of Open Access Repositories [EB/OL]. http://www.opendoar.org/countrylist.php? cContinent=Asia#China,2013-08-14

[7] 厦门大学学术典藏库 [EB/OL]. http://dspace.xmu.edu.cn/dspace/,2013-08-16

[8] Cybermetrics Lab. Ranking Web World Repositories[EB /OL]. http://repositories.webometrics.info/en/Asia,2013-08-15

[9] 台湾大学机构典藏[EB/OL]. [2013-08-13]. http://ntur.lib.ntu.edu.tw/

[10] Swan A, Brown S. Open access self-archiving: an author study [EB/OL]. http://cogprints.org/4385/,2013-08-13

[11] 郎庆华. 机构知识库自存储资源的获取策略研究[J]. 情报杂志,2009(7):169

(上接第 237 页)

个有效特征，通过加入该特征在准确率和召回率上均有提高。

表 1 不同的特征组合对 Hashtag 相关性检测的影响，符号 v 或 * 表示实验结果在显著水平 0.05 下优于(v)或不如(*)baseline 方法

	方法 1 (baseline)/%	方法 2/%	方法 3/%	方法 4/%
准确率	0.879	0.786 *	0.886	0.896
召回率	0.643	0.789v	0.868v	0.876v
F-值	0.741	0.787v	0.877v	0.886v

结束语 Hashtag 的话题相关性分析有助于从海量的微博信息中有效地挑选用户可能感兴趣的微博、帮助热点话题发现和聚合展示、方便用户沟通交流。本文根据微博文本的特点，不仅考虑了构成 Hashtag 的词汇重合度，还考虑了相关微博内容、Hashtag 的出现次数-时间分布、Hashtag 共现等特征来帮助 Hashtag 相关性分析。实验结果表明本文抽取的一系列特征都有助于 Hashtag 相关性判断。

由于存在着一些不相关的 Hashtag 共现的情况，在未来工作中，我们将考虑如何过滤这类噪音数据。另外，我们将挖掘其他特征来帮助判别 Hashtag 之间的相关性，如度量 Hashtag 之间的 google 距离[15]等。将 Hashtag 相关性判别技术用于微博聚类、Hashtag 推荐等也是未来工作之一。

参 考 文 献

[1] Rosa K D, Shah R, Lin B, et al. Topical clustering of tweets[C]// Proceedings of the ACM SIGIR: SWSM. 2011

[2] Sankaranarayanan J, Samet H, Teitler B E, et al. Twitterstand: news in tweets[C]// Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2009: 42-51

[3] 张晓艳. 新闻话题表示模型和关联追踪技术研究[D]. 长沙：国防科学技术大学，2010

[4] Pöschko J. Exploring Twitter Hashtags[Z]. 2011

[5] Antenucci D, Handy G, Modi A, et al. Classification of Tweets Via Clustering of Hashtags[Z]. 2011

[6] 郑斐然，苗夺谦，张志飞. 一种中文微博新闻话题检测的方法[J]. 计算机科学，2012，39(1)：138

[7] Cataldi M, Di Caro L, Schifanella C. Emerging topic detection on Twitter based on temporal and social terms evaluation[C]// Proceedings of the Tenth International Workshop on Multimedia Data Mining. ACM, 2010: 4

[8] Chang H C. A new perspective on twitter Hashtag use: diffusion of innovation theory[J]. Proceedings of the American Society for Information Science and Technology, 2010, 47(1): 1-4

[9] 随机森林-维基百科，自由的百科全书[DB/OL]. http://zh.wikipedia.org/wiki/随机森林，2013

[10] Leydesdorff L. On the normalization and visualization of author cocitation data: Salton's Cosine versus the Jaccard index[J]. Journal of the American Society for Information Science and Technology, 2008, 59(1): 77-85

[11] Laniado D, Mika P. Making sense of twitter[M]. The Semantic Web-ISWC 2010. Springer Berlin Heidelberg, 2010: 470-485

[12] Guo W, Li H, Ji H, et al. Linking Tweets to News: A Framework to Enrich Short Text Data in Social Media

[13] Wang A H. Don't follow me: Spam detection in twitter[C]// Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on. IEEE, 2010: 1-10

[14] Benevenuto F, Magno G, Rodrigues T, et al. Detecting spammers on twitter[C]// Collaboration, electronic messaging, anti-abuse and spam conference (CEAS). 2010

[15] Cilibrasi R L, Vitanyi P M B. The google similarity distance[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(3): 370-383