

●何华连

我国全文数据库建设 存在的主要问题和改进措施*

摘 要 我国全文数据库建设目前存在的主要问题有:完整性不足,时效性差,检索性有待加强,客观性欠缺,实用性不够理想等。提高全文数据库数据库质量的措施有:建立全文数据库建设的宏观调控机制,加快著作权立法与资源组织标准化的步伐,加强印刷型期刊编辑者与全文数据库生产者的联系,改进完善数据库生产的相关技术,提高全文数据库的质量。表1。参考文献6。

关键词 数字期刊 全文数据库 数据库建设

分类号 G250.74

ABSTRACT The author analyzes some problems in the development of full-text databases of periodicals and proposes some measures to solve these problems, such as macroscopic regulation mechanisms, intellectual property legislation, resource organization standardization, technology improving and database quality improving. 1 tab. 6 refs.

KEY WORDS Digital periodicals. Database of full-text periodicals. Database development.

CLASS NUMBER G250.74

传播信息新而快是期刊的特点。在这方面,较之印刷型期刊,电子期刊有着无与伦比、得天独厚的优势。本文根据我国电子期刊生产现状,以《中文科技期刊数据库》、《万方数据资源系统》与《中国学术期刊(镜像)》(以下分别简称维普、万方、同方)收录《中文核心期刊要目总览:2004年版》(以下简称《总览》)评选出的693种哲学社会科学方面的核心期刊的统计数据为依据,简要论述我国全文数据库建设存在的主要问题及其改进措施。

1 我国全文数据库建设存在的主要问题

1.1 完整性不足

期刊收藏必须遵循连续性、系统性和完整性原则,目前我国的全文数据库却都未能达到这一要求。《总览》及维普、万方、同方等库所收录的我国哲学社会科学方面的核心期刊的统计数据见表1。因为《总览》评选核心期刊是分类进行的,所以综合性类

目与专科性类目及交叉类目评选出的核心期刊有一定的重复。若将各类目罗列的核心期刊简单累计,总数为737种。在这737种核心期刊中,同时入选2个类目的为39种,同时入选3个类目的为5种。排重后,《总览》评选出的核心期刊实际为693种。就收录的期刊种数而言,《总览》评选出的693种哲学社会科学的核心期刊中,同方收录657种,收录率为94.81%;维普收录597种,收录率为86.15%;万方收录245种,收录率为35.35%。除同方收齐了哲学大类的核心期刊外,三库均未收录齐全其他大类的核心期刊。按大类统计,收录率最低的学科门类,所收核心期刊的种数仅占核心期刊总量的12.24%。就收录的论文篇目而言,维普库的“文学作品”类只收录了《山花》3篇、《钟山》2篇、《民族文学》2篇、《花城》1篇。同方、万方也有类似情况。就收录时间而言,上述三库也无一做到连续、系统收录。

表1 维普、万方、同方库收录我国哲学社会科学核心期刊的数量统计

(种)

类目名称	《总览》	维普	万方	同方
A/K 综合性人文、社会科学	99	96	32	98

* 本文为浙江省教育厅课题“国内电子期刊导航系统比较研究”(项目编号20040867)的研究成果之一。

续表

类目名称	《总览》	维普	万方	同方
B 哲学	23	22	6	23
C 社会科学总论	36	35	19	34
D 政治	83	75	27	80
E 军事	9	2	2	3
F 经济	183	181	63	166
G 文化、科学、教育、体育	150	147	51	136
H 语言、文字	28	11	10	26
I 文学	49	6	12	25
J 艺术	41	12	10	33
K 历史	36	10	13	33
合 计	737	597	245	657

1.2 时效性差

信息新颖、传播迅速是期刊的特点。但目前我国的全文期刊库的时效性很不理想。全文期刊库采用的是重新加工制作的方式,而不是数据链接或转换的方式,电子版期刊与读者见面的时间远远落后于印刷型期刊。笔者对《总览》收录的 1798 种核心期刊按出版周期作了分类统计,双月刊、月刊、季刊分别为 833、512、412 种,共占总量的 97.72%。而周刊、旬刊、半月刊、半年刊总共只有 41 种,仅占总量的 2.28%。针对中文期刊以双月刊、月刊、季刊为主的现实,笔者又随机抽样调查了《开放教育研究》等 16 种双月刊、《探索与争鸣》等 12 种月刊、《数学理论与应用》等 7 种季刊,通过印刷型期刊到图书馆时间与读者利用全文期刊库所收录相同资料的时间比对,推算出双月刊、月刊、季刊的全文数据分别滞后于印刷型刊物 5.0、4.5、5.6 个月,平均值约为 5 个月。也就是说,在印刷型刊物到馆 5 个月后才能看到全文期刊库中的同类资料。

1.3 检索性有待加强

众多的检索入口、强大的检索功能,可进行匹配检索、模糊检索,提示与外化文献中的任何一个知识点、自然词,涉及范围广、检索速度快,是全文期刊库的优势。然而就检索而言,我国的全文期刊库目前至少存在四方面问题。一是各异构数据库之间缺少统一的检索平台,读者查找特定资料时需逐一检索各数据库,检索过程烦琐。二是已有的全文期刊库检索界面设计得五花八门,读者检索特定库的资料时难以运

用自如,得心应手。三是各全文期刊库的刊名检索均未实现特性检索与族性检索的和谐统一。如在刊名检索的环境中输入“社会”一词,欲查找《社会》被全文期刊库收录了多少篇文章,但检索结果却包含全文库中所有冠有“社会”一词刊物发表的所有文章,涉及的篇目竟达 37 万余篇,如此等等。四是主题检索的设计制作并未到位。同方依据论文在原载期刊中所给定的关键词、摘要等著录,原作中无关键词、摘要或著录不规范的无法纳入该主题范围。维普采用图像扫描的方式建库,也无法实现全文检索。以上问题的存在极大削弱了全文期刊库的可检索性,严重减损了全文期刊库的检索功能。

1.4 客观性欠缺

客观地、实事求是地揭示特定的全文期刊库的内涵是数据库生产者应遵循的基本准则,然而目前我国的全文期刊库并非如此。一是各数据库还程度不同地存在着揭示所收录的期刊种数不明晰的问题。有人曾对维普收录的期刊实际数量作过调查。维普自称 7980 种,而实际调查结果是,收录基本齐全的期刊为 4872 种。二是上述三库在正确、及时地反映所收录期刊的停刊、改名、合并等方面的发展演变时取得的成绩虽然存在差异,但各库也都程度不同地存在一些问题,如《广西会计》已于 2003 年底停刊,但上述全文期刊库却都未能予以反映;称《湖南档案》改名为《档案时空》、《心理学动态》改名为《心理科学进展》同样有误,这里的“改名”其实是“合并”。此外,新版《总览》出版于 2004 年 7 月,而《天津社会科

学》、《浙江学刊》、《金融论坛》、《现代情报》、《科学新闻》等为数不少的刊物迄今还未被纳入“核心期刊”,行动过于迟缓。揭示期刊种数不清晰,不能正确、及时地反映所收录期刊发展演变的缺憾既影响了全文期刊数据库的客观性,也影响了数据库生产者的声誉。

1.5 实用性不够理想

上述全文期刊库的实用性不尽如人意,主要表现在4个方面。一是维普以扫描原文的方式建库,其检索结果输出时页面不够清晰,而通过OCR插件对选定区域的扫描图像格式进行文本转换,除文字外又会出现乱码较多的现象。二是各库所提供资料的差错率明显高于印刷型检索刊物,二次文献的篇目索引中的重复、误差尤其明显。三是现行的中文字库满足不了文字录入的实际需求,因文字缺失,获取的信息难以正确取舍。从全文期刊库的篇目索引中查到的篇名中,一些偏僻字由黑方块代替,想看懂题意并进而确定这些资料是否必需并非易事。四是全文期刊库在编制篇名索引时一味照录原载期刊的篇目而不加筛选的做法也不妥当。《山西档案》2004年第4期共刊载了23篇文章,其中5篇为:《老三毛上网》(24)、《投稿须知》、《有问有答》(16)、《简讯》、《捉拿错别字猎手榜》。无论从哪个角度看,提供学术论文的数据库为此类文字编制篇目索引都毫无意义。上述问题的存在,影响了全文期刊库实用性的发挥。

2 提高我国全文期刊库质量的应对措施

2.1 建立全文期刊库建设的宏观调控机制

我国的全文期刊库明显存在重复建设与空白缺门并存的现象。一方面,《总览》选定的693种哲学社会科学方面的核心期刊,维普、万方、同方三库同时收录的为209种,占总量的30.16%;两库同时收录的为383种,占总量的55.27%。重复收录的比例高达85.43%。如此多的重复建设,不仅造成巨大的人力、物力、财力浪费,也给读者逐库检索、去除重复增加了不必要麻烦。另一方面,上述三库均未收录的核心期刊为34种,占总量的4.91%。

全文期刊库建设中的文献过剩与文献稀缺并存局面的形成有多方面原因,最主要原因是未能建立宏观调控机制,没有统筹规划。因此,建立全文期刊数据库建设的宏观调控机制,是首先应当解决的问题。笔者以为,全文期刊库建设已形成的格局不宜过大调整,可在保持各库已形成的基本特色基础上,由国家

新闻出版总署牵头,组织专门机构,会同维普、万方、同方,共同商讨确定各库的收录范围、收录重点,最大限度地缓解文献冗余与文献不足的矛盾。通过宏观调控,达成有序生产,是我国全文期刊库建设的努力方向。

2.2 加快著作权立法与资源组织标准化的步伐

未来学家奈斯比特认为,失去控制和无组织的信息在信息社会里不再构成资源。在实现我国全文期刊库建设有效控制的基础上,加强期刊资源的标准化、规范化组织也十分重要。数字资源制作时间长、成本高、投入大和易被复制的特点造成了现有全文数据库的异构,纸本期刊与全文期刊库之间激烈的市场竞争又使文学期刊与部分权威性的学术期刊不愿被全文期刊库收录,经济效益的此长彼消也使纸本期刊编辑者与全文期刊库生产者的长期而稳定的合作关系受到冲击,从而影响了全文期刊库所收录资料的连续性、完整性与系统性。上述情况的出现与我国著作权法的不完善有关。完善著作权立法,以法律规范来约束、协调两者的关系,指导全文期刊库建设已成为当务之急。建设全文期刊库应遵循“以用户为中心”的原则、系统性原则、开放性原则、易用性原则和平衡性原则。期刊信息资源的标准化、规范化组织是全文期刊库建设的基础,著作权立法是解决目前全文期刊库建设中使用中存在问题的法律保障。

2.3 加强印刷型期刊编辑者与全文期刊库生产者的联系

全文期刊库在体现期刊信息的时效性方面原本有着与生俱来的优势。然而,由于我国的全文期刊库目前都是在印刷型期刊基础上加工制作的,因而其时效性明显逊色于印刷型期刊。数字期刊应下大气力提高时效性。加强印刷型期刊编辑者与全文期刊库生产者之间的沟通联系,密切两者的关系,是提高全文期刊库时效性的根本性举措。加强印刷型期刊编辑者与全文期刊库生产者的协作协调,争取印刷型期刊与全文期刊库的同步出版;印刷型期刊的编辑出版部门应在缩短审稿周期、加快编辑出版的工作流程上下功夫,而全文期刊库的生产部门应在缩短加工制作时间,保证相关数据及时上载方面采取更有力措施。只有这样,全文期刊库时效性强的优势才能真正发挥,对读者的吸引力也会大大增强。

2.4 改进并完善数据库生产的相关技术

异构数据库给读者查找专题资料带来了不便。统一检索平台、SFX技术让读者看到了(下转第107页)

设计与制造类的图书。与此同时,随着中国对外开放程度的不断提高,掌握外语对于人们能否在社会竞争中处于优势地位具有重要的作用,因此,以外语教育、外语应试为代表的教育类图书在大众读者中的需求量也呈现出上升的趋势。

3.3 从馆藏内容来看,公共图书馆的馆藏应以文学类、工业技术类、综合类图书为重点

情报学家 Kane 曾经指出:“将来的图书馆应根据 20-80 定律,将利用率高的核心文献收藏进馆,这些 20% 的藏书能够满足 80% 读者的需要,其余利用率不太高的文献可用馆际互借或电子信息源(包括 CD-ROM 产品,网上资源,全文数据服务等)获取。”^[3]从读者的阅读范围和阅读兴趣来看,文学类、工业技术类和综合类图书的借阅比例高达 50.8%,足以说明这三类图书的利用率是最高的;从读者的阅读需求来看,尽管公共图书馆的文学类图书馆藏相对丰富,读者仍然认为它最缺的是文学类图书;其次,有相当一部分读者认为图书馆最缺的是工业技术类图书;此外,文化、科学、教育、体育类图书和综合类图书均存在供不应求的情况。由此可见,从馆藏内容来看,公共图书馆的馆藏应以文学类、工业技术类、综合类图书为重点。

3.4 从馆藏管理来看,要及时了解图书出版信息,加快馆藏书目的更新速度

调查结果表明,馆藏更新太慢是当前公共图书馆的主要问题之一。更新太慢意味着图书馆无法有效满足读者及时获取最新知识的需求,因而也无法为读者提供高质量的服务。公共图书馆需要及时了解图书出版信息,加快馆藏书目的更新速度。

参考文献

- 1 红梅,张晓军,公共图书馆应加强对用户的吸引力. 图书馆学研究,1998(2)
- 2 石惠侠,谢林. 公共图书馆馆藏文献资源建设新思维. 中国图书馆学报,2000(2)
- 3 Kane L. T. "Access vs. ownership: do we have to make a choice?" College & Research Libraries. 1997, 58(1)

桂 胜 武汉大学法学院教授,博士生导师。通信地址:武汉市。邮编 430072。

田北海 武汉大学法学院博士生。通信地址:武汉市。邮编 710004。(来稿时间:2005-10-12)

(上接第 102 页)方便、快捷地检索、利用资料的希望。但目前统一检索平台的架构、SFX 技术并未成熟,还需不断探索,反复实践;已有的全文期刊库的检索界面设计目前也并未适应所有读者群的需求。只有改进并不断完善检索界面设计,读者检索资料时才能得心应手。各种全文期刊库在通过刊名检索相关资料时存在的特性检索的结果与族性检索的结果混为一谈的局面,也需要相关技术使之改观。文字录入技术是全文期刊库建设的基础性技术,应扩大中文字库以满足文字录入的需要。

2.5 提高全文期刊库的质量

全文期刊库的质量受多种因素制约。体制层面的原因,导致全文期刊库建设处于无序竞争状态;法制建设层面的原因,导致全文期刊库无法系统而完整地收录相关期刊文献;技术层面的原因,导致全文期刊库未能妥善解决特性检索与族性检索之间的矛盾,也未能圆满解决异构数据库的一站式检索问题及文字的录入问题;管理层面的原因,导致全文期刊库时效性低。要提高全文期刊库质量,必须妥善解决这些问题。此外,以下几方面的工作也值得重视并应尽快付诸实施。一是全文期刊库的生产者应密切关注所收录期刊的发展演变,及时准确地反映所收录期刊停刊、改名、合并等方面的情况,为读者提供及时可靠的信息。二是以原载期刊为依据,凭借数字文献的独特优势,按标

准化、规范化要求重新确定检索词,进一步增强全文期刊库的检索功能,满足读者通过知识点、自然词检索相关资料的需求。三是落实岗位责任制,加强督查,努力减少工作疏误,把全文期刊库中的篇目重复率和资料差错率降到最低限度。全文期刊库的经济效益和社会效益与它的质量密切相关,数据库生产者应该重视市场开发,但更应该重视产品质量的提高,视质量为生命的全文期刊库才堪寄予厚望,才有光辉前景。

参考文献

- 1 戴龙基,蔡蓉华. 中文核心期刊要目总览:2004 年版. 北京:北京大学出版社,2004
- 2 胡昌平,王翠萍. 基于个性化服务的信息资源组织目标、原则与规范. 图书馆论坛,2004(6)
- 3 马大川,杨红平. 信息资源的集成整合研究. 中国图书馆学报,2004(3)
- 4 白永革,彭佳. 提供一站式服务的电子资源融合系统的设计与实现. 大学图书馆学报,2005(1)
- 5 俞晓霞. 国内最具影响的两种综合性全文期刊库的比较分析. 江苏图书馆学报,2002(2)
- 6 何华连. 期刊信息资源开发利用有待解决的几个问题. 情报学报,2004(1)

何华连 浙江师范大学文献信息研究中心主任,图书馆副馆长,研究馆员。通信地址:浙江省金华市。邮编 321004。(来稿时间:2005-09-23)