

中文知识链接门户的构筑

曾建勋

(万方数据股份有限公司, 北京 100038)

摘要 文章在分析我国科学引文索引建设与利用情况的基础上,提出了利用学术文献引证关系和 WWW 链接机制构造中文知识链接门户的思想,介绍了中文知识链接门户的开发过程和基本功能,论述了其作为中文信息资源整合平台和信息分析工具的重要特征,并阐明和规划了其发展重点和方向。

关键词 知识门 开放链接 引文索引 期刊评价

Construction of Chinese Knowledge Linking Portal

Zeng Jianxun

(Wanfang Data Co., Ltd., Beijing 100038)

Abstract On the basis of analyzing the situation of the construction and utilization of some products similar to SCI in our country, the paper suggests a thought that the Chinese knowledge linking portal can be created by utilizing the academic literature citing relationship and the WWW linking system, and introduces the developing process and the basic functions of the Chinese knowledge linking portal, states its significant features as the integration platform of Chinese information resource. In addition, the article also illustrates and plans the key points and direction of the Chinese knowledge linking portal's development.

Keywords knowledge portal, open access, citation index, periodical appraising.

1 引言

1961年,美国科学信息研究所(ISI)研制成功了《科学引文索引》,1973年出版《社会科学引文索引》。近年来,随着因特网的不断发展,ISI继1997年推出基于Web的引文索引数据库Web of science,反映信息间内在联系以及开放扩展的结构,即通过因特网浏览器直接上网检索SCI扩展版、SSCI、A & HCI三大索引数据库。之后,ISI凭借其独特的引文机制和WWW链接特性,于2001年5月又推出了学术信息资源体系Web of knowledge。该体系收录200多个学科领域内最具影响力的学术出版物,以Web

of science为核心,不仅有效地整合了不同学科的数据库,而且还与图书馆OPAC系统、网络期刊全文以及精选的学术网站建立了相互链接,形成以知识为基础的学术信息资源整合体系。

我国于20世纪90年代开始,先后有几家单位研制科学引文索引。中国科技信息研究所研制《中国科技论文与引文数据库》、中国科学院文献情报中心研制《中国科学引文数据库》、2000年南京大学与香港科技大学联合研制《中文社会科学引文索引》。目前,引文索引已经成为评价我国各地区、机构、学者科研水平的重要工具,在各个学科中都得到了广泛应用。

虽然我国的引文研究异常活跃,引文数据库产

收稿日期:2005年4月29日

作者简介:曾建勋,男,研究馆员,1988年武汉大学图书情报学院硕士毕业,先后在首钢研究与开发公司、中国科技情报学会工作,现在中国科技信息研究所、万方数据股份有限公司从事信息资源建设工作。先后发表情报研究、期刊数据库建设、信息资源与网络媒体等方面的文章近70篇。

品也较多,但是引文研究力量分散,各引文数据库的产品很多局限在行业内部。总体来看,引文数据库产品市场占有率普遍较小,研发多数离不开各自行业主管单位的资金支持。“引文链接”是由传统的信息检索服务向知识服务转变的一个十分关键的功能环节,对信息资源的知识网络的形成有决定性的影响,而且是实现知识挖掘、发现、知识管理的基础。我国的主要引文数据库产品大多不支持“引文链接”,开发重心在于出版相关的引证研究报告,主要服务于科研管理评价,研究重点没有集中于信息资源的有机整合和知识管理上,引文索引缺乏关键词、文摘、以及与相关出版单位、文献传递单位有效的全文链接和服务合作,没有建立一种基于引文链接的知识整合平台。

近年来,国内很多单位虽然相继引进了 ISI 的 Web of Knowledge,然而,基于 SCI 的文献计量指标并不能代表中国的科学发展状况。由于特定的服务对象和定位,SCI 主要侧重于收录、检索以美、英等英语国家为主的基础科学类期刊,中国科学家的绝大多数学术论文没能被 SCI 收录。Web of knowledge 并没有整合中国的知识信息资源。所以,为了有效整合中文信息资源,提供中文知识成果的科学评价工具,有必要建立中国自己的基于期刊引文的知识链接门户。集期刊文献快速报道、引文关系揭示、学科状况分析和资源开放链接于一体,形成中文学术信息资源整合平台。

2 中文知识链接门户的设计思路

万方数据股份有限公司是科技部下属的我国第一家以信息服务为主体的股份制公司,其沿袭了中国科技信息研究所 40 多年的信息采集、数据库开发的主要业务内容,传承了上百个数据库产品和服务。在万方数据资源系统中,既有来源于国家“九五”重点科技攻关项目的中国数字化期刊群,还有中国科技论文与引文数据库、中国科技文摘数据库;既有中国学位论文数据库、中国学术会议论文数据库等高质量文献数据库,还有中国科技成果数据库、中国公司产品数据库、中国科技名人数据库等颇具特色的资源;既有自建的资源体系,又吸纳其他外部数据库加入。

近年中国数字化期刊群吸纳 5000 种学术期刊上网,其中,中文核心期刊科技类占 96%,社科类占 50% 以上,期刊论文量达到 500 多万篇、引文量达到

1400 多万条。加上中国科技论文引文数据库、中国科技文摘数据库,期刊论文总量也达到 1000 万条,为建立中文科学引文索引奠定了规模化的数据基础,也萌发了基于科学引文索引机制,利用开放链接标准,构建中文 Web of Knowledge 的思想;通过期刊中的各类型引文,把期刊、学位论文、会议论文、标准、专利、图书目次、外文期刊馆藏目录等关联和链接起来,通过论文的作者及其所属机构,把中国公司产品数据库和中国科研机构数据库、中国科技名人数据库和中国科技成果数据库沟通整合起来(如图 1 所示),从而形成中文知识链接门户。整个开发分以下步骤:

(1) 实现科技部下达的期刊全文上网的课题目标,实现中国科技期刊的网络化出版,丰富网上中文信息资源;

(2) 在中国数字化期刊群基础上,对中国科技论文与引文数据库、中国科技文摘数据库进行有机整合,形成更大范围的期刊引文索引,形成一种新型的文献检索工具,提供期刊论文与引文之间的关联检索。

(3) 建立一种基于 Web 的文献计量指标系统,打造中国的 JCR Web,并与引文索引条目有效挂接,成为我国科研成果的评价、学术期刊水平的测定、地区和机构科研实力的评估以及各学科核心作者群的确立的计量工具。

(4) 以期刊引文为核心,实现各种文献资源之间的整合与沟通。期刊引文中含各种类型的被引文献,利用多种类型文献之间的相互引证、相关参考的关系,对万方数据资源系统中,除期刊以外的会议论文、学位论文、标准文献、专利文献等资源进行有机整合,形成基于期刊引文的动态学术信息门户。

(5) 采用“一站式”信息服务的思路,实现对拥有使用权限的全文文献以及事实数据的链接和直接获取。

总之,以中国数字化期刊群为核心,将来源于学术期刊、技术专利、会议录、学位论文、技术标准及其它各种高质量信息资源整合在同一平台中,提供科学研究的全方位中文信息。兼具知识的检索、提取、管理、分析与评价等多项功能,扩展和加深信息检索的广度和深度,使研究人员不仅可以获得最新期刊的目次、文章题录和摘要等二次文献信息,而且可以获取完整的引文和被引信息;不仅可以从自身数据库或其他途径获取一次文献;而且可以获得期刊的评价统计指标,从而使研究人员获得各个领域学术

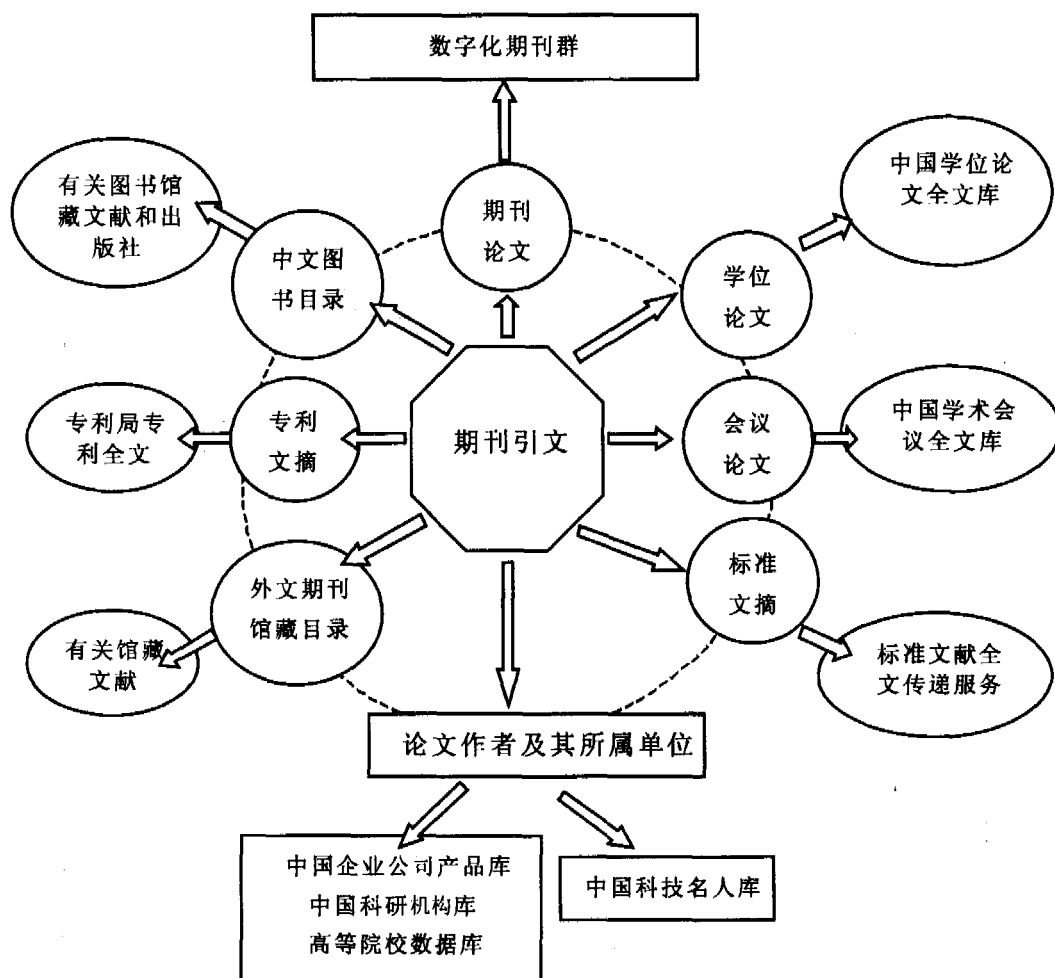


图1 各文献类型间的引文关系链接图示

研究的发展、影响和趋势报告,更好地掌握各个学术领域的渊源、变化、动态、走势及应用情况。

3 中文知识链接门户的开发过程

为了实现上述设计思想,在建立基于 DC 和 OpenURL 的数据库建设标准的基础上,确立利于链接的数据处理规范,设计适宜于指标统计的数据库结构,建立便于规范的关联字典。开发多条数据加工自动生产线系统,形成了一整套适应引文规范的数据处理流程和加工方案(如图2)。

整个系统数据库结构分为来源文献库、被引文献库、作者库、基金库、期刊载文表、期刊引文表等。通过“来源文献唯一标识”,将来源文献库与被引文献库、作者库中的相关记录联系起来。通过期刊规范字典、类主题字典、机构规范字典、基金规范词表等关联字典进行数据规范、优化检索(如图3),关

联字典设有规范词、非规范词、关联项、文献记录号、词频等字段。其中关联项主要记录规范词与非规范词之间的关系,记录号则提供关联字典和文献数据库中记录的联系途径,关联字典的应用将相同的错误全部一致修正,从而提高链接和统计的几率和效率,满足各类检索、统计、链接的需要。

数据处理是在中国数字化期刊群基础之上进行的。为了提高引文数据库的质量,主要以现有的数据库为参照标准,对参考文献数据进行逐一核查和规范处理,将每条参考文献与库中相应文献进行自动比对,对同一篇文献,通过人工判读检查作者、题名、刊名、年卷期、起始页等项目是否正确和齐全,发现错误或项目不全,则进行核查更正或用数据库中的正确记录替换。同时将被译成英文的中文参考文献逐条核查规范处理,再转换成中文,使得中国出版的英文版期刊能够全面统计,提高引文数据的规范化程度,保证检索的关联度、查准率和链接率。

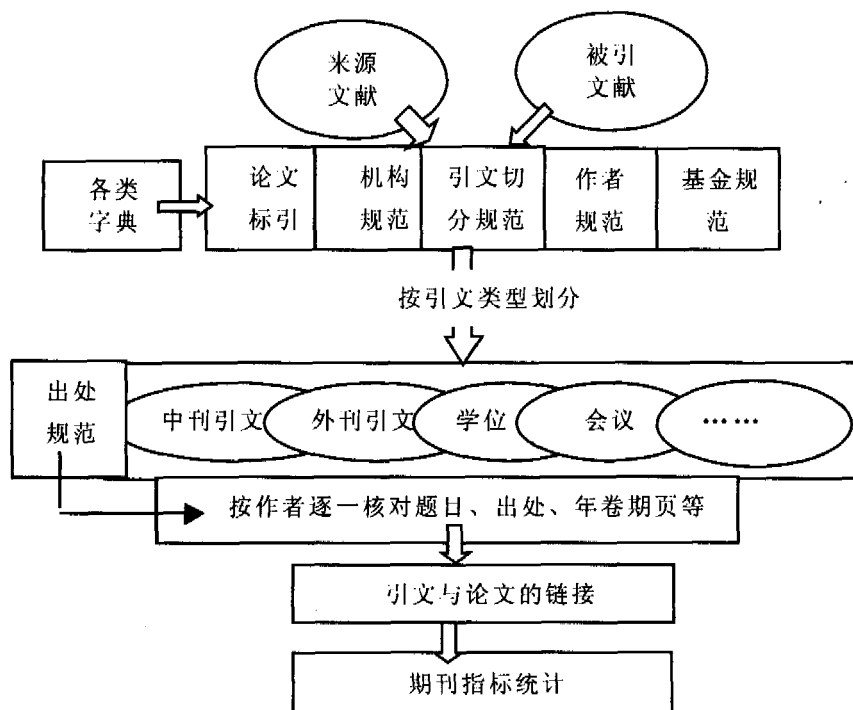


图2 论文引文规范链接处理过程

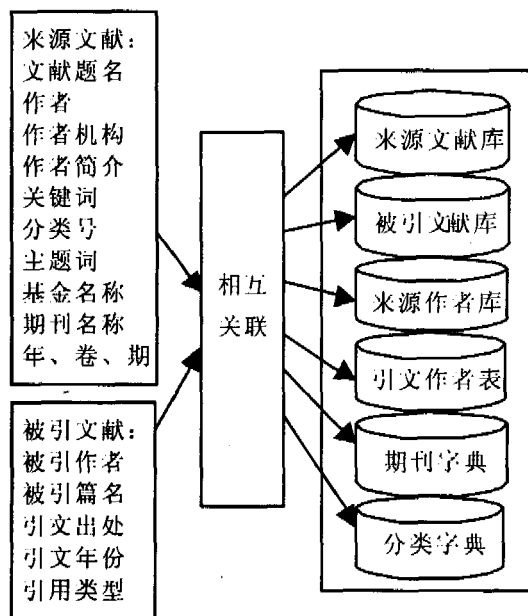


图3 论文引文间关联检索的数据库结构

期刊论文引文的统计从6个方面展开,即期刊、作者、学科、机构、地区和基金等方面来统计发表文章情况及被引用情况(如图4)。从而分析学科的核心期刊分布状况、遴选优势学科、评估优秀人才和成果、分析地区科研发展水平、基金资助分布和使用效果,以及测评同类机构科研能力等。

4 中文知识链接门户的基本功能

在整合中国数字化期刊群、中国科技论文与引文数据库、中国科技文摘数据库之后,便形成论文引文总量都达千万条记录规模的引文索引,继而,借助因特网 Web 的超文本技术,把自建的高质量学术文献数据库灵活地整合起来,不仅能从引文检索方式,还可从全记录页面的各种链接来获得相关文献,揭示了复杂的学术信息之间的逻辑关系。检索结果不是简单的排列与堆积,而是在引文索引基础之上的有机联系与综合,从中了解学术思想的创立及动态发展过程。整个中文知识链接门户基本功能分为:

(1) 关联检索:整个系统借用引文与论文间的相互关联关系,可以在一定时间范围内,从论文、引文和期刊等三方面,按照标题、作者、机构、关键词、基金、学科分类等入口进行检索。在被检中的来源记录即一篇论文的记录后都注明该篇论文被引次数、中文参考文献数和相关文献数。

点击“被引次数”热链,就会显示所有引用该篇论文的文献列表,由此,可以迅速了解该篇论文被其他哪些论文所引用,这些论文又出于何处。只要任意点击论文记录后注明的该篇论文“被引次数”,便又成为一组新的来源记录,点击 HTML 或 PDF 查看

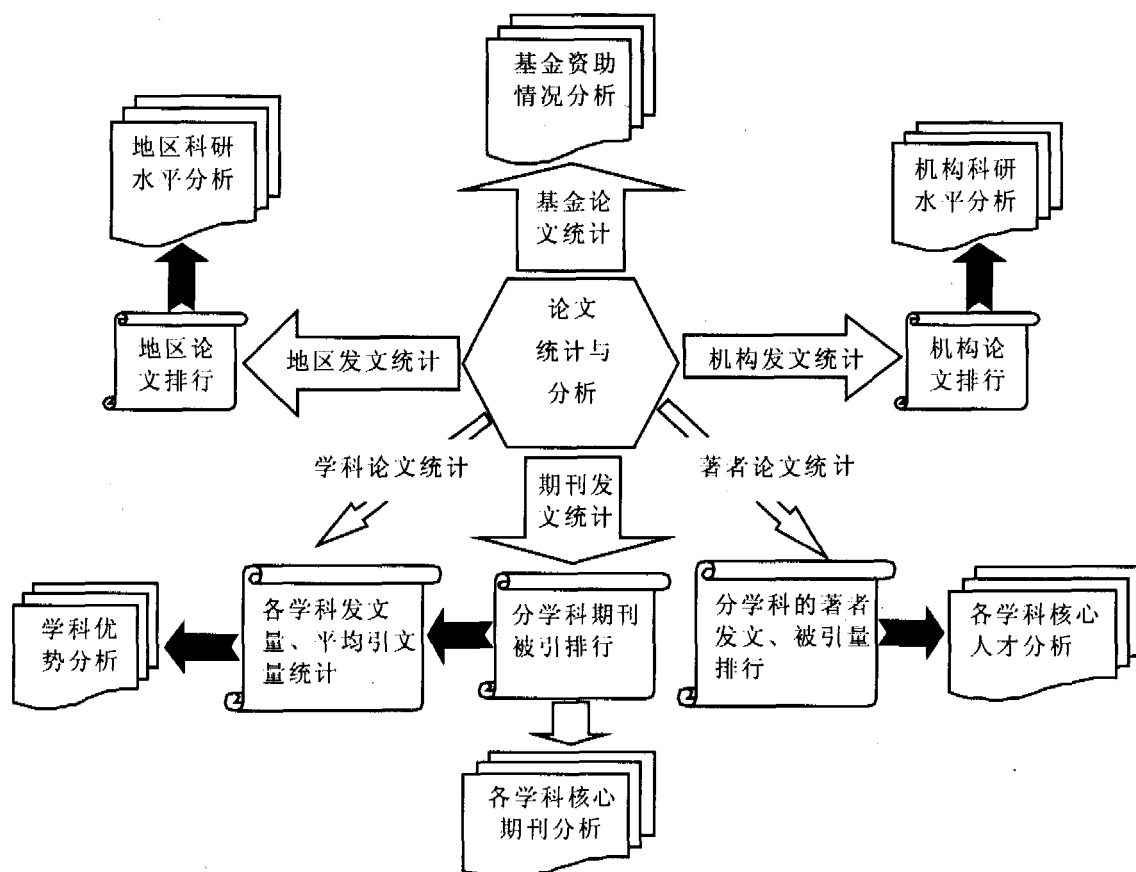


图4 论文引文统计分析中6个模块

到它的完整题录信息或全文。还可以看到它的被引次数,点击它的“被引次数”热链,同样也可以了解哪些论文引用了它的这篇论文,通过层层深入挖掘,使研究人员及时了解某一研究领域目前的进展状况和发展方向,跟踪国内外同行或权威人士的研究动态。随着更多的文献加入到数据库中,被引次数还会不断增加,所显示的每条引文文献记录都可以被层层激活。

点击“中文参考文献数”热链,就会显示当前来源记录所引用的中文参考文献的列表,列表中所有带 HTML 或 PDF 的记录也是万方数据的来源记录,可以被激活。作者引用的参考文献,诸如:图书、期刊论文、专利和其他文献都可以作为检索途径。参考文献列表的每条记录后,还可以看到它的中文参考文献数和被引次数,如此深入探究,便可以帮助研究人员了解某一研究课题的发展历史,揭示研究者在吸取前人研究成果时的去向。另外,那些没有标注 HTML 和 PDF 的引文记录,是未被万方数据收录,而无法被进一步点击查看。目前中刊引文链接率达 40%,即参考文献中有 40% 以上的中文期刊文献记录,同时也是万方数据收录的论文全文。

点击“相关论文数”热链,可以查看在不同年份中与当前所检索的记录共同引用同一篇或几篇参考文献的一组论文,即相关论文,并按共同引用参考文献的多少(即相关度)排序,和当前论文引用的相同文献越多,该文献与当前论文在主题上越靠近,因而该文献在列表中的位置就更靠前。“相关论文数”这一功能是快速高效的扩检手段,相关论文记录带有 HTML 和 PDF 的,可以被激活,揭开了研究课题之间的相关性,为研究某一课题提供具有连贯性的、相互关联的资料。

通过上述 3 种论文引文间的关联性,可以在其检索结果中所涉及到的论文资料里实现层层揭示,无限地浏览,帮助研究人员迅速找到所需信息。

(2) 专项检索:整个系统提供了从期刊、作者、主要机构、基金等,对某一特定对象的发表文章情况和文章被引用情况进行专项查询的功能。检索某特定作者后,点击查看“被引情况”数值,一次性查出某一作者的所有文章被引用情况。会按时间序列显示该作者历年来所发表的每一条著作记录,记录后带有 HTML/PDF 的,同时是被万方数据收录的来源记录,可以被激活。在每一条著作记录下,都有引用该

文章的论文列表,可以点击查看来源记录的具体内容,予以核实来源文献是否真实地引用该作者的文章。同时还显示了该作者发表文章总数以及总被引次数。

点击查看“发文情况”数值,会按时间显示该作者发表的论文记录,每条记录后如果显示中文参考文献数和被引次数,则可以具体查看,而层层浏览。同样,可以查某一种期刊在某年度发表的全部文献和某一种期刊某年度发表的全部文献的总被引情况,便于掌握期刊的全貌和对期刊来源文献进行计量分析。

在每条论文记录的具体内容中,既包含了记录的基本信息如题名、作者、作者机构、关键词、分类号、文摘、出处等字段的具体内容,还列出了来源记录的总参考文献数、中文参考文献列表、被引次数和被引论文列表,以及相关论文列表及其相关度。

(3) 指标查询:整个系统提供了期刊的文献计量指标,可以从期刊入手进行查询,同时按照期刊的学科分类,可以在某年份中,分指标进行排行榜。文献计量指标包括:载文量、总被引频次、影响因子、即年指标、被引半衰期、平均引文数、基金论文比、他引比、扩散因子等。当看到某一指标值时,如继续点击“总被引频次”值,便会显示出具体每一次被引用情况的论文与引文记录的对应关系列表。点击论文记录可以从各自的具体内容中得到进一步核实。为科研绩效的评价提供了科学的量化依据。

同时,系统还将提供作者、主要机构如大学、研究所、企业研发机构等和基金资助的发表文章和被引次数的检索和排行。

(4) 原文链接:随着越来越多的学术期刊走向Web,即时地提供全文将是信息服务的发展方向。中文知识链接门户的全文链接大多是双向的,既可以从论文直接授权链接调用全文,也可以从所引用的参考文献连接获取该被引文献的相关信息和全文内容。这样,经由这一枢纽,由不同编辑部提供的电子版全文文献通过相互引证的关系有机地联系在一起,提供一站式服务。

5 中文知识链接门户的扩展功能

Web技术的超链接特性与引文索引所特有的优势相结合,为建立以知识为基础的学术信息资源体系奠定了技术基础。中文知识链接门户的建设构筑了中文信息资源整合的逻辑平台,开创了一条中文

信息资源整合的模式和方法。要实现对中文知识资源的更广泛集成和更深入整合,真正形成中国的Web of knowledge,更需要以期刊引文为核心,进一步发扬共享理念,开放式发展,与其他的内外部学术资源进行整合,建立跨库链接机制,在充分发挥不同检索工具本身优势的基础上,揭示更多、更丰富的信息,并拓展评价分析功能,深化个性化、人性化服务理念,从而构成一个以知识为基础的既集中又开放的中文资源整合平台与分析工具。

5.1 进一步追溯过刊资源,实现新老文献资源的整合

为了更好地反映科学发展的轨迹,揭示旧文献对新文献的影响,新文献对旧文献的反馈,反映新旧文献在不同时期的发展特点和科学研究之间承前启后、前因后果的内在逻辑关系,使科技人员不仅可以追踪最新的研究进展和动态,更可以迅速回溯某一研究课题的历史性记载。需要更进一步地追溯过刊信息资源,逐步回溯加工历史的期刊论文,并不断拓展社科类期刊,形成更宽广的知识链接系统,反映现代科学发展的历史全貌和跨时域特征。

5.2 实现不同类型文献源的双向链接,拓展文献源间的内在联系

中文知识链接门户需要囊括各类型的资源体系,通过期刊引文中包含的各种类型文献,建立起包括期刊、专利、学位论文、标准文献、会议录等在内的多种类型文献之间的相互引证、相关参考的关系,实现对拥有使用权限的全文文献的链接,全方位地为科学研究提供文献信息保障,使科研工作者得以了解与其研究领域相关的各种类型文献,以及学科过去、现在和将来的发展脉络与交叉。

5.3 创建科学评价工具,实现二次文献与事实性数据库的对接

在中文知识链接门户中,汇集和分析学术文献所引用的参考文献,可以以作者、机构为主线,分析高被引频次的作者、单位和论文,建立“基本科学指数数据库”,分析各个领域学术研究的热点、影响和趋势。从中了解达到一定级别的科学家、研究机构(大学)、地区(城市)和学术期刊在某一学科领域的发展和影响。并且可以建立二次文献与事实性数据库之间的有效链接机制,使之与中国科技名人库、中国科研机构库相沟通。

5.4 利用 OpenURL 的链接机制,实现与馆藏资源(特别是英文馆藏)的链接,以及其他国内外数据库和出版社资源的整合

方便而深入的链接是实现集成检索的重要保障,链接的深度与广度直接关系到检索的效率。利用 OpenURL 标准和上下文相关的参考链接系统,可以有效地建立与所求资源之间的链接,实现不同供应商的或不同平台上内容关联的数据库间的相互链接,可统一检索不同网址上的多个数据库或信息资源,避免因网址改变或网络阻塞等故障而导致的“死链接”。中文知识链接门户的拓展需要本着开放合作的理念,加强与有关出版单位和数据库单位的合作,采用预先权限认证的连接技术,让用户直接链接到出版商提供的电子版全文。与数字图书馆单位联合,解决与图书馆馆藏目录系统的无缝连接,方便读者找到该文献所在期刊的馆藏记录。逐步形成以出版商及其全文数据库为基础的分布式、开放式和整合统一的数字化学术信息资源体系,为集成化检索的展开建立一个良好的运作环境。

6 结束语

人类的知识原本就是一个相互联系的整体。但是,我们使用的各种数据库都是以一种零散的、孤立的状态存在着,即使若干个库捆绑在一起,也仅仅局限在使用同一界面的层次上,体现不出文献内在的相互联系。中文知识链接门户的建设,利用论文之间相互引证的关系,建立起不同类型资源之间的关系,使之成为一个有机的整体,可以消除由于数据库

收录范围有限或数据库归属系统单位不同而造成的知识体系的割裂,最大限度地保持知识体系的完整性。

学术论文之间相互引证的关系最自然、有效地反映了学术研究之间的内在联系。中文知识链接门户所构造的引文检索机制将各个不同学科领域内,对于某一课题的相关研究成果轻而易举地揭示出来,方便用户掌握各种不同学科、不同领域相关研究的交叉与互动,将为科学研究的立项、规划、发展和深入提供知识层面的信息资源。同时也为我国知识资源整合、数据库集成和信息共享找到了一种有效的模式。

参 考 文 献

- 1 万方数据知识链接门户 <http://www.sci.com.cn>
- 2 范爱红,姜爱蓉.基于知识管理的学术信息资源整合体系.现代图书情报技术.2001(6):43~46
- 3 夏立娟,陈陶.从 Web of Science 的检索流程看引文索引的功能实现途径.图书馆建设.2003(4):75~77
- 4 蒋凌慧.Web of Science 数据库中的特色链接.现代图书情报技术.2001(4):44~45
- 5 郭丽芳.中外五大引文索引系统比较分析.现代图书情报技术.2005(1):36~39
- 6 郑德俊,叶继元.基于合作模式的引文数据库发展策略.大学图书馆学报.2005(1):79~83
- 7 潘有能,纪蔚蔚.中国警察科学引文索引系统的设计与实现.情报学报.2004(6):703~708

(责任编辑 王建平)