

NSTL文献信息加工10年概述

□ 孟连生 张建勇 刘筱敏 / 中国科学院国家科学图书馆 北京 100190

摘要: NSTL是中国最具影响的面向全国提供服务的公益性科技文献保障系统, NSTL的文献数据库建设和文献信息加工规模庞大, 影响深远, 构成了其文献信息保障系统建设的重要组成部分和信息服务工作开展的前提。文章概要地回顾了NSTL自建文献信息加工的发展历程, 从加工的类型、数据质量控制和加工平台等角度介绍了相关工作的发展概况。

关键词: NSTL, 文献信息揭示, 数据加工

DOI: 10.3772/j.issn.1673-2286.2010.10.005

国家科技图书文献中心(以下简称NSTL)成立于2000年6月12日。经过近10年的发展, NSTL已经发展成为中国最具影响的公益性科技文献保障系统, 其网络服务系统面向全国科技界用户, 提供文摘和引文数据检索、全文文献传递、网络版科技期刊全文利用、网络信息热点门户、开放获取期刊导航、网上参考咨询等多层次的文献信息服务。NSTL订购、收集和提供服务的文献类型涉及国内外科技期刊、会议文献、学位论文、科技报告、专利、标准、计量检定规程, 以及网上类型多样的文献信息资源等。为揭示和报道这些文献信息的内容, NSTL十分重视开展文摘、引文等文献数据加工和网络资源的搜集报道工作。

NSTL成立以来, 提供服务的文献数据库类型不断增加, 数据量快速增长, 截至2009年底, 网络服务系统提供检索和查询服务的二次文献数据量超过1.18亿条, 是系统开通之初的70倍。从文献书目、期刊目次、文章题录、文摘、引文和全文等不同层次, 全方位地揭示报道NSTL订购或收集的各类文献信息资源。提供服务的文献信息数据库的一部分是购置引进的, 另一部分则是由NSTL组织加工的, 后者由于与NSTL馆藏资源更为贴切, 用户可以通过检索结果直接请求文献原文, 因而在NSTL服务中发挥着十分重要的作用。本文主要介绍NSTL自建文献数据库及对网络信息进行加工揭示的有关情况, 力图从一个侧面较为客观地勾勒出NSTL成立10年以来的发展历程。

1 文献信息加工的主要类型

NSTL的文献信息加工工作, 伴随NSTL的诞生而诞

生, 伴随NSTL的发展而发展, 其原则是“统一标准、分布加工、集中建库、分布服务”。文献数据加工工作由9个成员单位共同承担, 各单位按照统一的信息揭示标准, 进行各种类型的文献信息加工, 再以集中建库的方式, 将二次文献数据及时加载到NSTL网络服务系统相应的数据库之中, 供用户查询检索文献信息。NSTL的服务部门根据用户的需求, 通过网络向用户提供文献全文传递服务。在这一过程中, 文献信息加工可以说是组织文献资源、揭示文献内容的一个重要的手段, 是NSTL科技文献信息资源与服务体系建设的重要环节, 对于促进文献信息资源的开发利用具有非常重要的意义。NSTL加工的文献类型主要包括期刊论文、会议论文、学位论文、科技报告、标准等, 加工规模庞大, 每年加工的文摘和引文数据量数以千万条计。随着用户获取信息渠道的变化, NSTL也十分重视对网上科技文献信息资源的搜集、加工和揭示工作。

1.1 文摘数据加工

鉴于文摘在揭示文献内容方面所具有的重要作用, NSTL始终将文摘加工, 特别是西文期刊文献的文摘加工作为其文献加工工作的核心内容。得益于NSTL一些成员单位具有多年的数据加工传统、丰富的数据加工经验和较强的数据加工能力, NSTL成立之初, 很快就在各单位组织起数据加工队伍, 有计划地加工本单位订购的西文科技期刊和西文会议文献的文摘数据。2002年之后, 又分期分批回溯加工了1998-1999年和1995-1997年出版的、各单位收藏的西文期刊论文的文摘数据。为了加强对文集汇编类文献的揭示与利

用, 2004年NSTL启动了馆藏西文文集汇编文献的文摘加工工作。

NSTL收藏的日、俄文期刊约占其订购外文文献总量的20%。2004年NSTL组织开发了相应的加工系统, 开始进行日、俄文期刊的文摘数据加工工作。同时, 在网络服务系统中建立了相应的文摘数据库, 为国内学者查询和利用日、俄文文献提供了便捷的条件。NSTL是国内唯一的一个通过互联网提供日、俄文科技文献文摘检索服务的机构。此外, NSTL还以文摘的方式组织加工所订购或收集的外文科技报告、中文学位论文、西文学位论文、西文文集汇编、计量检定规程, 以及国内外标准等文献。



图1 数据加工车间一角

1.2 引文数据加工

引文数据库是一种有别于传统文献检索方法的、基于引文索引法的文献数据库, 可以有效地把不同来源、不同类型的信息资源组织在一起, 方便用户查询和利用与研究主题相关的各类型文献。同时, 引文数据库还可用于文献分析和科技活动评价等, 是一种十分重要的文献数据库类型。

经过一年多时间的可行性论证, 2006年NSTL正式启动了国际科学引文数据库(Database of International Science Citation, 简称DISC)建设, 同年10月DISC建设被列入国家科技支撑计划项目。在深入调查研究的基础上, 自主研发引文数据加工和引文检索服务系统, 确定了DISC数据加工流程, 制定了相关标准规范, 添置了加工设备, 引文数据加工工作全面展开。综合考虑各方面因素, DISC首批来源期刊的数量确定为1,000种左右。2007年, 完成了DISC检索服

务系统的开发, 累计加工引文数据1,000余万条, 实现了引文数据库第一阶段建设目标。12月6日, 在NSTL二届理事会四次会议上, 师昌绪理事长点击开通了DISC试用系统。2008年, 为增强引文数据库的服务能力, DISC来源期刊品种数增加到2,500种, 提供服务的引文数据量增加到2,400余万条。同时, 对DISC试用系统进行了升级, 提高了引文数据的归一化精度, 用户界面更为简洁、清晰、友好。2009年, 为了适应引文数据大规模增长的发展趋势, 对引文数据库服务器的系统结构进行了全面调整升级, 同时进一步改进和完善了系统的服务功能。新系统采用数据库集群、分布式索引等系统框架, 调整了数据结构, 实现了来源文献和引文文献的归一, 增加了国际通用的访问统计功能, 建立了数据可视化模型, 体现了检索过程方便快捷、服务功能个性化和检索结果可视化等特点, 数据处理和服务功能均达到国际先进水平。新的国际科学引文服务系统于2010年6月正式上线。2010年, 系统提供服务的数据总量达到4,800万条, 系统访问量逐月攀升, 服务效果显著提升。DISC来源期刊品种数将增加到3,000种, 一年产生的引文数据量将超过3,000万条。

开发建设具有自主知识产权的国外科技期刊引文数据库, 为用户提供国际科学引文检索服务, 标志着NSTL在提供深层次、知识化服务方面又迈出了新的一步, 不仅可以在很大程度上增强NSTL馆藏资源的利用效果, 而且可以逐步减轻国内科技信息机构和广大用户对国外相关数据库的依赖程度, 从整体上减少国家对订购相关资源的经费投入, 具有长远的社会效益和广泛的应用前景。

1.3 热点门户建设

2002年, NSTL在服务系统中开辟了网络科技信息资源指南栏目, 大量搜集国内外科技信息资源网站网址, 为用户查询链接相关网站提供方便。2004年, NSTL组织对网上近2,000种医科领域的免费全文电子期刊进行了整理和揭示, 受到用户的欢迎。为了进一步拓展网络信息揭示服务, 同年NSTL启动了科技热点门户建设, 以门户网站的方式, 围绕国内学术界关注的热点科技问题, 组织对互联网上相关文献信息的搜集与报道。报道的科技信息类型包括专业文献信息资源、科技动态、专业新闻和机构信息等。至2007年

底,先后开通了纳米科技、认知科学、食物与营养、艾滋病预防和控制等17个热点信息门户网站,累计搜集报道相关信息10万余条,为用户了解和利用网上有价值的专业信息资源提供了有益的引导与帮助。

1.4 网上开放获取期刊的搜集与加工

近年来,国际上开放获取运动发展迅速,开放获取的服务模式为用户提供了一种全新的、无障碍的学术信息交流渠道。国外相继出现了DOAJ、Open J-Gate、PubMed Central等开放获取期刊数据库,开放获取资源日益受到学术界的广泛重视。开放获取文献的传播,缓解了学术期刊及数据库价格逐年上涨和文献出版时滞给学术交流带来的不便,促进了学术信息的广泛交流与共享。

认识到这种新型文献发布形式对于信息传播与利用的意义,2008年NSTL开始组织对网上开放获取科技期刊进行搜集整理、加工描述和集成整合等处理,建设开放获取科技期刊集成揭示平台,为用户提供针对互联网上各类型开放获取期刊检索和链接的一站式服务。2009年10月,NSTL“开放获取期刊集成检索试用系统”开通服务,搜集报道4,000余种开放获取期刊,学科范围涉及理、工、农、医各个领域。系统提供期刊名称字顺和学科分类两种浏览方式,具有刊名、ISSN、主题、出版者及期刊其他元数据信息的检索功能,为用户了解开放获取期刊的发布信息、开放获取资源的利用与获取方法提供了便捷渠道。

2 数据加工管理与数据加工质量控制

在文献数据库建设和数据加工方面,NSTL先后颁布了《国家科技图书文献中心数据库建设管理办法》、《国家科技图书文献中心数据加工工作管理办法》、《国家科技图书文献中心数据加工工作组工作条例》等多项规章制度,从宏观管理的角度规范此项工作的健康发展。NSTL设有数据加工工作组,全面负责数据库建设的组织与协调,推进数据加工工作的开展。

2.1 数据加工质量与时效

数据质量是数据库的生命,直接影响到数据库的利用效果及网络化文献服务的质量。NSTL在数据库建设和数据加工过程中十分重视数据质量控制工作,

2003年颁布了《国家科技图书文献中心文献数据加工质量监督管理试行办法》,2007年颁布《国家科技图书文献中心数据加工质量控制管理办法》,取代前面的试行办法。NSTL在组织加工数据的初期,便设立了数据质量控制小组,建立了三级数据加工质量控制体系,开发了数据录入软件和质量控制软件。每年针对数据加工中存在的问题,组织形式多样的质量检查,提出数据加工质量分析报告,积极落实整改措施,不断提高数据加工管理水平和数据加工质量。

2008年,为配合网络服务系统升级,NSTL开始组织对历史数据进行全面清洗,系统地检查和修订自建数据库中存在的各种加工质量与规范问题。截至2009年底,总计清洗各类型文献的文摘数据5,000余万条,从整体上大幅度提升了NSTL在线服务的各类型数据库的数据质量。历史数据清洗是一项十分复杂而艰巨的工程,在大约2年的时间里,完成数量如此庞大的文摘数据清洗,在国内外同行中十分罕见。

提高数据加工的时效性,一直是数据加工管理的重要内容。为此,NSTL为各单位配置了优质高速的扫描仪和文献数字化加工系统,文摘数据的加工时差要求由最初的文献到馆后2个月,逐步缩短到2周时间。文摘数据上传频率也由最初的每月报送一次,提高到每周报送一次。文摘数据的及时更新,使用户能够在文献出版后较短的时间内检索和利用最新文献。

2.2 数据加工规范

NSTL十分重视数据加工的标准化和规范化建设,2001年制定了《国家科技图书文献中心文献数据加工细则》,规范和约束数据加工的格式和内容。2004-2007年,NSTL组织协调的“我国数字图书馆标准规范建设项目”,先后推出了与数字图书馆建设相关的一系列研究成果,其中多项标准涉及到NSTL建设的数据资源。2007年,为进一步提高数据库建设的标准化和规范化水平,NSTL启动了文献数据加工细则的修订工作,遵循数字图书馆标准规范项目的相关标准和其他研究成果,从数字图书馆资源组织和现代网络信息服务的角度,重新审视、修改和细化NSTL的数据加工规范。

经过两年多时间的努力,凝聚了各单位数据加工人员智慧的《NSTL文献资源加工规范》于2009年11月由知识产权出版社正式出版。该规范充分考虑元数据

创建者、元数据管理者和元数据使用者各方的多层次需求,参考DC元数据的组织体系,结合资源对象特征,以业界通用的XML作为数据交换格式。规范的体系结构具有开放性和可扩展性,符合当前网络化文献数据加工的要求,达到国内领先水平。促进了国内本领域学术研究的深化与拓展,对其他单位或系统的文献数据加工工作具有参考或借鉴作用。

2.3 数据加工系统

NSTL的文献数据加工系统是多年来逐步部署开发的,其早期使用的二次文献加工系统、数据质量检查系统和用于集成各单位不同来源数据的数据转换系统,都是2000年前后开发的。目次扫描加工系统开发于2004年,以Web方式上传目次数据。不同类型数据的加工处理,以及对不同类型数据的质量控制都分散于不同的系统之中,缺乏统一的、可以相互配合、集成管理的数据处理环境,从而造成数据质量控制能力较弱、各类型数据一致性较差、数据加工和管理成本较高等问题。

2007年,NSTL启动了联合数据加工系统的研制工作,目的是以中心千兆城域网为基础,建设一个集数据加工、质量检验、过程管理于一体的集成数据加工平台,所有成员单位在同一个加工平台上加工处理各类型数据。联合加工系统由中心和分中心两级管理系统组成,于2009年投入试运行。该系统支持对各单位数据加工工作的集成化、协同化管理,支持对NSTL订购资源、第三方资源、免费网络资源等多种类型数据资源的导入和整合,支持对题录、文摘、引文等不同类型数据的加工处理,有效地提高了NSTL数据加

工的管理与质量控制水平。作为NSTL业务管理流程的重要环节,在NSTL整个业务工作体系之中,该系统前后分别与联合编目系统和仓储系统相衔接,采用OAI数据收割方式传输数据,构成了一个从文献采集、文献编目、登到加工、文摘、引文加工到文献数据仓储及发布服务的、具有国际先进水平的、网络化的综合业务管理体系,从根本上提高了NSTL的网络化业务管理水平。

3 结语

10年来,NSTL数据库建设及数据加工工作走过了一条不平凡的艰苦创业之路,取得了一定成绩,但是与国际先进水平相比,仍然存在较大差距,还面临艰巨的任务。当今世界,国内外各种类型文献信息机构的服务方式和服务手段日益更新,新的检索技术与知识挖掘组织系统不断涌现,具有个性化和知识化服务特征的信息系统迅速发展。智能化、个性化和知识化的服务功能,大多依赖对文献信息的深层次挖掘、规范化加工和知识化组织而实现。今后NSTL数据库建设和数据加工工作不仅要进一步推进科学化、规范化管理,全面揭示各类型馆藏文献资源,不断提高数据加工的效率和质量,而且要加强应用国际最新数据加工与知识组织技术方法,全面提升对网络信息资源的整合揭示能力,加强对网络信息资源的集成报道能力,为用户提供更加便捷的文献信息查询、获取与利用环境。回顾历史,成效显著,展望未来,任重道远。预祝NSTL数据加工工作不断取得新的更大的成绩,为NSTL向广大用户提供特色化、智能化和知识化服务奠定良好的数据基础。

作者简介

孟连生,中国科学院国家科学图书馆研究馆员,学术委员会主任,国家科技图书文献中心数据库建设部主任。通讯地址:北京中关村北四环西路33号100190。E-mail: menglsh@gmail.com
张建勇,中国科学院国家科学图书馆研究馆员,资源建设部主任,国家科技图书文献中心数据加工工作组组长。通讯地址同上。E-mail: zhangjy@mail.las.ac.cn
刘筱敏,中国科学院国家科学图书馆研究馆员。通讯地址同上。E-mail: liuxm@mail.las.ac.cn

Outline of Document and Information Processing by NSTL for the Past 10 Years

Meng Liansheng, Zhang Jianyong, Liu Xiaomin / National Science Library, Chinese Academy of Sciences, Beijing, 100190

Abstract: As a public welfare agency and one of the most influential scientific and technical document service systems in China, NSTL is offering information services to its clients all over mainland China. As an important part of its document and information Security Systems and a prerequisite of information services, the task of database construction and information process of NSTL is of a large scale and has a far-reaching influence. The brief history of database construction of NSTL was recalled and the types of data processed, the data quality control and data processing platform were introduced in this paper.

Keywords: NSTL, Document and information retrieval, Information data processing

(收稿日期: 2010-08-30)