

基于元数据仓储构建科技文献共享服务平台

王 飞 殷 铭 高 璇

(江苏省科学技术情报研究所, 江苏南京 210042)

摘要: 随着互联网的普及与发展, 各种信息资源铺天盖地而来, 信息资源结构以及对资源结构的需求也在发生着变化, 传统的文献信息服务已经不能完全满足用户的需求。通过对现有的异构整合技术进行分析比对, 提出基于元数据仓储构建区域科技文献共享服务平台, 结合平台的建设实践, 探讨元数据仓储知识库的构建, 并对平台的功能模块设计、实现以及应用成效等进行详细说明, 最后指出科技文献共享服务平台未来的建设和发展方向。

关键词: 元数据; 仓储; 科技文献; 共享服务平台; 科技文献共享

中图分类号: G250

文献标识码: A

DOI: 10.3772/j.issn.1674-1544.2015.01.009

Construction of Science and Technology Document Sharing Service Platform Based on Metadata Storage

Wang Fei, Yin Ming, Gao Xuan

(Jiangsu Science and Technology Information Institute, Nanjing 210042)

Abstract: With the popularity and development of the Internet, various information resources are overwhelming. Information resources structure and the demand for resources structure are changing. The traditional information service has been unable to fully meet the demand of users. Through the comparison of existing integration technologies, the way of building science and technology document sharing service platform based on metadata storage is proposed. Combined with the practice of science and technology document sharing services platform construction, the construction of metadata storage is discussed. The design and implementation of the platform are detailed described. Finally, the development trend of science and technology document sharing service platform is put forward.

Keywords: metadata, storage, science and technology document, sharing service platform

1 引言

科技文献资源是国家宝贵的战略资源, 是创新发展的重要前提和保障。推进科技资源共享作

为建设创新型国家的重要组成部分, 已成为提升科技、经济社会发展“软实力”的重要支撑^[1-4]。

我国十分重视科技文献资源的开发和利用, 早在《2004-2010年国家科技基础条件平台建设纲

作者简介: 王飞* (1976-), 女, 江苏省科学技术情报研究所副研究员, 研究方向: 科技文献共享与服务平台建设、科技情报研究等; 殷铭 (1969-), 男, 江苏省科学技术情报研究所研究馆员, 研究方向: 科技信息资源建设、信息平台开发与服务工作; 高璇 (1983-), 女, 江苏省科学技术情报研究所助理研究员, 研究方向: 科技咨询、科技情报研究。

基金项目: 江苏省科技基础设施建设计划 (BM2004101, BM2004714, BM2010611)。

收稿日期: 2014年6月5日。

要》中明确提出要重点建设六大平台,以促进全社会科技资源高效配置和综合利用。“科技文献共享平台”就是其中之一。目前,在全国31个省、自治区、直辖市中(不包括港、澳、台),除北京外其余均建立了省级科技文献共享平台^[1],西安、杭州、南宁、宝鸡、常州等地区也建立了市级科技文献共享平台。现已初步建成了国家、省、市三级科技文献共享平台。为科技创新、经济发展提供了强有力的资源保障^[1]。

江苏省工程技术文献信息中心(简称“中心”),通过集成、整合社会优质资源,以少量的杠杆式引导资金,调动了江苏省科技、教育、文化三大系统10家单位的存量文献信息资源,在全国率先建成省一级跨部门、跨系统的共享共建科技文献信息服务体系^[5]。但随着现代科技信息服务业的深化和平台建设工作的推进,原有平台的功能架构和服务体系已经无法满足用户日益增长的需求,主要体现在以下几个方面:一是平台服务用户数出现了级数增长,资源品种也在不断变化,用户对平台的服务水平和服务层次提出了更高的要求;二是国内各文献资源平台在建设初期大多采用基于页面分析的网页搜索代理模式,虽然实现了跨库统一检索,但是存在着检索速度慢、维护复杂、数据著录缺乏统一标准、难以扩展等弊端,对于外网用户使用数据库的知识产权问题也难以解决;三是在当前建设创新型国家的大环境下,政府、科研院所以及企事业单位的科技人员更加关注能否从海量级资源中高效、便捷的获取最有价值的知识服务,以改变资源超载与知识饥渴并存的现状。

因此,在上述需求的推动下,迫切需要对原有平台进行升级改造。引入元数据整合技术,并将知识服务的理念融入到文献资源整合中来,在此基础上集成互联网采集、文本挖掘和热点发现技术,开发基于元数据仓储的科技文献共享服务平台,从而在实现资源统一检索的同时,进一步扩大检索范围、提高查全率、提升检索效率,为实现传统文献服务向知识服务转变提供强有力的资源支撑,为实现知识深层次挖掘与分析提供依

据。本文将重点论述新构建的江苏科技文献共享服务平台的结构、功能及其成效。

2 元数据仓储知识库

江苏省工程技术文献信息中心是由10家单位共同建设以实现资源共享与服务的。由于各单位订购资源之间缺乏关联再加上资源商的服务范围限制,无法实现资源的充分共享与利用。为了解决此问题,中心启动了二期改造工程,通过采用元数据整合技术,将来自于不同系统的异构资源通过元数据规范标准进行采集、整理与加工,构建元数据仓储知识库。

2.1 元数据整合技术

传统的文献资源平台大多采用基于页面分析的网页搜索代理模式,即通过对不同检索系统的页面进行特征分析,提取检索结果的跨库检索方式。元数据是关于数据的组织、数据域以及关系的信息,也就是“关于数据的数据”^[6-9]。元数据整合技术通过制订统一的元数据规范,屏蔽系统之间的差异,可以实现对分布在不同物理存储空间中的多种资源元数据的采集和集中管理,并通过对抓取数据的转换、清洗、质量检查、数据抽取和深度标引构建海量的仓储知识库^[9]。通过对上述两种整合技术进行对比(表1)不难发现,基于页面分析的整合技术虽然实现了跨库统一检索,但是存在着检索速度慢、维护复杂、数据著录缺乏统一标准、难以扩展弊端。而基于元数据的整合技术相对页面分析技术来说,通过定义资源元数据规范,将资源进行有序整合、集中式管理,从而有效地解决了跨库检索系统检索速度慢、维护复杂等问题,同时由于其具有数据集中存储、检索效率高等优点,在用户统一身份认证、资源的二次开发以及深层次文献服务等方面更具优势,能够为用户提供更高层次的知识挖掘和分析服务,将大大提升平台的资源利用率和服务效果。

2.2 元数据仓储的架构

元数据仓储知识库是基于统一的元数据规范构建不同资源的元数据库,通过对分散的、异构

表1 页面分析整合技术与元数据整合技术对比

技术指标	页面分析整合技术	元数据整合技术
检索效率	数据组织松散，结构不统一，受原检索系统约束，检索效率低	数据集中管理，结构统一，标准规范，检索效率高，速度达毫秒级
资源利用情况	由于数据分布式管理，检索结果重复率高	数据整理加工时清洗，去重去噪，查全率、查准率高
数据更新情况	与原有系统同步	依赖元数据加工频率
服务拓展	文献检索，原文获取	统一检索，用户统一认证，文献二次开发，深层次文献服务

数据的整合，屏蔽系统之间的差异，满足对元数据的存储需求，实现元数据管理和共享。基于平台资源的现状，根据元数据仓储整合原理，构建了平台元数据仓储架构图，如图1所示。

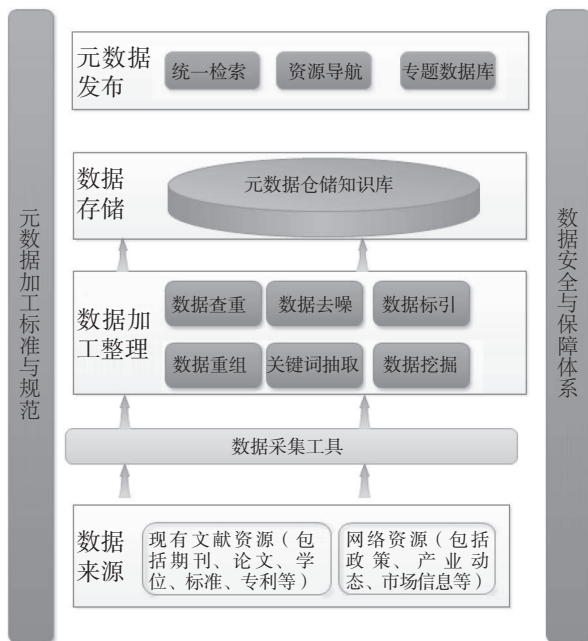


图1 元数据仓储架构图

构建元数据仓储知识库首先要做好资源分析，对现有文献资源和外部网络资源进行梳理，确定资源类型，并针对各种不同的类型明确资源的采集方法、加工模型、加工方式、采集接口以及开放协议等，建立统一的元数据规范与加工流程标准。元数据的采集与加工是仓储知识库建设的重要环节，本平台建设中采用关键词抽取、关键词分析、信息重组分类等多项技术，成功开发了数据加工采集系统工具。通过在元数据仓储知

识库中建立各个资源元数据加工规范模板，配置相应的采集参数，由元数据仓储系统自动实现元数据的采集与存储。采集过后的数据通过去重去噪、自动标引等功能完成二次加工后再进入元数据仓储知识库。

本平台以元数据仓储技术为核心，整合各类资源，形成了满足江苏省需求的工程技术文献信息资源体系，数据量达1.7亿条（截至2013年12月），国内外重要工程技术文献中文满足率达90%以上。基于元数据整合模式的跨库检索解决了原数据库系统资源记录孤立、关联度不高、用户使用时难以形成知识间联系网络等问题，同时元数据的使用仅涉及文献的题录文摘等表征数据，不侵犯知识产权，而原文则采用传递等国内外常规方法解决，从而做到在不影响知识产权的情况下将国内外优质数据库资源整合起来对用户开放服务。

3 平台的设计与实现

元数据仓储知识库的构建为实现科技文献共享服务平台提供了基础保障。在此基础上集成WEB2.0、AJAX、WebService、RSS、中文自动分类等技术，应用门户系统把各种应用系统、数据资源和互联网资源统一集成到通用门户之下，根据每个用户使用特点和角色的不同，形成个性化的应用界面。通过对事件和消息的处理传输把用户有机地联系在一起，设计建立了“江苏省工程技术文献信息中心科技文献共享服务平台”（以下简称“平台”）门户网站，建成了支持跨平台、分布式、异构数据库环境的综合性工程技术

文献检索系统和信息服务系统。其主要功能如下。

3.1 文献统一检索

平台通过构建元数据仓储知识库,对存在于不同检索系统的资源根据元数据规范进行数据处理、关联整合,经过元数据分析,构建了期刊、学位论文、会议论文、专利、图书、科技成果等11个元数据框架。通过抽取基本相近的元数据,生成统一元数据框架,以满足用户对所有数据资源统一检索功能,提高文献检索效率以及查全率。在此基础上,开发了全文检索服务系统,采用开源的全文检索引擎Lucene作为全文检索引擎框架,实现了类似百度、Google的检索界面,简单易用。通过对元数据进行分类处理,将同一类别的元数据进行统一分类,实现了基于年代、学科、资源类型等聚类分析及快速的资源导航功能(图2)。

3.2 用户统一登录

通过对资源元数据的整合,将各系统用户进行统一管理,开发用户统一认证模块,以用户管理为纽带,将平台中包括万方知识服务平台、维普信息资源系统、中国知网(CNKI)、国研网在内的文献检索系统集成成为一个有机整体,实现了平台用户集中管理与认证,用户相互认证,极大地方便了用户的使用(图3)。相对国内一般的文献平台必须进入各自登录系统,实现了登录一个平

台就能快捷、高效使用不同网络环境、不同数据结构的文献资源,解决了不同系统用户名、密码不统一、难以记忆的问题。

3.3 特色专题库建设

元数据仓储系统的构建大大提高了特色专题数据库建设的速度和准确率,改变了原先从不同的检索系统中组织人工批量下载再进行数据清洗与整理等一系列的繁杂过程,而且可以根据用户的服务需求个性化定制特色专题服务系统,快速形成各类特色专题数据库系统。以此为基础,根据江苏省十大新兴产业特点,建设行业特色专业网站,开发了特色专题数据库开发工具,实现行业专题数据库的数据下载、加工、组织与发布的功能,从而为用户提供了成熟的特色专题数据服务。到目前为止已建成清洁生产、科技咨询、新材料等特色专题数据库服务系统。

3.4 协同服务模块

通过元数据整合技术不仅将平台10家共建单位的优质资源进行了有效整合,而且在服务上也同时实现了整合应用。平台拥有一支由10家共建单位服务馆员组成的优质服务队伍,应用竞争性抢答与文献定向服务相结合的体系架构,建立了快速响应的协同服务系统。平台上一旦有文献请求出现,任何一家共建单位都可进行“抢答”,来竞争该请求的服务权(图4)。“抢答”机制大大提高了服务效率,传递速度具有明显优



图2 统一检索界面

势。对一些特殊文献（如标准）系统指定服务单位，有了该类文献请求之后，系统将会自动分配给该单位的服务馆员，这不仅提高了平台的服务效率，也大大保证了文献服务的质量。

3.5 NSTL接口模块

利用WebService技术与元数据整合技术，打通了与国家科技文献中心（NSTL）的接口，在全

国率先实现了国家资源与地方资源两套系统的完全融合，使得用户可以在平台中直接访问NSTL资源，实现了用户的统一管理、原文直接索取以及费用统一结算等功能（图5）。

4 应用成效

江苏省工程技术文献信息中心科技文献共

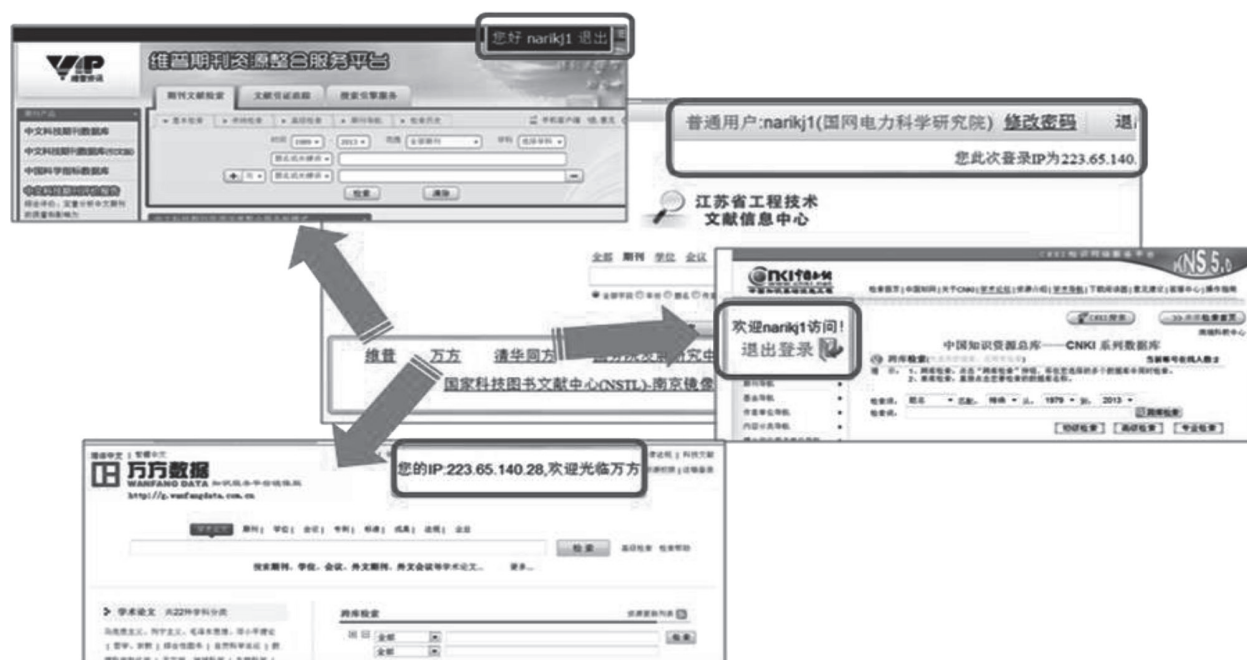


图3 统一登录实现界面

文献咨询任务->全部待处理任务

文献咨询记录查找: 邮箱 查找

共找到 1 条记录

类型	订单号	标题	产生时间	数据来源	认领任务
外文学位	141010000000205	A pattern-driven process	2014-10-10	,Proquest	认领任务

快速响应

标题	订单号	类型	读者名称	回复状态	提交时间	处理时间	单位
Postmortem MRI of b	141010000000045	外文期刊	-	已处理	2014-10-10	2014-10-10	南京农业大学
Magnetic resonance	141010000000044	外文期刊	-	已处理	2014-10-10	2014-10-10	省科情所
Pollen morphology o	141010000000009	外文期刊	南瑞公司	已处理	2014-10-10	2014-10-10	省科情所
Mutagenesis induced	141010000000008	外文期刊	南瑞公司	已处理	2014-10-10	2014-10-10	省科情所
The quinohaemoprote	141010000000007	外文期刊	南瑞公司	已处理	2014-10-10	2014-10-10	省科情所
Amorphous manganese	141010000000006	外文期刊	南瑞公司	已处理	2014-10-10	2014-10-10	省科情所
FT Raman investigat	141010000000005	外文期刊	南瑞公司	已处理	2014-10-10	2014-10-10	省科情所
Spectral characteri	141010000000004	外文期刊	南瑞公司	已处理	2014-10-10	2014-10-10	省科情所
The impact of urea-	141010000000003	外文期刊	南瑞公司	已处理	2014-10-10	2014-10-10	省科情所
Spectroscopic chara	141010000000002	外文期刊	南瑞公	未回复	2014-10-		东南大学

图4 协同服务界面

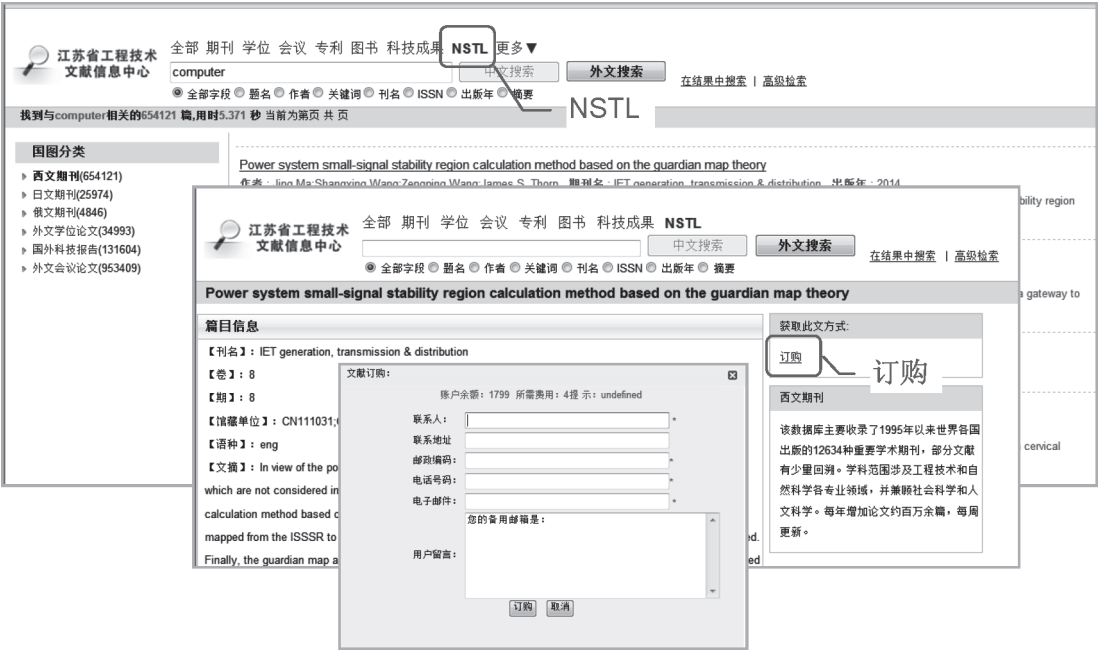


图5 NSTL访问界面

享服务平台已于2012年12月3日通过了专家组验收，总体运行平稳，各项服务指标呈现了突破性增长，如表2所示。据最新统计，2013年度平台新增元数据1600万条，实现原文传递量达364696页，新增用户数8200户，访问量突破300万次，服务用户总数达3万家，数据量和服务量再创历史新高。

可见，通过采用元数据技术对科技文献资源进行有序整合，加速了资源检索的效率，提高了

资源覆盖率和利用率，大大提升了科技文献平台的整体服务能力。2012年11月，在中国科技信息研究所组织开展的全国科技文献平台共享评估中，平台综合排名位列省级文献平台第一^[10]。2014年2月，在江苏省科学技术奖励大会上，平台荣获2013年度江苏省科学技术三等奖。

5 结语

在现有文献资源的基础上，江苏省科技文献

表2 平台改造前后指标对比情况

技术指标	平台改造前（基于页面分析技术）	平台改造后（基于元数据技术）
检索速度	秒级	毫秒级
一次性可检索数据量	3000万	1.48亿条
共享数据库量	21个	35个
年人工原文传递量	14886页（2011年）	181718页（2012年1-10月）
统一账号登录	不支持	支持
支持在线并发用户数	20人	500人
平台承接深层次文献服务量	50	1800
检索频道数	6	12
分中心建站与同步服务	没有	有

（下转第74页）

享、协作共用的和谐机制,只有坚持整体规划和科学论证,致力于平台体制机制创新,才能真正保证大型科学仪器共享平台的可持续发展。平台的创新发展需要分级分步建立大型科学仪器共享服务体系,逐步形成大型仪器资源平台制度体系,创新服务模式支持经济发展。大型科学仪器区域协作共享平台的建设,是一个循序渐进的过程,相信在各有关部门的共同努力下,大型科学仪器设备共享工作能越走越好,为学科发展、科研实力的提升和企业科技创新创造更好的条件。

参考文献

- [1] 孙绪华.关于促进我国大型科研仪器设备共享的思考[J].实验技术与管理,2006,11(23):10-14.
- [2] 孙绪华,伊彤,高鲁鹏,等.以机制创新推进大型科学仪器设备共享[J].中国科技资源导刊,2009,41(4):14-18.
- [3] 胡晓萍,钟叶龙,徐军明.大型科学仪器设备共享管理方法的研究与实践[J].实验技术与管理,2012,29(7):215-220.
- [4] 谢敏蓉.浅谈大型科学仪器设备共享平台的建设[J].技术与市场,2011,18(9):224-225,229.
- [5] 方松,赵红萍.我国大型科学仪器设备共享研究初探[J].科技管理研究,2011(2):39-41.
- [6] 关丽萍,宋皓.探索高校实验室大型仪器设备管理新模式[J].中国现代教育装备,2011(1):21-23.
- [7] 张欣欣.全国科学仪器门户平台的设计与实现[D].北京:北京邮电大学,2011.
- [8] 吴澄,刘连臣,肖田元,等.大型科学仪器设备资源的建设与整合[J].中国科技成果,2010,11(15):23-24.
- [9] 闻星火,李明,黄乐,等.高等学校仪器设备和优质资源共享系统的建设与成效[J].实验技术与管理,2006(10):6-11.
- [10] 刘连臣,吴澄,闻星火.高等学校仪器设备和优质资源共享系统的规划与建设[J].中国教育信息化,2008(11):34-36.

(下转第62页)

共享平台整合科技政策、科技成果、科技专家、大型仪器、地方科技信息、科技金融、产业动态、项目申报指南等与企业创新各个环节紧密相关的科技资源,以国内外知识管理、知识发现、知识挖掘、知识分析、知识服务等理论为指导,对整合的各类科技资源进行挖掘、整理、重组,形成科技资源共享知识库并研发出科技资源知识服务系统成为了平台未来的发展方向。目前,江苏省科学技术情报研究所已经开始了知识服务平台的相关研究和建设工作,在不久的将来,基于元数据仓储构建的中心服务平台不仅能够为科技创新提供文献保障的平台,而且能够为企业创新提供增值性、智能化的知识管理与知识服务的平台。

参考文献

- [1] 杨红婕,韩瑞平,赵隽.我国省级科技文献共享平台现状及对策研究[J].图书馆工作与研究,2013(11):48-50.
- [2] 李健,杨淇蓓.科技文献共享服务平台建设与信息服务实践探讨[J].图书情报工作,2011(增刊1):144-146.
- [3] 杨淇蓓.科技文献共享服务平台建设与信息服务探讨[J].情报科学,2011,29(9):1374-1377.
- [4] 陈晓玲,张博.云计算下的吉林省科技文献信息服务平台应用研究[J].信息系统工程,2011(12):25-26.
- [5] 孙斌,徐建民.江苏科技、教育、文化系统文献信息资源初步实现共知共享—江苏省工程技术文献信息中心共建初见成果[J].江苏科技信息,2006(6):33-35.
- [6] 辜寄蓉,苗放,王成善.基于元数据的信息共享机制研究[J].物探化探计算技术,2006,28(1):75-79.
- [7] 林瑞峰,陈平华,林锦川.面向科技管理的数据共享平台关键技术研究[J].现代计算机,2009(9):104-106.
- [8] 熊建斌,李振坤,陈平华,等.元数据技术在数据共享平台中的应用[J].微型机与应用,2010(9):13-16.
- [9] 刘军,牛争艳.科技文献共享服务平台科技创新决策分析服务系统的研究与应用[J].情报科学,2013,31(8):81-83,89.
- [10] 殷铭,高璇,王飞.江苏省工程技术文献信息中心共享服务现状及展望[J].中国科技成果,2013(15):11-13,17.