

科研人员引用科学数据的问卷调查

张英杰¹ 彭洁^{1,2} 张新民¹ 吴国雄¹

(1. 中国科学技术信息研究所, 北京 100038; 2. 武汉大学信息管理学院, 武汉 430072)

摘要: 文章认为, 科学数据引用是科技论文与科学数据之间的桥梁, 是促进科技文献与科技数据融合的方式之一。伴随着数据密集型科研模式的兴起, 科技人员在论文写作中保持科学数据的可追溯性变得非常重要, 而国内科研人员针对科学数据引用依然存在着认识不充分、引用行为不规范等诸多问题, 在对4337份调研问卷的基础上, 对当前进行科学数据引用的需求层次、障碍因素等12个问题进行调研, 提出进一步促进科学数据引用的建议。

关键词: 科研人员; 科学数据引用; 认知与实践; 问卷调研

中图分类号: G35

文献标识码: A

DOI: 10.3772/j.issn.1674-1544.2013.01.008

Awareness & Practice Survey on Scientific Data Citation from R&D Staff

Zhang Yingjie¹, Peng Jie^{1,2}, Zhang Xinmin¹, Wu Guoxiong¹

(1. Institute of Scientific and Technical Information of China, Beijing, China, 100038; 2. School of Information Management of Wuhan University, Wuhan, China, 430072)

Abstract: Scientific data citation is a bridge between S&T paper and scientific data, which is a convergence channel for both of them. With the upsurge of data-intensive science, it is becoming very important to keep trace of scientific data in S&T papers. Meanwhile, there exist some problems, such as insufficient awareness and irregular citation behaviors. Through the 4337 survey questionnaires, it does research on twelve subjects around needs hierarchy, obstacle factors, etc.. At last, it comes to the advice on how to promote scientific data citation.

Keywords: R&D staff, scientific data citation, awareness & practice, questionnaire survey

1 引言

科学数据是科学研究的重要部分, 可以通过实验、观测和模拟等方式获取。在数据密集型研究特征日益显现的今天, 开放的科学数据库不仅可以加强交流, 启发新思路, 减少重复劳动, 而且通过与期刊论文等文献资源的融合, 揭示科学本质与规律, 探索科学的奥秘。然而, 在科技论文中保持科学数据的可追溯性是非常重要的, 正确而最新的科学数据对科学研究是很有帮助的。2012年10月, 汤森路透宣布推出Data Citation Index™ (DCI, 数据引文索引)。作为汤森路透搭建的Web of KnowledgeSM信息平台上的一个新的研究资源,

数据引文索引 (DCI) 将使数据集的发现和利用方式发生革命性变化。DCI能够使研究人员在数据研究支持的文献上下文中发现、引用和查看相关数据, 从而推动对数据集和数据研究的发现、使用及归属^[1]。面对科学数据引文索引这一新事物, 国内科研人员对在论文中规范引用科学数据的态度如何? 不同学科的差异性有多大? 科研人员的引用实践中还有什么问题? 针对这些问题, 笔者以科学数据引用为切入点, 设计了面向科研人员的调研问卷, 重点调研国内开展科学数据共享、实施科学数据引用、建立科学数据引用索引的可行性, 从而为下一步建立国内科学数据引文索引奠定基础。

第一作者简介: 张英杰 (1979—), 男, 中国科学技术信息研究所博士后, 研究方向: 科技信息资源管理。

基金项目: 国家科技支撑计划项目“电动汽车技术预测与决策支持系统开发”(2013BAG06B00)。

收稿日期: 2013年1月10日。

2 问卷的设计与回收

本问卷主要针对在科研过程中可能涉及科学数据的各学科科研人员,调研其对开展科学数据引用的态度以及相关实践引用的情况。其中问卷中的科学数据指的是各类科技活动产生的原始性、基础性数据及其分析研究信息,如地质、气象、海洋、水文、环境、基因、地震调查和观测等。同时,为提高问卷回复率、有效率和回答质量,本次问卷设计遵循了客观性、必要性和自愿性3项原则,同时设置了答题抽奖,采用网络在线填写的方式,接收各调研对象的答题。

问卷标题是“科学数据引用实践及认知现状调研(科研人员版)”,总共包含了17个问题,涵盖了3个背景性问题。主要涉及被调研人员的基本情况,如所属学科专业、从事科研工作的时限以及联系方式;3个客观性问题,主要了解被调研人员所熟悉的国内外科学数据库的情况;11个主观性问题,主要用于调研期刊编辑部对科学数据引用的态度、意愿等主观思想状况,其中包括2个单选题问题,7个多选题问题,2个开放式问题。

为控制问卷的填写质量,有效收集用户反馈,问卷的设计经过了预填写和正式发布两个阶段。在预填写阶段,邀请6名同事对在线调研问卷进行预填写,共填写了6份问卷。通过与填写人员的沟通,对问卷进行了调整优化,包括对问题的多选题设置了最多填写项和可填内容的其他项,增加了填写帮助说明文字,调整了部分选题的排列顺序。此外,为了防止在线问卷的恶意填写,从问卷的后台设置了防重复填写的功能,对每个IP设置了最多只能填写2次的限制,最终的结果显示独立IP数为3886个,IP的重复率为10.4%。

本次问卷调研的时间跨度为2012年12月11日-2012年12月24日,为期14天,如图1所示。发放对象主要是各个学科的科研人员。问卷的发放方式主要是通过开展在线有奖答题的方式进行。问卷标签是通过数据堂这一第三方公益性的科学数据平台首页进行投放。问卷的在线访问地址是<http://www.sojump.com/jq/2078525.aspx>,各类科研人员可通过访问上述地址,直接填写,方便快捷,答题基本可以在150~300秒内完成。

本次调研问卷总共回收问卷4337份,通过问卷星调查网站^[2]提供的后台统计发现,问卷的填写

区域已经覆盖了全国所有省份,包括了香港、澳门和台湾地区,如图2所示。来自北京的调研样例有635个,占14.64%。其次分别是广东349个,占8.05%;江苏340个,占7.84%。此外,还有来自国外的95份问卷,占总数的2.19%。

3 调查对象的背景分析

(1) 学科类别

本次调研的学科类别是参照《学位授予和人才培养学科目录(2011年)》(学位[2011]11号)^[3],采用了包括哲学、经济学、法学、教育学、文学、历史学、理学、工学、农学、医学、军事学、管理学等12大学科门类。考虑到艺术学的特殊性,没有设置该选项,但增加了“综合性跨学科专业”,以方便一些新兴跨学科专业人员填写。在参与调研的对象中,位居前五位的学科分别是工学、理学、管理学、医学和经济学,其中工学2175人,占50.15%;

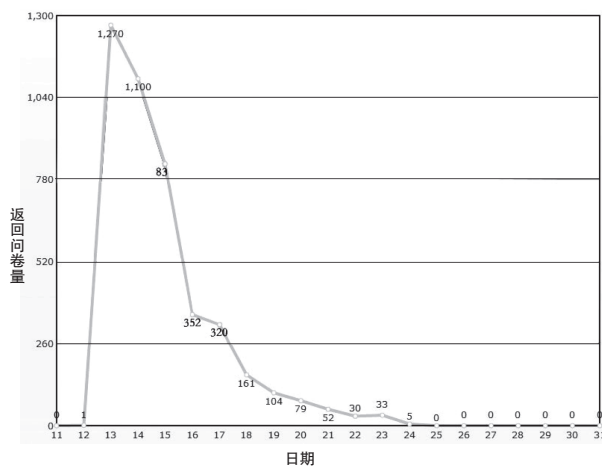


图1 调查问卷填写时序图



图2 调查问卷填写的地理位置分布图(以北京市为例)

理学767人，占17.69%；管理学260人，占5.99%；医学239人，占5.51%；经济学232人，占5.35%。综合性跨学科专业有204人，占4.7%。最少的是法学、历史学和军事学，小计70人，占1.61%。如图3所示。由此可见，本次调研的参与人员已经基本覆盖了所有学科类别。

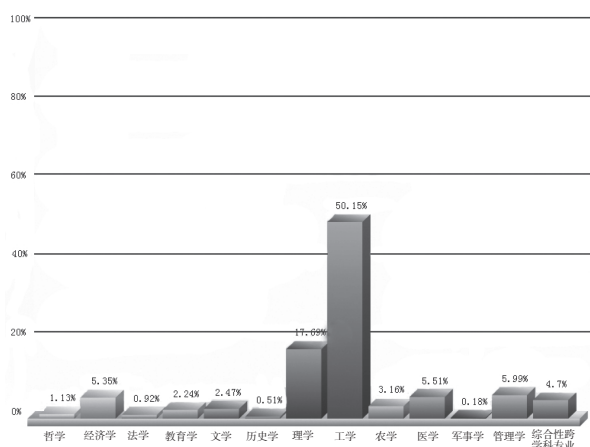


图3 调研对象学科类别分布图

(2) 科研工龄

在参与调研的科研人员中，2年以内的科研人员有1822人，占42.01%，2年以上的有2515人，占57.99%，其中10年以上的科研人员有554人，占12.77%。由此可见，本次调研的参加人员主要以中青年科研人员为主。

4 问卷的客观性

在4337名回答并返回问卷的人员中，针对科学数据引用情况的调查，有高达89.53%的人引用过科学数据，且有1856人“经常引用”，占42.79%。只有454人没有引用过，占10.47%。这也从另一个角度说明数据驱动型科研的现实存在^[4]。

当选择国内外熟知的科学数据库或平台时，国内的“中科院科学数据库”认知度最高，高达3240人，占74.71%。数据堂位列第二，有1608人，占37.08%。而在国家科技基础条件平台支持的科学数据平台中，认知程度依次是中国气象科学数据共享服务网（1087人，25.06%）、国家农业科学数据共享中心（841人，19.39%）、海洋科学数据共享中心（781人，18.01%）、国家地震科学数据共享中心（726人，16.74%）、地球系统科学数据共享网（631人，14.55%）、交通科学数据共享网（630人，14.53%）。如图4所示，而在国外所列的科学数据

库或平台中，WikiSpecies作为一种以维基方式存在的生物品种目录^[5]，认知度最高，有920人（占21.21%），人们可以自由编辑更新各类动植物物种信息，目前拥有354746种物种信息。欧洲分子生物学实验室建立的DNA数据库（European Molecular Biology Laboratory-DNA，EMBL-DNA）（672人，15.49%）和美国的GenBank（603人，13.9%）都是DNA领域的知名科学数据库，在本次针对国内科研人员的调研中，分别位列第二和第三。如图5所示。

同时发现，在所列的国内外众多科学数据库中，分别有11.41%（495人）和33.32%（1445人），选择了“不了解，没听说过”，从参与对象的学科类别来看，主要分布在理学、管理学、教育学和经济学等学科，从事科研工作的时限主要都是少于2年的调研对象。

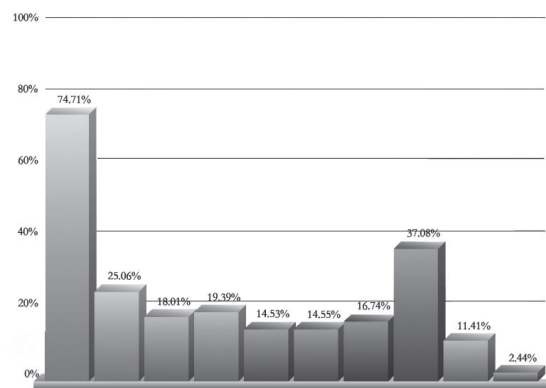


图4 国内科学数据库或平台的认知情况

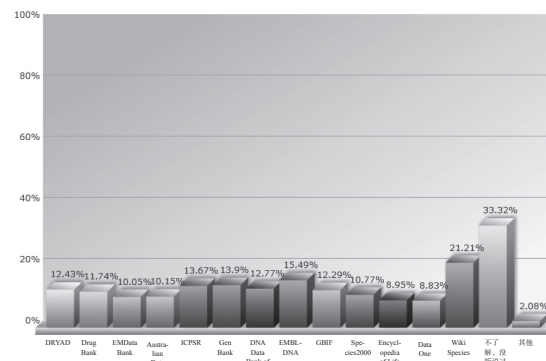


图5 国外科学数据库或平台的认知情况

5 问卷的主观性

(1) 引用科学数据的提交

在文章与数据结合的创新学术发表模式下^[6]，科学数据作为一种特有类别，需要发布在公共平台供其他读者引用或二次应用，那么科研人员更易接

受哪种性质的平台呢?经调研发现,科研人员更易接受专业领域内的科学数据平台,首选是“领域内的专业科学数据库中心”(3035人,69.98%),其次是期刊联盟(2490人,57.41%)或期刊数据库(1697人,39.13%),而由政府管理部门和商业公司建设的数据库都不在首选之列。具体如图6所示。

此外,在涉及科学数据提交的时间点时,参与调研的对象更倾向于在“论文公开发表后的阶段”提交(2118人,48.84%),“论文提交杂志评审的阶段”(758人,17.48%)、“课题研究过程关键节点”(754人,17.39%)和“课题验收后的阶段”(707人,16.3%)则基本持平。

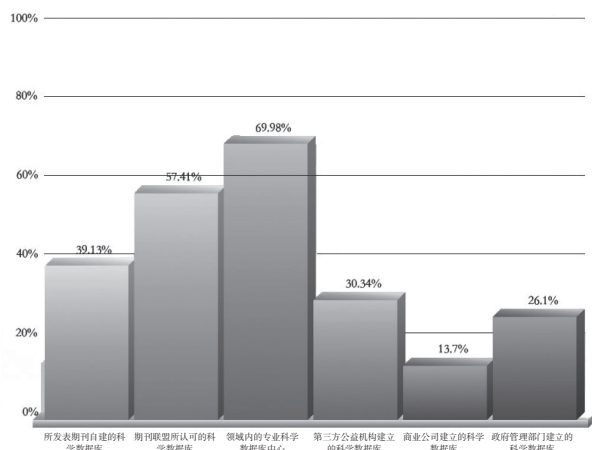


图6 作者提交论文相关科学数据集意向图

(2) 引用科学数据的描述与标注

在设计用于描述科学数据的10个元素中,数据名称(3089人,71.22%)和数据创建者(2826人,65.16%)这两个元素是调研人员选择的第一梯队,它们构成了科学数据描述的核心必备元素;数据来源网址、数据版本、科学数据类型、数据存储机构、数据创建日期、数据授权方式等8个元素则是非核心必备元素。数据获取号作为数据的唯一标识,如DOI可以方便的用于标识、解析和定位科学数据集^[7],在科学数据系统中具有非常重要的作用,而调研人员在选择时,并没有引起充分的重视,只有829人,占19.11%。这也从另外一个角度说明人们更注重表意的内容,而对于适用于系统的识读型编号并没有引起足够的重视。具体如图7所示。

而对于引用科学数据的标注位置,被调研人员认为,“数据说明部分”(2757人,63.57%)和“图表说明部分”(2191人,50.52%)都是比较合理的

标注位置,而作为文章的参考文献部分(1700人,39.2%)却并没有获得调研对象的认可,由此也说明现有的参考文献规范并不能完全适用于科学数据引用,需要根据科学数据引用情况进一步修订完善。

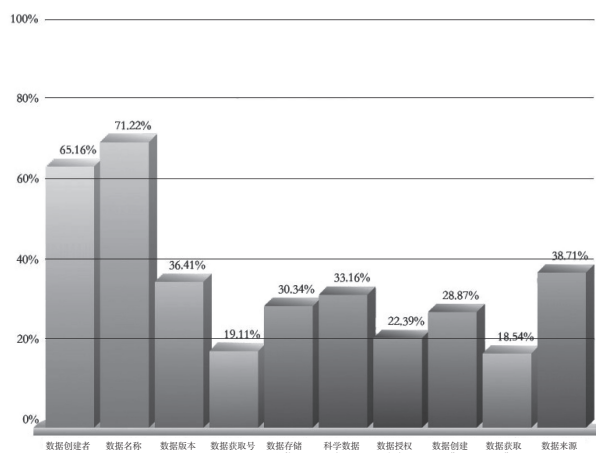


图7 科学数据引用的著录元素序列图

(3) 科学数据引用的动机

依据美国心理学家亚伯拉罕·马斯洛的需求层次理论^[8],科学数据引用行为的背后,隐含着科研人员怎样的引用动机,为此,根据5种需求层次,分别设计了6个选项,除社交需求设计了两个问题外,其他的都是一个选项。其中作为最基本的生理需求“有利于科研人员可以更好地去佐证自己的研究或他人的成果”获得了2818人认可,占64.98%。作为高级的成长需求“有利于发现大量当前还无法实现的新的研究机会”则获得了1193人认可,占27.51%。具体如表1所示。

此外,在论文中可能阻碍科学数据引用的三大因素分别是“没有开放可靠的数据存缴系统”(2785人,64.21%)、“没有提供规范的数据集引用格式”(2578人,59.44%)和“提供的数据提交工具不方便”(2205人,50.84%)。同时,可能促进科学数据引用的三大因素则分别是“方便的数据提交工具”(2298人,52.99%)、“通用规范的科学数据引用规则”(2290人,52.8%)和“投稿期刊提供清晰的论文关联科学数据提交流程”(2064人,47.59%)。

(4) 科学数据引用的功能

科学数据引用的功能包括了6个方面,其中科学数据的“引用使用”居第一位(2553人,58.87%),其次分别是“科学融合”(2235人,51.53%)与“明确出处”(2081人,47.98%)。如图8所示。

表1 科学数据引用需求层次表

选项	小计	比例	需求层次
有利于科研人员可以更好地去佐证自己的研究或他人的成果	2818	64.98%	生理的需求
有利于相关科学数据的检索和传播	2457	56.65%	社交的需求
有利于保证数据（或数据相关信息）链接的持久性和唯一性	2158	49.76%	安全的需求
有利于通过共享和被引用数据可获得更高的信誉和认可度	1359	31.34%	尊重的需求
有利于发现大量当前还无法实现的新的研究机会	1193	27.51%	成长的需求
有利于认识潜在的研究同行	937	21.6%	社交的需求

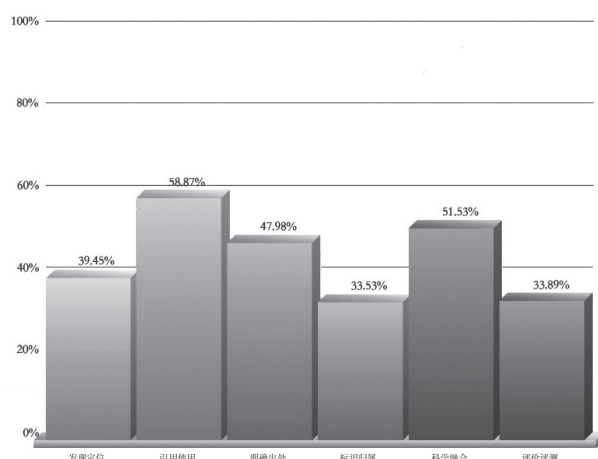


图8 科学数据引用功能排序图

（5）科学数据的商业化

在是否愿意授权对提交的科学数据进行商业化运作时，有4142人、占95.51%的调研对象愿意授权，其中有69.5%的人选择在不影响数据知识产权等的情况下，可“有条件授权”。选择“不同意”的只有195人，占4.5%。

6 答卷人的建议

科学数据作为科技论文的重要支撑与有机组成，二者的融合具有潜在的互动促进意义。一方面能够通过完整的内容表达，帮助读者更准确地获取论文传达的信息；另一方面科学数据可以作为相对独立的存在，为其他研究者提供一些重要参考。

在回答开放性问题“您对推动科技论文与科学数据融合引用有什么好提议呢？”时，总共有1535条记录，其中有333条为无效回答，如“无”“暂无”“还没想好”等。将剩余的1202条建议归纳整理成以下7个方面。

（1）科学数据的公开与共享

坚持“数据共享，科研至上”的基本理念，提倡数据公开，通过有计划、分阶段地执行科学数据开放获取策略，引导科研人员形成良好的数据共享

氛围，提供便携可靠的科学数据获取、共享、分析与利用服务。具体包括论文中须标明实验数据集的来源；将商业提供与论文作者自愿提交二者融合起来，形成互补；成立有利于数据共享的跨学科研究中心；在发布数据的同时，附带提供该数据的访问接口，而不用论文编写者自己开发接口来读取数据；组织期刊共同合作，并在高端论文上推广使用，形成标杆习惯；国家项目支持下的基础研究数据应强制向国家有关管理部门提交，否则不予验收等。

（2）科学数据引用的规范

规范数据格式和说明，制定统一通用的科学数据引用规则，在条件成熟后，可推出国家推荐标准，要求科技论文须有科学数据的支撑。共享数据格式，促进及时获取数据与数据内容的透明化。提供人性化的科学数据存储与引用服务，提高引用时效性和知名度。

（3）科学数据质量与知识产权

建立完善齐全的引用数据库，对科学数据库要求做到，数据完整真实，合理学科分类，数据标注详尽，数据审核严谨，规范可用有效。规范数据集，建立具体研究领域公认的benchmark数据集；数据版本多样化，能够被不同行业、不同学术水平的学者利用；推动数据库间的融合，便于科技论文引用时查找和使用；建立检索期刊专门的文章所引用科学数据库，定期更新数据。

数据提供机构需要对数据真实性进行认证，提高数据发布单位的可信度，强化数据库的广泛性，权威性，知名度；作为数据提供者与数据使用者的桥梁，在提供者与使用者之间进行匹配；建立良性循环的数据授权和使用方式；促进科研数据流通，采用激励方式促进数据的共享。

科学数据注明出处，重视科学数据知识产权保护，提出完善的知识产权保护措施。

（4）软件与应用平台

建立一个可信、规范、权威的科研数据共享平

台;提供各种专业数据的一站式搜索,能根据检索意图推荐合适的数据集;完善数据提交工具,提供更人性化的提交界面;建立国内外科学数据库的索引,利用类似于PAGERANKE技术,对各类数据集进行详细介绍、评价(或评级),减少科研人员查找数据时的困难;进一步规范数据格式及数据应用格式,提供多样化的存储格式和格式转换工具;增强与国际科学数据库的合作与链接;提供专业化的数据利用工具,提供更方便的权限;提供统一标准的引用接口和使用工具。

(5) 培训与宣传

通过讲座、会议、电子专题期刊等方式宣传科学数据引用,对科技人员进行相关培训,培养科学数据引用的习惯;可对高年级或者研究生开设必修课,教授如何在科技论文使用和看懂文中的数据和图表,使得科学工作者从起步就懂得如何规范地融合同行的数据和科学结论。

(6) 组织管理机制

科技论文与科学数据的融合发展需要政府管理部门、科研机构、出版单位、商业机构和科研人员的共同合作,在机制建设方面需要建立并完善科学数据引用机制,引导科学论文以科学数据为研究基础;建立数据审核与甄别机制,防止科学数据滥用;建立数据共享激励机制,保障数据提供者的适当收益;建立数据库监管机制,保障数据引用与数据安全。通过上述机制的建设,提升引用科学数据的价值,保护科学数据知识产权。

(7) 期刊的作用

期刊采取相应激励措施鼓励使用科学数据,同时作为投稿的一项要求,投稿同时提供原始的科学数据;加强对科技论文严谨性的审核,突出科学数据的作用,提高论文质量;提供完整、清晰的论文科学数据提交流程,在期刊编辑部投稿平台上提供数据提交平台,生成提交所需的规范化的数据;提高论文评审专家在评审的过程中对引用相关科学数据的重视程度;制定论文发表行为规范,为科学数据融合引用达成共识。

7 结论

通过本次针对科学数据引用的网络调研,基本了解了国内科研人员对开展并促进科学数据引用的态度和现状。

首先,科学数据的重要性已经获得科研人员的

普遍认同,很多研究人员由于科学数据共享程度不高,而无法进行跨区域比较、跨学科研究,更无法形成长时间周期的规律性认识。因此,积极呼吁通过多种方式宣传培训科学数据引用规范,大力倡导营造良好的科学数据引用文化氛围。

其次,在论文中,准确规范的标识科学数据,需要建立并推广统一的科学数据引用规范,建立共享的科学数据库和软件平台,提供人性化的数据提交工具,开发专业性的数据应用工具,从而丰富科学数据应用服务,最大程度的发挥各类科学数据的价值。

最后,期刊作为传统的出版媒体,在促进科技论文与科学数据引用中,发挥着关键的“平台”作用。既要通过制定激励措施,在论文投稿和编辑的过程中,要求作者提交相应的科学数据,同时,也发挥着宣传推广的功能,促使更多的科研人员用数据、引数据,形成良性互动行为,促进全社会科学素养的提升。

参考文献

- [1] 汤森路透发布 Data Citation Index 促进对全球研究数据的发现 [EB/OL]. [2013-01-10]. <http://ip-science.thomsonreuters.com.cn/press/press20121102/>.
- [2] 问卷星 [EB/OL]. [2012-12-11]. <http://www.sojump.com>.
- [3] 关于印发《学位授予和人才培养学科目录(2011年)》的通知 [EB/OL]. [2013-01-10]. http://www.moe.edu.cn/publicfiles/business/htmlfiles/moe/moe_834/201104/116439.html.
- [4] 微软《第四范式:数据密集型科学发现》发布 [EB/OL]. [2013-01-10]. <http://it.chinabyte.com/252/1246-3252.shtml>.
- [5] WikiSpecies [EB/OL]. [2013-01-10]. http://species.wikimedia.org/wiki/Main_Page.
- [6] 大数据杂志《GigaScience》正式发刊,实现文章与数据结合的创新学术发表模式 [EB/OL]. [2013-01-10]. <http://www.ebiotrade.com/newsf/2012-7/201271291628859.htm>.
- [7] 吴立宗,涂勇,王亮绪,等.浅谈科学数据出版中的数字对象唯一标识符[J].中国科技资源导刊,2010,42(5):22-29.DOI:10.3772/j.issn.1674-1544.2010.05.005.
- [8] 马斯洛需求层次理论 [EB/OL]. [2013-01-10]. <http://baike.baidu.com/view/690053.htm>.