

基于第三方引用标注的科学数据库影响力分析

朱艳华 胡良霖

(中国科学院计算机网络信息中心, 北京 100190)

摘要: 通过独立的第三方机构“同方知网”统计分析中国科学院科学数据库2006至2011年被学术论文全文引用标注的情况。结果显示, 科学数据库被引用标注1082次, 已经在科学研究、人才培养、政府决策和经济建设等诸多领域赢得了大量的用户, 产生深远的影响力。

关键词: 科学数据库; 影响力分析; 应用研究; 引用标注

中图分类号: N945

文献标识码: A

DOI: 10.3772/j.issn.1674-1544.2012.06.004

Analyzing the Influence of Scientific Database Based on the Third-party Cited Marks

Zhu Yanhua, Hu Lianglin

(Computer Network Information Center of CAS, Beijing 100190)

Abstract: Based on research results from National Knowledge Infrastructure (CNKI), we analyze cited frequency of scientific database from 2006 to 2011. The results show that scientific database of CAS was cited 1082 times by various papers in the past six years, which has won a large number of users and is playing an important influence in many areas, such as scientific research, personnel training, government decision-making and economic construction.

Keywords: scientific database, influence analysis, applied research, cited marks

1 引言

中国科学院从1982年就开展了科学数据库及其应用系统建设项目。经过30年的持续发展, 参与数据库建设的单位从最初的几家研究所扩展到院内62家研究所, 几乎覆盖了中国科学院所有的研究领域; 科学数据库工程已经建成为一个庞大的、资源类型多样的科学数据库群。科学数据库在国家经济建设、国防建设、规划决策、科学研究、学科发展、国家合作等诸多方面得到广泛应用, 已经初步取得显著的社会效益和一定的经济效益, 并在国内外产生了积极的影响^[1]。

科学数据库的数据资源量、访问量、下载量和服务科研项目数从不同方面反映了科学数据资源建

设和应用服务的情况, 已经在国内外产生了广泛的影响力。科技文献是科技论文的重要组成部分, 它表明科研工作的继承性和相关性, 是科技期刊和论文评价的最重要依据^[2]。科技文献引用指数是评估学术论文价值的最重要依据。同样数据库在学术论文中被引用和标注情况也可以作为其影响力分析的一个重要参考指标^[3]。

本文通过独立第三方机构“同方知网”, 调研科学数据库被学术论文全文引用和标注的情况, 期待从另外一个角度客观分析和评价科学数据库的综合效益和影响力。

2 科学数据资源体系和应用服务

在“十五”的基础上, 中国科学院系统规划了

第一作者简介: 朱艳华 (1982-), 女, 中国科学院计算机网络信息中心工程师, 硕士, 主要研究方向: 数据库技术与标准规范、数据应用服务。

基金项目: 中国科学院信息化专项项目“科技数据共享应用机制与效益评估研究”(XXH12510-02)。

收稿日期: 2012年8月17日。

科学数据资源体系,并于“十一五”期间将科学数据基础设施列入中国科学院信息化基本环境,进行重点建设^[4]。基于院内有特色和长期积累的数据资源,通过严格质量控制与管理建成了具有完整性和权威性的化学参考型数据库和中国植物物种信息数据库等2个参考型数据库。根据国家和中国科学院部署的重大研究计划或项目,建设了冰雪冻土环境数据库、聚变数据库、青海湖联合科研基地数据库和生态系统与功能区划数据库等4个专题数据库。面向特定的学科和应用领域,整合若干逻辑相关的数据库,建设了人地系统数据库、资源环境遥感数据库、空间科学数据库、天文科学数据库、微生物与病毒数据库、动物数据库、材料科学数据库、化学数据库等8个主题数据库,并从“十五”期间已支持且服务比较好、使用比较广泛的数据库中择优确定了37个重点专业数据库。此外,还积极引进/镜像常用的国际优质数据资源,努力建成国际重要科学数据资源集结点。

截至2011年,科学数据库累计建设各类资源量超过150TB,经由资源整合形成528个数据库^[5]。其数据资源领域分布情况如图1所示^[3]。

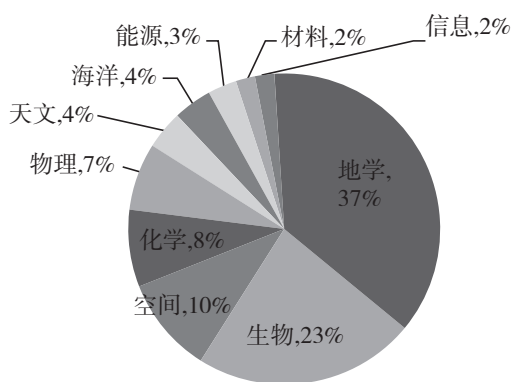


图1 中科院科学数据库数据资源领域分布状况

中国科学院科学数据服务已逐渐建立起一套完整的数据服务体系,即:一般数据在线服务和专业服务团队支持的深层次数据服务两个层次。其中,一般数据在线服务包括满足用户基本需求的查找数据库、数据检索、数据在线分发和信息服务等形式;深层次数据服务包括为满足用户咨询要求而开展的参考咨询服务,为用户科学数据代检代查要求而开展的委托查询服务,针对用户无法在线获取数据而开通的数据订阅和数据传递服务以及面向用户重大应用需求而提供的工具环境或数据服务空间的特定应用服务等。

截至2011年,“十一五”期间科学数据库访问量累计达3931万人次,数据下载量累计达353TB,直接服务重大科研项目210余项,取得了丰硕的科技成果和良好的社会效益。科学数据库网站年度访问人次持续增长(图2)。

3 第三方的引用指标来源

笔者在同方知网学术文献总库中选择了期刊、博士学位论文、硕士学位论文、会议论文和报纸五类文献资源,以参与“十一五”数据库项目的51个数据库名称为检索词,在全文字段中进行检索,检索时间限定为2006年1月1日至2011年12月31日。

在检索过程中,针对一些特殊情况,笔者进行了相应的处理。如有些数据库是“十五”建设的延续,在全文检索中就包括了该数据库“十一五”和“十五”期间的数据库名称,如“人地系统主题库”是在原“中国自然资源书库”基础上形成的,因此在检索中输入这两个库名,并选择“或包含”关系。有些数据库名称过于宽泛,虽然检索得到的文献全文包含了该数据库名称,但在对全文进行审核的过程中,发现文中所引用的数据库并不是该数

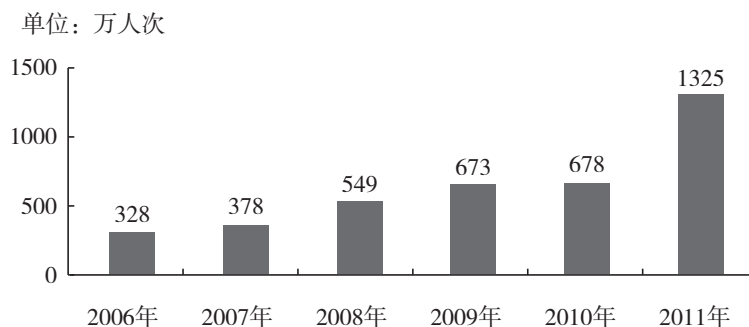


图2 中国科学院科学数据库网站年度访问人次增长趋势图

数据库，在这种情况下就加上建库单位的限定。如“材料数据库”检索结果中有些文献全文虽然包含了“材料数据库”，但所指的是英国或美国的“材料数据库”，因此在检索时，加上了建库单位“中国科学院金属研究所”的限制。

经过一系列的分析和筛选，笔者一共获得引用数据库的全文文献记录1082条，每条记录包括题名、作者、作者单位、文献来源、发表时间、被引用频次和下载频次等基本信息。同时，同方知网还对检索结果设计了9种分组分析方法。本文选用了其中的5组指标，即发表年度、来源数据库、作者单位、研究获得资助和研究层次，以综合考察科学数据库的影响力。

4 论文引用整体情况分析

4.1 论文年度分布

2006年至2011年全文中引用科学数据库的论文共1082篇，其中2006年119篇，占论文总量的11.0%；2007年180篇，占论文总量的16.6%；2008年203篇，占论文总量的18.8%；2009年190篇，占论文总量的17.6%；2010年180篇，占论文总量的16.6%；2011年210篇，占论文总量的19.4%。从论

文的年度分布来看，科学数据库平均每年支持发表论文180篇，总体呈现稳定且略有上升的增长趋势（图3）。

4.2 论文类型分布

在科学数据库被引的1082篇论文中，期刊论文582篇，占论文总量的53.8%；硕士学位论文334篇，占论文总量的30.8%；博士学位论文111篇，占论文总量的10.3%；会议论文41篇，占论文总量的3.8%；报纸14篇，占论文总量的1.3%（图4）。

从论文的类型分布看，科学数据库主要用于支持科学研究，引用数据库的论文中期刊论文占了一半以上；科学数据库在人才培养方面也发挥了积极作用，支持的硕士和博士学位论文占了论文总量的41.1%。

4.3 论文作者单位分布

根据论文第一作者所在单位的行业性质划分，在1082篇论文中，有782篇来自高等学校，占论文总量的72.3%；有216篇来自科研机构，占论文总量的20.0%；有43篇来自政府管理部门，占论文总量的4.0%；有12篇来自公司企业，占论文总量的1.1%；未标明单位的有29篇，占论文总量的2.6%（图5）。

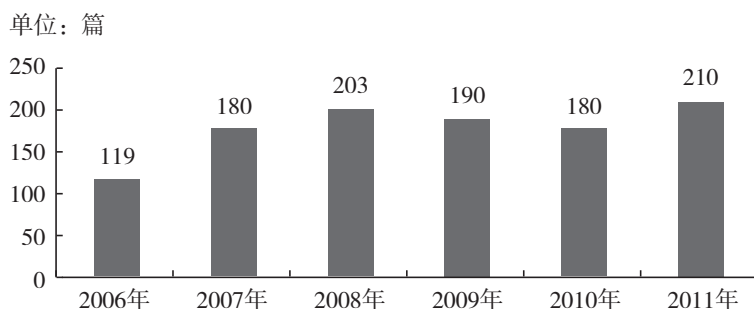


图3 科学数据库支持发表的论文年度分布

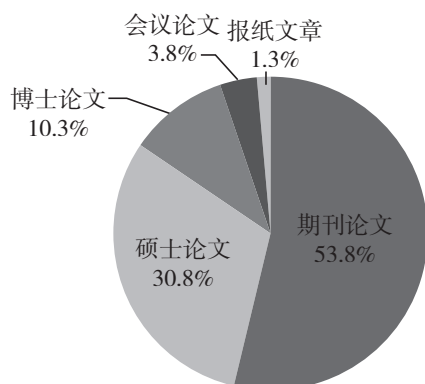


图4 科学数据库支持的论文类型分布

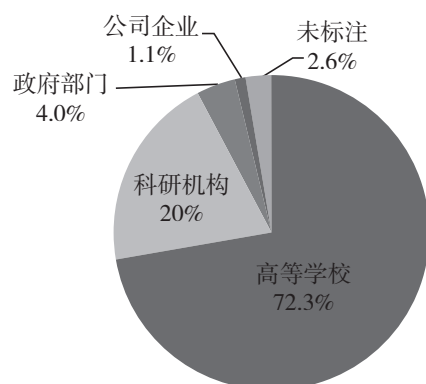


图5 科学数据库支持的论文作者单位分布

引用论文的行业分布与科学数据库服务于科学研究和培养人才的宗旨相符。虽然科学数据库主要是中国科学院研究所联合建设,但是其服务范围远远超出中国科学院内的研究人员,更多地为高校广大师生提供科研支持,其中来自高校的论文引用占了70%以上。此外,科学数据库还服务于政府管理部门,为部门决策提供一定程度的参考支持。

4.4 论文研究获得资助分布

根据同方知网对研究资助的划分,引用科学数据库1082篇论文共支持基金550项,平均每两篇就获得一项基金项目。其中,国际基金9项,占基金总量的1.6%;国家部委基金407项,占基金总量的74.0%;科研院所基金58项,占基金总量的10.5%;高校基金13项,占基金总量的2.4%;各地方政府基金63项,占基金总量的11.5%(图6)。

引用科学数据库的论文获得基金资助,在一定程度上也可以理解为科学数据库支持了这些基金项目研究,统计结果彰显了科学数据库的服务成效。

4.5 论文研究层次分布

同方知网将论文所属的研究层次划分为自然科学层次、社会科学层次和其他三大类。其中,自然科学研究层次包括基础与应用基础研究、工程技术、行业技术指导、政策研究、专业实用技术等;社会科学研究层次包括基础研究、行业指导、政策研究、职业指导等;其他层次包括基础教育与中等职业教育、高等教育和大众科普等。

在1082篇论文中,有4篇没有标引研究层次。在标引研究层次的1078篇论文中,有861篇属于自然科学研究层次(其中,基础与应用基础研究、工程技术、行业技术指导、政策研究、专业实用技术分别有510篇、321篇、21篇、5篇、4篇),占标

引总量的79.9%;203篇属于社会科学研究层次(其中,基础研究、行业指导、政策研究、职业指导分别有139篇、42篇、19篇、3篇),占标引总量的18.8%;14篇属于其他研究层次(其中,基础教育与中等职业教育、高等教育和大众科普分别有1篇、9篇、4篇),占标引总量的1.3%(图7)。

以上统计数据显示,科学数据库主要支持自然科学研究,占论文引用总量近80%;社会科学研究也占了一定比例,科学数据库也同样支持相关社会科学研究。从论文内容来看,科学数据库不仅为研究者们提供了基本的数据支持,还利用数据资源引导他们探索新发现的领域和科学问题,提供多种分析视角。从论文支持力度来看,自然科学研究层次中基础研究有510篇,占该研究层次的59.2%;社会科学研究层次中基础研究有139篇,占该研究层次的68.5%。这共同说明了科学数据的基础支撑作用,在科研活动中具有基础性和战略性地位,业已参与国家基础科学数据共享网建设,成为国家科学数据共享平台的重要组成部分。

5 被引数据库的学科分析

5.1 被引数据库学科分布

根据科学数据库学科领域划分,在1082篇论文中,659篇引用地球科学领域的数据库,占论文总量的60.9%;159篇引用生命科学领域的数据库,占论文总量的14.7%;116篇引用物理化学领域的数据库,占论文总量的10.7%;60篇引用材料科学领域的数据库,占论文总量的5.5%;56篇引用天文与空间学科领域的数据库,占论文总量的5.2%;32篇引用其他领域的数据库,占论文总量的3.0%(图8)。

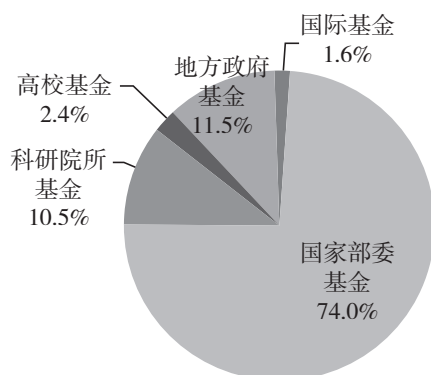


图6 科学数据库支持的基金类型分布

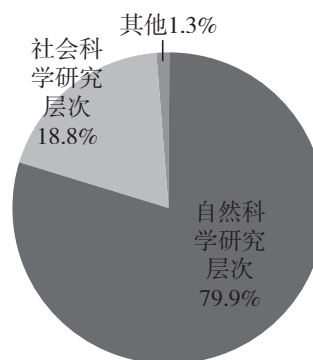


图7 科学数据库支持的论文研究层次分布

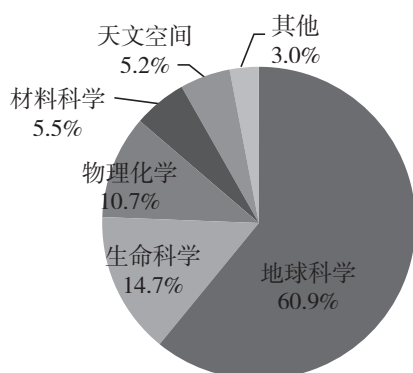


图8 论文引用数据库的学科分布

如图1所示科学数据资源体系中地球科学领域数量最多，占全部资源的37%，而其论文引用量则占了总量的60%以上。事实上，地球科学领域的数据库不仅支持自然科学研究，而且支持社会科学研究。论文所属的研究层次统计显示，在203篇社会科学研究层次的论文中引用地球科学数据库的论文162篇，占该研究层次的79.8%。而天文空间等学科领域虽然用途重要，但由于专业性强、关注度相对较低等原因致使论文引用量较少、引用率较低。

5.2 论文引用最多的10个数据库

按照科学数据库被引用论文数量排序，得到被引用最多的前10个数据库（图9）。从图9中可以看出，论文引用量最多的10个数据库分布在各个学科领域，其中以地球科学领域最多，有5个数据库。论文引用多的数据库一般具有以下特点：建库时间长，资源丰富，很大程度上满足了用户的需求，逐步形成了稳定的用户群；所服务的用户不仅分布

在自然科学领域，还分布在社会科学等其他领域。相反，那些论文引用数量较少的数据库建库时间相对较短，可能有些还停留在自建自用阶段，能提供的资源有限，服务形式单一；也可能由于那些数据库专业性强，从事相关研究的人较少，导致关注度低。

6 结语

本文基于独立的第三方机构“同方知网”全文检索数据，从论文年度分布、类型分布、作者单位行业分布、论文获得资助分布和论文研究层次角度客观分析了中国科学院科学数据库在“十一五”期间被学术论文引用标注的情况，并对被引数据库的学科领域进行了比较分析。结果显示，科学数据库在学术论文中具有较高的被引频次，已经服务于科学研究、人才培养、政府决策和经济建设等诸多方面，产生了广泛的影响力和综合效益。

本文分析所用的数据仅来源于同方知网，并没有从专家和用户定性评价角度考察科学数据库的影响。尽管如此，作为世界上全文信息量规模最大的网络出版平台，同方知网发布的数据具有较强的代表性和客观性。

本文对中国科学院科学数据库的综合分析表明，引用较多的数据库都具有建库时间长，资源丰富，更新及时；服务形式多样，满足特定用户需求；领域专家和信息技术人员形成稳定的团队，提供联合服务等特点。从中我们得出如下几点启示。

一是科学数据建设是一项需要长期、持续、

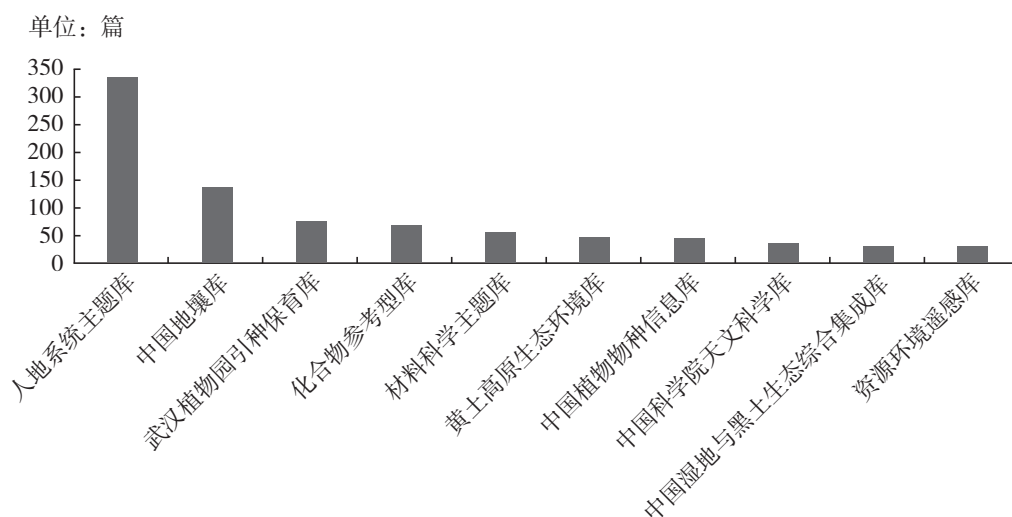


图9 论文引用最多的10个数据库

足量投入的系统工程。尤其对于那些反映客观世界发展规律的数据,长期收集整理具有更大的分析、研究和利用价值。二是随着科学数据的快速积累和信息服务技术的不断成熟,用户对数据服务的要求不断提高,数据浏览、查询和下载等常见功能模式很难满足其特定需求,数据服务向多元化、集成化和主动式的专业服务方向发展。顺应这种趋势的数据库开发了很多针对用户需求的应用软件工具,例如,人地系统主题库形成了一套数据在线可视化与分析工具集、化学数据库开发了特殊化学数据(结构和谱图)的可视化展示方法、武汉植物园数据库集成了地图搜索等功能,这些服务模式让用户的体验更方便、更快速,继而提高了这些数据库的利用率和被引率。三是这些被引用较多的数据库都有领域专家和信息技术人员联合服务的良好基础和传统,他们都有固定的学科领域专家和技术人员组成的稳定服务团队,利用多种方式采取在线、离线相结合的手段,开展面向特定领域、特定项目的专家咨询以及面向一般用户的技术服务咨询。

论文引用情况表明,经过30年的建设,中国

科学院科学数据库不仅为科学研究提供基础数据支持,还引导研究者们探索新发展的领域和科学问题;不仅为国家高层次人才培养提供支持,还为国家重大战略决策提供重要参考。

鉴于科技数据的基础性和战略性地位,不断加强中国科学院科学数据的整合与共享是今后的重要任务。

参考文献

- [1] 孙九林.科学数据库发展战略思考[C]//科学数据库与信息技术论文集.北京:兵器工业出版社,2010:3-7.
 - [2] 董建军.参考文献引用分类标注与科技期刊和论文的评价[J].编辑学报,2006,18(6):406-409.
 - [3] 廖顺宝.从数据库的引用标注看科技资源的应用服务成效[C]//科学数据库与信息技术论文集.北京:科学出版社,2012:391-196.
 - [4] 中国科学院信息办.2010年中国科学院信息化资源报告[R].北京,2010.
 - [5] 陈明奇.中国科学院科技数据资源现状及其发展思考[C]//科学数据库与信息技术论文集.北京:科学出版社,2012:3-9.
-
- (上接第16页)
- Sharing and Retrieval Using OAI-PMH[J]. Earth Sci Inform, 2011(4):1-5.
 - [8] Stanislav Grill, Michal Schneider. Geonetwork Open-source as an Application for SDI and Education[EB/OL]. [2012-12-09]. http://gis.vsb.cz/GIS_Ostrava/GIS_Ova_2009/Sbornik/Lists/Papers/039.pdf.
 - [9] Kristin Stock, Gobe Hobona, Carlos Granell, et al. Ontology-Based Geospatial Approaches for Semantic Awareness in Earth Observation Systems. Geospatial Semantics and the Semantic Web[J]. Semantic Web and Beyond, 2011(12):97-118. DOI:10.1007/978-1-4419-9446-2_5.
 - [10] Chaitan Baru, Sandeep Chandra, Kai Lin, et al. The GEON Service-oriented Architecture for Earth Science Applications[J]. International Journal of Digital Earth, 2009, 2(21):62-78.
 - [11] Xin Li, Zhuotong Nan, Guodong Cheng, et al. Toward an Improved Data Stewardship and Service for Environmental and Ecological Science Data in West China[J]. International Journal of Digital Earth, 2011,4(4):347-359.
 - [12] 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.
 - [13] 涂勇,彭洁.基于DOI技术的科学数据与科技文献融合的研究[J].数字图书馆论坛,2007(10):28-31.
 - [14] Robert G Raskin, Michael J Pan. Knowledge Representation in the Semantic Web for Earth and Environmental Terminology (SWEET)[J].Computers & Geosciences, 2005(31):1119-1125.
 - [15] 贺德方,张旭.服务于科技信息资源共享的数字对象唯一标识应用研究[J].现代图书情报技术,2007(8):26-29.
 - [16] 诸云强,孙九林,宋佳,等.地学e-Science研究与实践——以东北亚联合科学考察与合作研究平台构建为例[J].地球科学进展,2011,26(1):66-74.