我国科学数据共享研究的文献计量分析 *

Bibliometric Analysis on Scientific Data Sharing in China

王巧玲[1,2]　钟永恒[2]　江　洪[2]

(1. 中国科学院研究生院 国家科学图书馆　北京　100190;
2. 中国科学院国家科学图书馆武汉分馆　武汉　430071)

摘　要　以《中国期刊全文数据库》和维普《中文科技期刊数据库》为信息源,对我国科学数据共享研究的论文进行检索,经去重、去杂处理后,进行时间、主题、作者和论文刊载期刊分布的计量分析。总结我国科学数据共享研究的进展和不足,如我国科学数据共享研究的主题分布不平衡、研究的机构分布不平衡等,并提出相应的建议。

关键词　科学数据　科技数据　数据共享　文献计量

中图分类号　G350

美国自 20 世纪 90 年代以来,实行"国有科学数据完全与开放共享国策",极大地推动了美国经济的增长。我国于 2002 年启动"科学数据共享工程",力求构建面向全社会的共享服务体系,实现科学数据的规范化管理和高效利用。科学数据的共享研究得到了普遍重视,本文拟对科学数据共享研究的情况进行文献计量分析,以便为国家科学数据共享工程和相关研究提供参考。

1　统计来源和方法

1.1　**科学数据定义**　根据《OECD 关于公共资助科学数据获取的原则和方针》[1]中科学数据(Research data)的定义:作为科学研究基本来源的事实记录(数值,文本记录,图像和声音),被科学团体所共同接受的对研究结果有用的数据。但不包括这些内容:实验室笔记、初步分析、科学论文的草稿、未来的研究计划、同行评论以及个人和同行的交流,以及实物(如实验样本、细菌和测试的动物)等。另外还强调数据为数字化的计算机可读的科学数据。

公共资助的科学数据(Research data from public funding)的定义:政府部门或机构指导的研究或者利用公共资助资金进行研究而获得的科学数据。

我国《科学数据共享工程》提到的科学数据的定义:科学数据是人类社会科技活动所产生的基本数据、资料,以及按照不同需求而系统加工的数据产品和相关信息,具有明显的潜在价值和可开发价值,并在应用过程中得以增值,是信息时代最基本、最活跃、影响面最宽的科技资源[2]。

鉴于我国《我国科学数据共享工程》中提到的科学数据的定义比较宽泛,笔者采用《OECD 关于公共资助科学数据获取的原则和方针》报告中明确提出的关于公共资助的科学数据的定义,它也符合我国科学数据共享工程所研究的内容。

1.2　**检索策略及结果**　笔者以中国期刊全文数据库和维普中文科技期刊数据库为数据源进行分析,两个数据库的检索策略和检索结果如下:从中国期刊全文数据库,以关键词:科学数据或者科技数据并且关键词:共享的高级检索模式进行精确检索,时间范围 1994～2007 年,共搜到 191 篇论文,检索时间:10 月 25 日。

从维普中文科技期刊数据库,以题名或关键词:科学数据或者题名或关键词:科技数据并且题名或者关键词:共享的高级检索模式进行精确检索,共搜到 180 篇论文,时间范围 1989 - 2007 年,检索时间 10 月 25 日,将两者的结果综合起来进行去重去杂处理后,得到论文 248 篇,用 Excel 进行数据统计,以文献计量学方法对所得到的论文进行时间分布、主题分布分析(作者分析和期刊分析基于中国期刊全文数据库,下文有说明),总结我国科学数据共享研究工作的进展和不足,为我国进一步的科学数据共享研究工作提供参考。

2　统计结果与分析

2.1　**论文年代分布及其数据增长趋势**　经统计得出,我国科学数据共享研究论文时间分布如表 1。

由表 1 可以看出我国最早于 1996 年在地球科学领域出现了研究科学数据的论文,地球科学研究需要大量的科学数据作为基础,1957 成立的世界数据中心(WDC)亦是以地球科学、空间科学和天文学数据为重点,这与科学研究的需要是必

基金项目:湖北省科技厅"科学数据共享机制与制度的研究"项目成果之一。

作者简介:王巧玲,女,1983 年生,硕士研究生,研究方向为信息资源组织与建设研究;钟永恒,男,1965 年生,研究员,馆长,研究方向为信息资源组织与建设研究;江　洪,女,1968 年生,副研究员,业务处处长,研究方向为图书馆学研究。

然联系的,从而,可以在一定程度上说明我国的科学数据研究起步较晚。1996 年的研究工作处于萌芽阶段,有 3 篇论文,之后到 2002 年前几乎没有相关研究,直到 2002 年,即我国科学数据共享工程要实施的一年,科学数据共享研究论文出现突增,得到论文 14 篇,2003、2004 年论文持续增长,2006 年我国科学数据共享的研究论文出现高潮,这与第 20 届国际科技数据委员会会议在我国举行是分不开的,同时也是我国科学数据共享工程进行全面推进阶段[3]的开始。2007 年的研究论文和期刊的发文周期有一定的关系,截至检索时间 2007 年 10 月 25 日,仅得到论文 31 篇,不足以反映 2007 年论文增长情况。但笔者认为随着我国科学数据共享工程的开展,会引起越来越多的研究人员对科学数据共享工作的关注,因此,我国科学数据共享研究论文会出现持续增长的趋势。

表 1　1994～2007 年科学数据共享研究论文分布

年份	1994	1995	1996	1997	1998	1999	2000
论文数	0	0	3	1	0	1	0
年份	2001	2002	2003	2004	2005	2006	2007
论文数	0	14	42	45	35	76	31

2.2　**主题分布**　笔者通过对研究论文的文摘或者全文的浏览,把研究主题划分为理论研究、技术工作、政策法规、新闻消息、机制研究等五类,统计得出的主题分布图 1、表 2,如下所示:

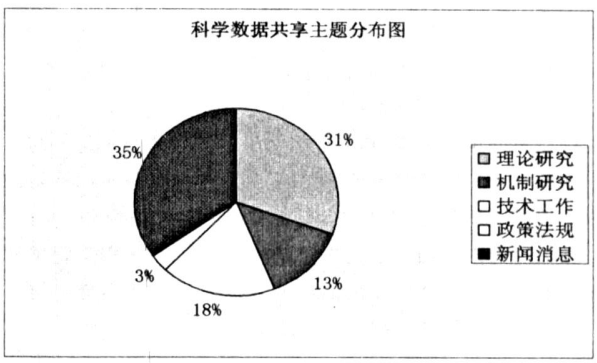

图 1　科学数据共享主题分布图

表 2　科学数据共享主题分布年份表

主题分布	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	论文数
理论研究	1	0	0	1	0	0	5	21	12	7	20	9	76
机制研究	0	0	0	0	0	0	0	6	9	4	10	4	33
技术工作	1	1	0	0	0	0	1	4	13	11	7	7	45
政策法规	0	0	0	0	0	0	0	1	2	0	4	1	8
新闻消息	1	0	0	0	0	0	8	10	9	13	35	10	86

理论研究包括科学数据共享的意义、作用、共享中会遇到的问题和挑战、国内外现状、综述以及科学数据共享和其他工作的关系等。2003 年和 2006 年论文比较多,这与我国科学数据共享工程的实施和国际科学技术数据委员会的会议是分不开的,二者促进了我国科学数据共享的研究工作。总的来说,我国在科学数据共享工作方面时间还不是很长,理论研究在所有研究论文中占到了 1/3 以上是情理之中,借鉴其他国家成功经验的理论研究,对我国科学数据共享工作的意义重大,随着我国科学数据共享工作的开展,在"数据科学成为一门独立的学科在科学界正式地被确立下来[4]"的背景下,我国科学数据共享的理论研究会有更大的发展。

机制研究包括科学数据共享的协调管理机制、资源整合平台、科学数据分级分类管理和发布策略、数据共享的标准体系研究等。刘闯指出"科学数据共享重在机制建设",机制建设是科学数据共享工作的保障。从论文数量上来说,机制研究在科学数据共享所有论文中占的比重还不是很大,仅占到 13%的份量,从内容上来看,机制研究集中在国外经验介绍和国内宏观方面的探讨,以及从宏观方面论述了我国科学数据共享的机制问题,形成了比较一致的看法,科学数据共享机制一般应包括:政策法规保障机制、资源整合和分级分类共享机制、技术保障机制、协调管理机制等,但就具体的科学数据整合、分级分类、发布、保存机制等科学数据管理流程中机制的研究还比较少。

技术工作包括元数据研究、Web Services 技术、信息系统、网格研究等。我国在科学数据共享技术方面已经形成了《元数据标准化基本原则和方法》、《元数据内容》、《元数据 XML/XSD 置标规则》、《元数据检索和提取协议》等关于元数据的技术标准[5]。在当前网络技术迅速发展的环境下,数据获取和共享最主要的障碍不再是技术问题,而是制度和管理、资金和预算、政策法规和文化行为的问题[6]。

政策法规研究包括国家政策方针、信息立法以及知识产权法研究等。从表 2 可以看出我国在这方面的研究论文仅有 8 篇,介绍了欧美科学数据共享的法制建设,提出了科学数据共享与知识产权冲突与协调的问题,并对构建我国科学数据共享的政策法规体系等,均提出了重要的问题,但就提出问题的解决和相关问题的研究还没有跟踪研究。我国科学数据共享管理和立法工作与国际先进国家相比,与日益高涨的共享呼声相比,还有相当的差距,必须加快科学数据共享政策和法规体系研究的脚步[7]。

在检出的所有论文中,新闻消息占 35%,这是因为科学数据数据共享工作在我国开展的时间比较短。广泛的科学数据共享宣传工作有利于调动起全社会对科学数据共享的关注,从而为全社会的科学数据共享打下坚实的"共识[8]"基础。

2.3　**作者分布**　鉴于统计的可行性,作者分布和机构分布的分析是以中国期刊全文数据库为信息源,对检索出的 191 条数据进行初步处理得到 147 条数据,即 147 篇论文,统计后得出:共有 221 名作者独立或者合作参与科学数据共享的研究工作,论文合著情况及占论文总数的比例如表 3。

表 3　论文合著情况

独著	两个作者	三个作者	四个作者及以上	篇均作者	
篇数(篇)	70	36	18	23	1.50
百分比(%)	47.62	24.49	12.24	15.65	

篇均作者数、合著率反映作者的合作程度,进而可以反映本学科与其它学科的交叉情况以及本学科领域内研究的深入情况。篇均作者数是指在一定时期内,某期刊、某学科的每篇

论文的平均作者数;合著率是指在一定时期内,某期刊、某学科多著者论文数与总论文数之比。篇均作者数、合著率与学科的综合性、研究的难易程度有关,虽然社会科学文献的合著程度小于自然科学,但其合作率也正在逐步提高。娄策群在《社会科学评价的文献计量理论与方法》一书中写到:1987～1996 年国外社会科学各学科著作的平均合著率为 25.26 %,《中国社会科学》所载论文的合著率为 20.73 %[9]。

从表 2 可以看出我国在科学数据共享研究方面合著率很高,占到一半以上,篇均作者为 1.5 人,说明科学数据共享工作涉及的专业领域比较广,计算机、图书情报、知识产权等;牵涉的部门比较多,如国家政府部门、科研单位、信息管理部门以及科学家个人等,工作的难度比较大。

作者发表文章数越多,在一定程度上可以说明该作者对这一领域的研究进行得越深入,有跟踪研究,对这一领域的贡献越大。从表 4 可以看出,发表四篇及四篇以上论文的作者人数有 8 个人,发表一篇论文的作者数占总人数的 77.83 %,说明我国在科学数据共享研究方面还比较薄弱,高产出作者还不多。相对高产出作者的信息,如表 5。

表 4　作者发表文章数量的分布

论文数	四篇及以上	发表三篇	发表两篇	发表一篇
作者人数	8	11	30	172

表 5　发表论文在四篇及以上的作者信息

作者	作者单位	论文数
刘闯	中国科学院地理科学与资源研究所全球变化信息研究中心	7
尹岭	解放军总医院神经信息中心	6
黄鼎成	中国科学院地质与地球物理研究所	5
孙九林	中国科学院地理科学与资源研究所	5
李集明	中国气象局气象信息中心	5
朱星明	中国水利水电科学研究院	4
李晓波	国土资源部信息中心	4
魏淑艳	东北大学文法学院	4

2.4　**论文所在期刊分布**　147 篇论文发表在 79 种期刊上,其中有 5 种期刊发表文章在 4 篇以上,共发表文章 49 篇,占全部论文数 1/3;发表文章最多的期刊是《中国基础科学》,共发表文章 30 篇,如表 6;发表一篇论文的期刊 58 种。从表 7 可以看出,我国科学数据共享论文一半以上发表在自然科学综合类期刊上,图书情报类期刊发表论文相对也比较集中,9 种期刊共发表论文 17 种,如表 8。表 8 对图书情报同行投稿和参考已有研究有所帮助。

表 6　刊载论文 4 篇以上的期刊

期刊名	论文数
中国基础科学	30
中国科技论坛	5
太原科技	6
地球科学进展	4
科学中国人	4

表 7　147 篇论文在不同期刊上的分布

期刊类型	期刊种数	刊载论文数	占总论文数比例(%)
自然科学综合类	28	76	51.70
专业学术期刊	14	17	11.56
图书情报类	9	17	11.56
学报	11	16	10.88
其他	17	21	14.29

表 8　图书情报类期刊上的论文分布

期刊名	论文数
现代情报	3
图书情报工作	3
农业图书情报学刊	3
图书馆论坛	2
情报学报	2
中华医学图书情报杂志	1
医学情报工作	1
林业科技情报	1
中国档案	1

3　存在的主要问题与建议

3.1　**科学数据共享研究的主题分布不平衡**　从以上的主题分布分析可以看出,我国在科学数据共享方面还处于初步阶段,研究论文一方面是宏观的,谈共享意义多,另外一方面是谈建设的技术多,而科学数据共享的关键在机制建设,保证科学数据共享顺利进行的法宝是政策法规建设,而我国在这些方面做得还是不够的。

对不同学科来说,科学数据对科学研究的作用是不同的,科学数据管理的策略也是不同的,如科学数据收集的策略,是收集数据流还是数据事件[10],在 HIV/AIDS 研究中,要收集一长段时间内的相对不变数据流中的数据;在地震工程领域,倾向于收集不连续的离散的事实数据而不是数据流。因此,我国应加强对科学数据共享管理流程中各环节(数据产生、收集、汇交、整理、保存以及使用等)、各要素(机构、组织、人员、资金等)以及产生的具体问题的研究,从微观方面入手对科学数据共享机制进行深入研究。

王正兴、刘闯等在《科学数据可持续共享:关键是利益的均衡》[11]一文中指出,限制科学数据共享的根本原因是与科学数据生产、流通和利用的各方的利益失去平衡,深层次的原因是没有建立相关的法律和政策,而从上文的统计结果可以看出,我国在科学数据共享政策法规方面的研究不多,为保障我国科学数据共享工程的顺利进行,我国要加强科学数据共享的政策法规研究,完善科学数据共享的政策法规体系,明确科学数据的知识产权保护问题等。

3.2　**科学数据共享研究的机构分布不平衡**　纵观上文的统计数据发现,科学数据共享的研究主要集中在中科院各研究所,高等院校单位参与的研究不是很多,这不利于我国科学数据共享工程的开展,我国高等院校,特别是进入国家"211 工程"的若干研究型大学的科学家承担了国家　(下转第 134 页)

3　讨论和建议

联合国等权威机构的研究表明,电子政务建设有很高风险。建立完整和体系化的电子政务绩效评估模型对正确引导电子政务建设和保证投资效益有重要价值。研究首先对官方机构、大学和商业咨询公司三类六家机构建构的电子政务评估模型比较分析,发现电子政务评估中以公民为中心和以结果为导向的基本原则,本研究引入比电子政务评估发展成熟得多的 Delone 和 Mclean 等人在企业信息系统成功模型以指导电子政务评估模型的建立,以使模型更加完整和体系化,更能针对中国的电子政务发展中的"重硬轻软"等不足起纠正作用。我们分析认为一个完整和体系化的电子政务系统绩效评估应该包括电子政务系统质量、信息和服务的质量、电子政务的应用基础环境、三方用户的感知有用性、用户在使用系统中提高满意度和电子政务的目标组成。参考国内外有关的权威文献和访谈 30 多位专家的基础上建构出本研究的电子政务绩效评估参考模型。另外,本研究的目的在于做出评估的参考框架,因此不进一步做出详细的指标体系,因为在实际运作中,指标体系和对评估权重的具体赋值,必须根据具体情况来决定。比如,国家一级、地市级以及县级的电子政务系统在安全性的要求上的不同的,涉外服务与非涉外服务的系统的要求也是不同的等等。在具体运用中可以参考本研究的模型进行细化和量化。

参 考 文 献

1　UN. E - Government at the crossroads[R]. www. un. org ,Aug. 2003

2　国务院信息化办公室[R]. 中国信息化发展报告,2006

3　Accenture ,e - government leadership: high performance and maximum value[R]. http :/ / www. accenture. com/

4　UN. UN Global E - government Survey 2003[R]. http :/ / www. unpan. org/

5　UN. Global E - Government Readiness Report 2004[R]. http :/ / www. unpan. org/

6　Brown University , Global E - Government ,2004[R]. http :/ / www. brown. edu/

7　Fortuneage ,中国电子政务研究报告[R]. http :/ / www. fortuneage. com/

8　张维迎,刘　鹤. 中国地级市电子政务研究报告[R]. 北京:经济科学出版社,2003

9　王长胜. 中国电子政务发展报告[M]. 北京:社会科学文献出版社,2003

10　Delone W H. Mclean E R. Information System Success: The Quest for the Dependent Variables[J]. Information Systems Research ,1992 (3) :60 - 95

11　Seddon P B. A Respecification and Extension of the Delone and Mclean Model of IS Success[J]. Information System Research ,1997 (3) :240 - 253

12　Seddon P B. Kiew M Y. A Partial Test and Development of Delone and Mclean 's Model of Is Success[J]. Australian Journal of Information System ,1996(1) :90 - 105

13　Shang ,S. Sedden , P. B. A comprehensive Framework for Classifying the Benefits of ERP System[C]. The Proceedings International Conference on Information System ,1998 :165 - 174

14　徐维祥, 张全寿. 从定性到定量信息系统项目评价方法研究[J]. 系统工程理论与实践,2001 ,21(3) :124 - 128

15　张玲玲,佟仁城. 企业信息系统项目综合评价指标体系探究[J]. 中国管理科学,2004 ,12(1) :95 - 101

16　OMB , E - Government Strategy [R]. http :/ / www. whitehouse. gov/ omb/ inforeg/ egovstrategy. pdf

17　Delone W H. Mclean E R. Information System Success: The Quest for the Dependent Variables[J]. Information Systems Research ,1992 (3) :60 - 95

18　吴敬琏. 电子政务:用信息化手段推进社会主义民主政治建设[EB]. http :/ / www. 50forum. org. cn/ meeting. asp ? meetId = 118 #

(责编:刘影梅)

(上接第 130 页)大量的研究基金项目[12],产生和掌握了大量的科学数据资源。因此,高等院校要参与到科学数据共享工程的建设和研究中来,明确其拥有的权利和应承担的义务,为科学数据共享提供意见和建议,要和各研究机构一起共同推动科学数据共享的全面协调开展。

参 考 文 献

1　Oecd Principles and Guidelines for Access to Research Data From Public Funding [EB]. http :/ / www. oecd. org/ dataoecd/ 9/ 61/ 38500813. pdf ,2007 - 12 - 06

2　科学数据共享工程简介[EB]. http :/ / www. sciencedata. cn/ index. php ,2007 - 12 - 06

3　科学数据共享调研组. 科学数据共享工程的总体框架[J]. 中国基础科学,2003(1) :63 - 68

4　孙鸿烈,刘　闯. 国际科学技术数据前沿领域发展研究[J]. 中国基础科学,2003(1) :18 - 23

5　科学数据共享工程标准规范研究[EB]. http :/ / www. sciencedata. cn/ index. php ,2007 - 12 - 06

6　Oecd Follow Up Group on Issues of Access to Publicly Funded Research Data. Promoting Access to Public Research Data for Scientific , Economic , and Social Development [EB]. http :/ / dataaccess. ucsd. edu/ Final _ Report _ 2003. pdf ,2007 - 12 - 06

7　路　鹏. 我国科学数据共享现状[J]. 国际地震动态,2007(6) :26 - 32

8　数据共享重在机制建设[EB]. http :/ / shgy. jhgl. org/ shownews. asp ? newsid = 923 ,2007 - 12 - 06

9　李长玲,化柏林. 我国网络计量学研究的文献计量分析[J]. 图书情报工作,2006(9) :46 - 50

10　Jeremy Birnholtz ,Matthew Bietz. Data at Work: Supporting Sharing in Science and Engineering[EB]. http :/ / www. crew. umich. edu/ publications/ 03 - 01. pdf ,2007 - 12 - 06

11　王正兴,刘　闯. 科学数据可持续共享:关键是利益的均衡[J]. 中国科技论坛,2005(6) :92 - 96

12　陈传夫. 中国科学数据公共获取机制:特点、障碍与优化的建议[J]. 中国软科学,2004(2) :8 - 11

(责编:白燕琼)