基于文献计量学的国际地球科学数据管理研究进展

王淑强[1]，王卷乐[1,3,5*]，李 扬[2,4*]，王 晶[1]，王玉洁[1]，李海英[2]

1. 中国科学院地理科学与资源研究所，北京 100101；2. 中国科学院文献情报中心，北京 100190；
3. 中国-巴基斯坦地球科学研究中心，伊斯兰堡 45320，巴基斯坦；4. 中国科学院大学经济与管理学院，图书情报与档案管理系，北京 100190；5. 江苏省地理信息资源开发与利用协同创新中心，南京 210023

摘 要：科学数据是重要的科技基础条件和国家战略资源。随着大数据时代的到来，全球各国纷纷将科学数据管理纳入到国家发展战略。面对地球科学数据管理的紧迫需求，首先要深入了解地球科学数据管理研究的现状和发展态势。本文基于 Web of Sciences 数据库，遴选了被科学引文索引（SCI）、社会科学引文索引（SSCI）和会议录索引（CPCI）收录的文献，采用文献计量分析方法，对 1900–2018 年国际地球科学数据管理研究相关文献进行统计分析，揭示了全球国家/地区和研究机构的论文产出数量、论文影响力、机构科研实力，以及学科分布情况。借助知识图谱分析发现，地球科学数据管理发展为六大研究主题领域。其中，近 10 年来地理空间数据管理研究形成了最为完备的理论和方法体系。“科研数据的开放获取政策研究”发展迅速成为热点。未来，地理空间数据管理研究将会对地球科学大数据研究、数据管理决策模型等研究领域起到引擎驱动力的作用。

关键词：地球科学；数据管理；文献计量；研究热点；发展态势与研究进展
DOI: https://doi.org/10.3974/geodp.2020.03.13

1 前言

科学数据是在科技活动中所获取的反映客观世界的本质、特征、变化规律等的原始基本数据。科学数据已经成为科技创新、经济发展和相关决策活动不可缺失的基础科技支撑条件，被公认为继物质和能量之后的第三类资源，亦视为重要的国家科技战略资源[1–2]。地球科学研究是典型的数据密集型科学研究，其在解决科学和应用问题的过程中需要大量的科学数据支撑，同时又在相关科研活动中不断产出新的衍生数据和产品[3–4]。因此，地球科学数据管理研究对推动我国地球科学乃至全学科科学数据管理的发展具有十分重要的战略意义。

地球科学领域数据的来源主要包括两种：一是通过对地球科学研究与实践直接获取的科研数据，二是政府部门长期采集和管理的业务数据，如国土资源管理部门的地质调查数据、水利部的水文数据、气象局的气象和气候数据等。地球科学数据管理是利用计算机硬

收稿日期：2020-08-10；**修订日期：**2020-09-20；**出版日期：**2020-09-25
基金项目：国家自然科学基金项目（41842061）；中华人民共和国科学技术部（2020WT22），中国科学院（XXH13505-07）
***通讯作者：**王卷乐，中国科学院地理科学与资源研究所，wangjl@igsnrr.ac.cn；
李扬，中国科学院文献情报中心，liyang@ mail.las.ac.cn
作者 ID：王淑强 0000-0002-2432-8161，王卷乐 0000-0002-5641-0813 李扬 0000-0002-2890-9000 王晶 0000-0002-9669-358X 王玉洁 0000-0002-7531-2880 李海英 0000-0002-1520-516X
引用格式：王淑强，王卷乐，李扬等. 基于文献计量学的国际地球科学数据管理研究进展[J]. 全球变化数据学报，2020, 4(3): 299–313. https://doi.org/10.3974/geodp.2020.03.13.

件和软件技术对地球科学数据进行有效的收集、存储、处理和应用的过程，其目的在于充分有效地发挥数据的作用，对这两类数据进行有效的管理，并促进其广泛共享，使之价值最大化[5]。科研人员不仅要基于对数据实时、动态的监测与分析来处理难以解决或不可触及的科学问题，更将数据作为科学研究的对象和工具，基于数据来思考、设计和实施科学研究[6]，这也引发了一系列具有地球科学领域研究特殊的科学数据管理问题。

20 世纪中期以来，地球科学数据管理与共享逐渐引起国际科学界的关注。1957 年，在国际科学联合会理事会的组织下成立了以地球科学、空间科学和天文学数据为重点的世界数据中心[7]。1960 年，美国成立国家大气研究中心，最早开始了对地球科学数据的建模、收藏和保存工作[8]。1969 年 White[9]回答了为什么要进行地球物理科学数据管理，地球科学的数据管理研究已经成为驱动科学发现和决策支持的重要科学平台，相关研究问题集中在数据的存储、共享、管理政策与信息挖掘。科学数据的开放共享为科研成果的广泛传播和重复利用打通了渠道，在国际科学联合理事会的组织下，美国及欧洲的一些发达国家建立了国家级科学数据中心群和数据共享服务网络，如美国航空航天局（NASA）主持的全球变化数据和信息系统（DAACs），全球变化主目录（GCMD）等[10–12]。多元化的数据形态对管理提出了很大挑战。随着数据密集型科研范式兴起，以实体数据为核心的数据出版、数据仓储及数据枢纽受到了许多科研机构与学者的关注。地球系统科学数据（ESSD）于 2009 年起发行出版，目前已与德国海洋数据中心（PANGAEA）等多个数据中心合作建设数据存储库[13]。2019 年美国地球物理学会（AGU）启动期刊数据存储计划，要求旗下学术期刊公开出版与论文关联的原创数据，且数据必须存储于 AGU 认定的 226 家数据仓储中心[14]。作为地球观测领域规模最大、最具权威和影响力的政府间国际组织，地球观测组织（GEO）发起和推动数据枢纽建设，预期通过一个建立在云端的平台，将开放获取的数据、论文、算法、模型和计算能力融合在一起。我国地球科学数据管理研究内容与国际研究相似，但起步较晚。直到 1981 年，李善芳等[15]将美国地学 STATPAC 数据管理系统概念引入我国，这对开启我国地球科学数据管理研究具有重要意义。之后，国内外科研人员持续关注我国地球科学数据管理及其对地球科学的影响。1996 年李军等[16]指出，地球科学数据的元数据系统的建立有助于地球科学数据的开发和利用，说明了元数据在地球科学数据管理中的应用。2002 年孙九林等[17]指出，地球信息科学也因计算机和遥感技术而产生了革命性的进展，每天获得 Tb、Pb 量级的地学数据得不到有效利用的问题日益困扰广大地学工作者，将网格技术引入地学数据存储和共享系统将有助于解决这一难题。2003 年我国地球化学数据管理信息系统升级版问世[18]。之后，杜云艳等[19]基于来自空间和常规的海岸带及近海多源信息，旨在满足国家空间数据基础建设和应用的迫切需求，在对诸多信息特征进行分析的基础上，建立了中国海岸带科学数据平台概念模型，并在此概念模型的基础上进行了具体的逻辑结构、遥感影像数据的 ARCSDE 存储、遥感数据的元数据存储等模型设计。王卷乐等[20]针对地学数据共享的实际需求，分析通用地学元数据的构架模式与方法，由此构建的元数据框架包括 3 个层次，即核心元数据、模式元数据和应用领域专用元数据。肖建华等[21]针对当前国内地理时空大数据生产管理与应用面临的数据存储组织难、数据吞吐处理难、数据集成应用难等问题，同时从地理实体产生消亡与地理数据生产服务两个全过程管理角度出发，研究了地理时空大数据全生命周期管理与应用的相关方法。

顺应全球科学数据出版与存储发展的大趋势，《地理学报》2014 年增刊、《中国科学数据》、《全球变化数据学报（中英文）》、《Big Earth Data》等数据期刊率先建立了较完善的数据论文审核、存储、同行评议等流程，快速推动国内数据出版。其中，“全球变化科学研究数据出版系统”（中英文）在保护数据知识产权和促进数据共享方面做出实践性案例，在数据产权认证、数据质量标准、同行专家评审、数据长期保存、数据开放共享、国际资质联网等方面的成果为数据的增值起到关键性作用[22]。2018 年以来，国内中国天文数据中心、可再生资源与环境世界数据中心、地质科学数据出版中心、中国地球物理学科中心、国家空间科学数据中心先后成为国际认可的数据仓储、出版中心或数据枢纽[23]。2020 年 3 月 4 日中共中央中央政治局常务委员会召开会议，指出要需加快数据中心等新型基础设施建设进度，作为数字经济时代的枢纽，建设大数据中心已经成为大势所趋。

国内外地球科学数据管理研究已经历了几十年的发展并积累了一定的研究成果，但多是在具体领域的技术方法进展，缺少从文献计量视角对地球科学数据管理研究的综合分析。面向新时期地球科学领域大数据技术和数据管理规范化的发展需要，本文拟开展国际地球科学数据管理发展态势和研究进展分析，为进一步促进和发展我国地球科学数据管理提供决策参考。

2　数据源与研究方法

2.1　数据源

地球科学的研究对象是地球的整体，是为人类合理开发自然资源、充分利用自然条件，避免和减轻自然灾害，适应自然规律，使人口、资源、环境相互协调，可持续发展的重要基础自然科学[24]。地球科学数据管理既包含自然科学又涉及人文社会及管理科学，因此本文文献数据选择了覆盖面较广、影响力较大的被 SCI、SSCI 和 CPCI 索引收录的 1900–2018 年发表的地球科学数据管理研究相关文献。

2.2　检索思路与策略

“地球科学数据管理”的主题词可分解为地球科学和数据管理。因此，本文适宜采用地球科学学科和数据管理主题组合检索策略。地球科学学科检索词包括 20 个学科分支：生态环境科学、地球化学与地球物理学、地质学、遥感、天文学和天体物理学、气象学与大气科学、公共环境职业卫生、水资源、农业、自然地理学、海洋学、采矿选矿、林学、渔业、地理学、矿物学、城市研究、区域研究、生物多样性保护、影像科学摄影技术。数据管理主题检索词则依据表 1 中要素和数据管理相关法律法规、政策加以限定。

2.3　数据处理与分析指标

2019 年 6 月 20 日，依据上述检索策略，得到 3,202 条文献。经专家识别排除不相关文献，共得到 2,391 条文献记录。为了更加精确地进行数量统计，我们对机构和关键词等信息进行了清洗。然后采用科睿唯安的 DDA 文本挖掘软件、微软公司的 Excel 软件，以及美国德雷塞尔大学信息科学与技术学院陈超美研发的 CiteSpace 软件[25]和荷兰莱顿大学科学技术研究中心研发的 VOSviewer 软件[26]等分析工具，定量分析全球数据管理研究的综合发展态势，以及研究领域的进展情况。

本研究采用的评价指标主要包括发文量、总被引频次、平均被引频次等。发文量是指某一特定范围内科研工作者、科研机构或国家在一定时间内发表的文献数量。总被引频次是指检索到的某一特定范围内的所有文献被引次数。平均被引频次是指在某领域检索到的一定时间内某一特定范围的所有文献被引次数与文献记录数的比值。

表 1 数据管理主题列表

1.**数据治理**	7.**数据仓储和业务智能管理**
数据资产	业务智能
数据治理	数据集市
数据专员	数据挖掘
2.**数据架构、分析和设计**	数据移动（提取、转换、加载）
数据分析	数据仓储
数据架构	8.**文件、记录和内容管理**
数据模型	文件管理系统
3.**数据库管理**	记录管理
数据库管理	9.**元数据管理**
数据库管理系统	元数据
数据维护	元数据发现
4.**数据安全管理**	元数据出版
数据存取	元数据注册
数据擦除	10.**接触数据管理**
数据保密	业务持续性规划
数据安全	市场运作
5.**数据质量管理**	用户数据集成
数据清洗	身份管理
数据完整性	身份盗取
数据丰富度	数据盗取
数据质量	ERP 软件
数据质量保证	CRM 软件
6.**参考数据和主数据管理**	地理位置
数据集成	邮编
主数据管理	电子邮件
参考数据	电话号码

注：该表来源于 https://encyclopedia.thefreedictionary.com/data+management

3 结果与分析

3.1 基于论文产出和引文数据的全球综合态势分析

3.1.1 研究热度与总体影响力分析

某研究领域的年发文量可以在一定程度上反映该领域的研究热度。图 1 展示了全球地球科学数据管理研究领域发文量趋势。最早的一篇文献可以追溯至 1953 年，《Sources of Legal Information in Poland》，发表在英文期刊《Law Library Journal》上。从 1953 年到 1974 年，该领域全球文献数量处于 10 以下。1975–1997 年期间发文量虽有波动但整体呈现上升状态。1997–2001 年期间发文量略有下降。2002 年至今发展较快，发文量呈上升趋势，其中 2014 年发文 99 篇，研究热度达到历史最高点。近几年发文量仍保持在高位。

研究成果的影响力一定程度上可以通过引文来判定。地球科学数据管理研究从 1953 年第一篇发文至 2018 年共跨越 66 年，本文以 5 年跨度分段（最后一段为 6 年）计算平均引用次数。如图 2 所示，在 2002 年前，平均被引频次在 0–3.5 之间波动变化，从 2003 年开始，平均被引次数快速上升，2008–2012 段达到最大，为 10.85 次/篇，影响力达到峰值。

然而，与地球科学整体研究影响力相比，地学科学数据管理方面的影响力仍不高。表 2 显示了近 10 年地学科学数据管理领域平均被引次数与 ESI 公布的地球科学被引基线对比。可以看出，地学科学数据管理领域近 10 年只有 2010 年和 2014 年超过了 ESI 的 50%基线。

发文期刊的分布也可反映文献影响力的一个侧面。按照文献类型统计，地球学数据管理领域 2,391 篇文章发表在 342 个期刊上。其中 249 个期刊可在最新版 2018 版 JCR 中查得影响因子，包括 6 个中国期刊。这些期刊的影响因子区间分布如表 3 所示。影响因子大于 7 的期刊有 5 种：《Environmental Health Perspectives》（8.309），《Frontiers In Ecology and the Environment》（8.302），《Bulletin of the American Meteorological Society》（7.804），

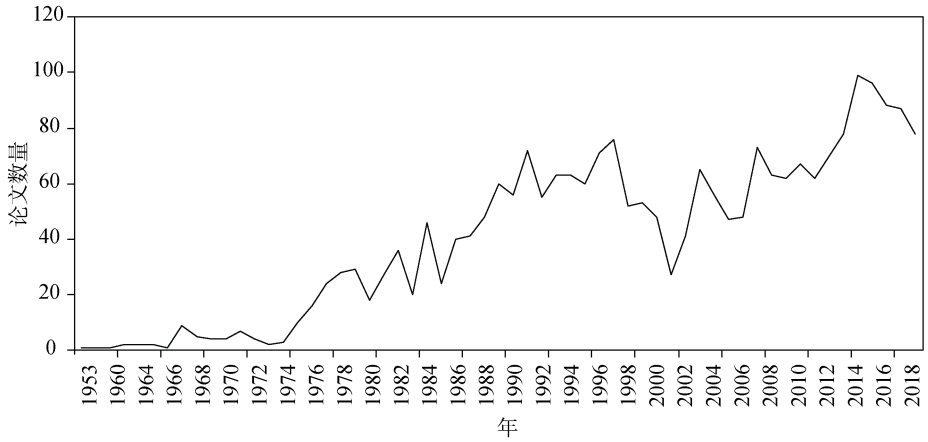

图 1　地球科学数据管理研究发文量趋势

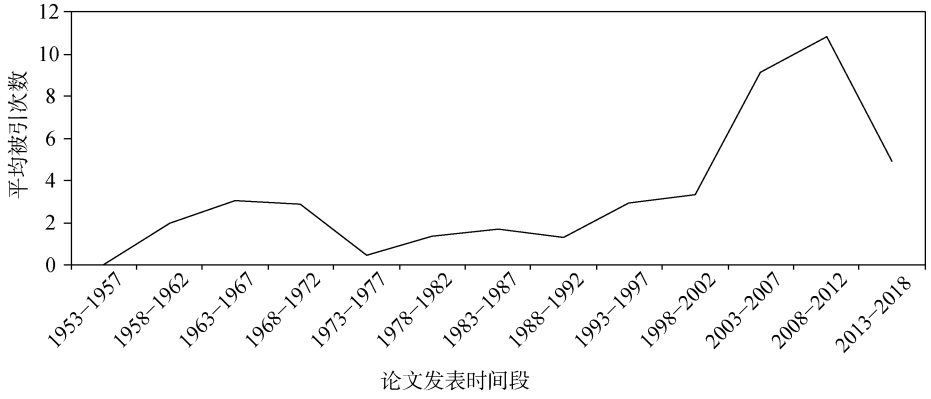

图 2　各时区平均引用次数

表 2　2009–2018 年 ESI 被引基线及地学科学数据管理研究领域平均被引次数

学科	基线	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
地学	0.01%	2,028	1,401	1,211	696	759	541	567	256	110	43
	0.10%	609	530	448	372	328	239	174	108	55	22
	1.00%	195	163	153	131	111	85	63	44	24	9
	10.00%	59	52	48	41	35	28	22	15	9	3
	20.00%	37	33	30	27	23	19	15	10	6	2
	50.00%	15	13	12	11	9	8	6	4	3	1
平均被引次数		12.82	18.07	6.74	9.24	8.24	9.04	5.74	3.52	2.13	0.88

《Conservation Letters》(7.279),《Water Research》(7.051)。多数期刊分布在 2≤IF<4 和 1≤IF<2 这两个区间内。1≤IF<2 区间的期刊发文量是 378 篇，居第一位；2≤IF<4 区间期刊发文量 327 篇，位居第二。

3.1.2 国家/地区与机构科研实力分析

（1）国家/地区发文排名

地球科学数据管理研究领域发文分布在全球 88 个国家/地区。如图 3 所示，在 TOP20 国家发文量有显著差异，其中，美国研究独占鳌头，发文为 655 篇，占总发文量的 27.39%。英国、中国发文量虽然分别排名第二、三位，发文 177 篇和 123 篇，但两国占比之和（英国 7.40，中国 5.14%）不到美国的一半。其后是德国（89），加拿大（75）等。

表 3 各影响因子区间内期刊数量及发文量

影响因子区间	期刊数量	发文量
IF≥7	5	11
4≤IF<7	29	237
2≤IF<4	82	327
1≤IF<2	74	378
IF<1	59	258

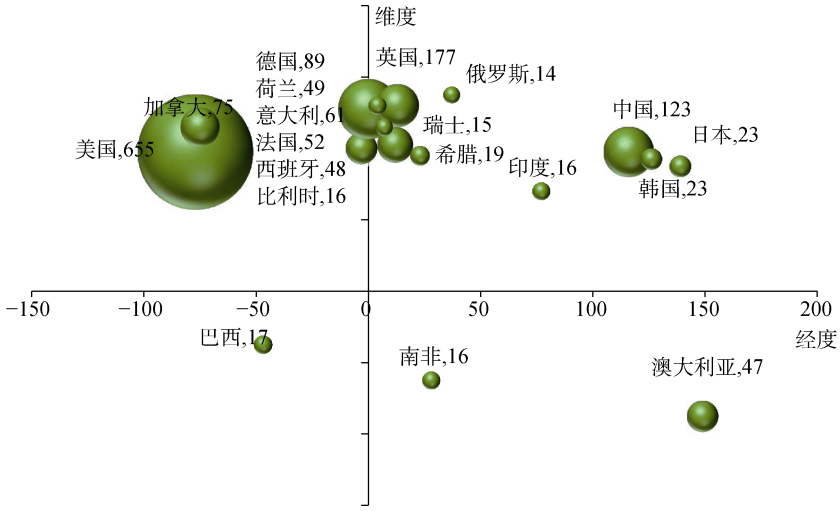

图 3 TOP20 国家论文产出数量分布

如图 4 所示，在 TOP20 国家中，美国从 1966 年就开始对该领域进行研究，研究时间最早。其后是英国、意大利、法国，始于 1972 年，比利时为 1975 年，加拿大、澳大利亚、印度为 1977 年。中国相比其它国家起步较晚，但研究发展迅速，从 1998 年发表第一篇国际刊物论文开始，到 2004 年之后发文量快速上升，可见中国近几年在该领域研究热度比较高。美国作为最主要的研究阵地，从 1966 年发表第一篇文章后，一直处于低速不平稳发展状态，直到上世纪 90 年代后，发文量和研究热度保持较高水平。

（2）重点机构科研实力分析

对地球科学数据管理研究领域全部论文的第一作者机构进行统计，发文量大于等于 6 篇的机构共有 25 个，下称 TOP25 重点机构，如表 4 所示。重点机构集中分布在美国（17）、英国（5）、中国（2）、加拿大（1），各国机构占比分别为 68%、20%、8%、4%。其中，中国进入 TOP25 的两个机构是武汉大学（11 篇）和北京大学（7 篇）。

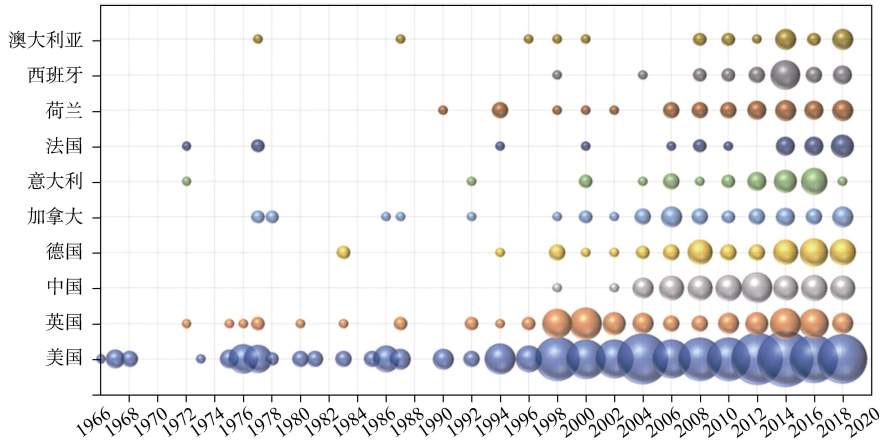

图 4　TOP20 国家论文产出-时间矩阵

表 4　TOP25 第一作者机构发文情况

第一作者机构	发文量	被引次数	被引占比（%）	篇均被引
拉夫堡大学	29	117	1.21	4.03
伊利诺伊大学芝加哥分校	16	56	0.67	3.50
美国国家海洋与大气管理局	15	19	0.63	1.27
武汉大学	11	11	0.46	1.00
肯塔基大学	10	70	0.42	7.00
伦敦城市大学	9	53	0.38	5.89
佛罗里达州立大学	9	83	0.38	9.22
印第安纳大学	9	110	0.38	12.22
伊利诺伊大学厄巴纳-香槟分校	9	56	0.38	6.22
谢菲尔德大学	9	118	0.38	13.11
南加州大学	8	58	0.33	7.25
惠灵顿维多利亚大学	8	7	0.33	0.88
密歇根州立大学	7	103	0.29	14.71
北京大学	7	7	0.29	1.00
密歇根大学	7	56	0.29	8.00
多伦多大学	7	17	0.29	2.43
哥伦比亚大学	6	304	0.25	50.67
美国国家光学天文台	6	12	0.25	2.00
普渡大学	6	24	0.25	4.00
纽约州立大学奥尔巴尼分校	6	186	0.25	31.00
伦敦大学学院	6	2	0.25	0.33
马里兰大学	6	67	0.25	11.17
匹兹堡大学	6	68	0.25	11.33
田纳西大学	6	58	0.25	9.67
威斯康星大学	6	32	0.25	5.33

发文量排名第一的拉夫堡大学首次在1998年发文4篇，虽然发文较晚，数量却较多；1998–2009是拉夫堡大学的研究活跃期，到2010年发文减少，近三年没有发表论文；美国伊利诺伊大学芝加哥分校的发文活跃期是1998–2007年，2008年后发表论文明显减少；美国国家海洋和大气管理局的研究活跃期为2002–2010年；中国的武汉大学研究活跃期为2003年至今；肯塔基大学的发文开始时间更晚，但从2014年至今一直处于研究活跃期。

拉夫堡大学虽然发文数量排名第一，篇均被引只有4.03，低于领域平均被引次数（6.55）；哥伦比亚大学虽然发文较少（6篇），篇均被引为50.67，居首位；纽约州立大学奥尔巴尼分校，发文数量为6，篇均被引为31。TOP25重点机构中，共有11个机构的篇均被引次数超过领域平均被引次数，这些机构大多来自美国。中国的武汉大学和北京大学，篇均被引均为1。

3.1.3 学科分支及其关系分析

根据“Web of Science Categories”学科分类标准，本研究对地球科学数据管理研究领域的文献进行了学科统计，列出了文献数量排名前20的学科（简称TOP20学科）名称，详见表5。统计结果显示，全球地球科学数据管理研究论文共涉及118个学科。其中，信息科学与图书馆学论文最多，计1,411篇；其次是信息系统计算机科学，计579篇文章。两个学科论文数累计1,990篇，占总论文数的绝对比例。

表5 TOP20学科发文量

学科	论文数	学科	论文数
信息科学与图书馆学	1,411	通讯	78
信息系统计算机科学	579	电气和电子工程	76
环境科学	223	自然地理学	73
遥感	188	地理学	65
跨学科应用计算机科学	164	环境工程	62
多学科地球科学	150	海洋学	62
成像科学与摄影技术	117	法学	58
水资源	112	天文学和天体物理学	56
环境研究	101	人工智能计算机科学	55
电信	98	管理	54

将地球科学数据管理研究领域发文量排名前15（简称TOP15）的地球科学WoS学科及其与“地球系统科学数据共享平台分类编目体系”一级分类[27]对应关系列入表6。TOP15学科总发文量1,174篇，其中发文量超过100篇的WoS分支学科是环境科学、遥感、地质学、多学科综合地学、水资源和环境研究。按照该分类，可将TOP15WoS分支学科归为6个一级分类。其中陆地表层包括8个分支学科，论文数计为703篇，占TOP15论文总数的59.9%；遥感数据188篇，占比16.0%；自然资源包括3个分支学科，论文数计为154篇，占比13.1%；海洋62篇，占比5.3%；大气圈47篇，占比4.0%；固体地球与古环境20篇，占比1.7%。

需要指出，Web of Science学科分类体系是复分体系，一篇文章有可能属于多个学科。

为了揭示这种关系，本文绘制了 TOP15 学科的论文所属学科关系图谱，结果如图 5 所示。图中圆圈大小代表了学科发文数量，圆圈间连线粗细代表了同时属于两个学科以上的论文的相对数量。可以看出，环境科学与研究、遥感领域不仅发文量大，而且与其他学科联系紧密，尤其是遥感与环境研究及自然地理学关联非常密切。需要指出，Web of Science 学科分类体系是复分体系，一篇文章有可能属于多个学科。为了揭示这种关系，本文绘制了 TOP15 学科的论文所属学科关系图谱，结果如图 5 所示。图中圆圈大小代表了学科发文数量，圆圈间连线粗细代表了同时属于两个学科以上的论文的相对数量。可以看出，环境科学与研究、遥感领域不仅发文量大，而且与其他学科联系紧密，尤其是遥感与环境研究及自然地理学关联非常密切。

表 6　地球科学 WoS 分支学科 TOP15 学科分布及其与一级分类对应关系

序号	学科分类	记录数量	一级分类
1	环境科学	223	陆地表层
2	遥感	188	遥感数据
3	多学科地学	150	陆地表层
4	水资源	112	自然资源
5	环境研究	101	陆地表层
6	自然地理学	73	陆地表层
7	地理学	65	陆地表层
8	海洋学	62	海洋
9	气象学与大气科学	47	大气圈
10	生态学	43	陆地表层
11	城市研究	26	陆地表层
12	多样性保护	22	陆地表层
13	林学	22	自然资源
14	多学科农业	20	自然资源
15	地球化学与地球物理	20	固体地球与古环境

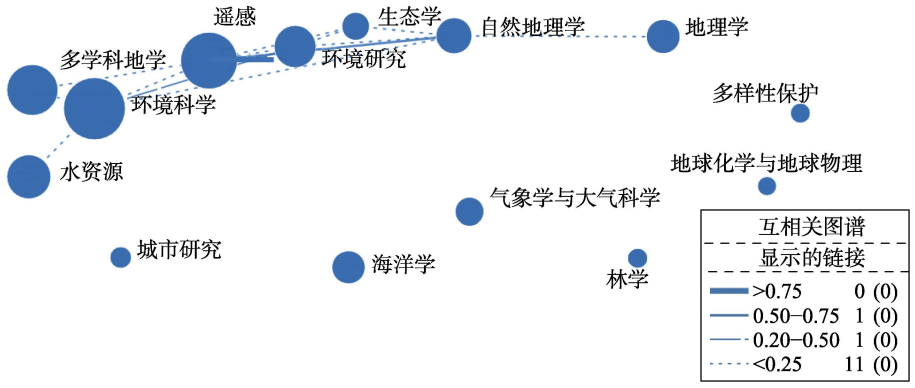

图 5　TOP15 重点学科关系图谱

3.2　基于图谱分析的研究领域划分及进展评述

3.2.1　研究领域划分

基于文献检索结果，将地球科学数据管理研究领域所有数据视为一个数据集，采用

DDA 软件将论文的“作者关键词”字段经过机器与人工清洗，从 2,609 个关键词中选取出现频次大于 4 次的 101 个关键词作为分析对象，利用 VOSviewer 软件对出现的高频主题词数据进行聚类，生成关键词共现关系网络图谱（图 6）。分析此图，结合专家判读，可将所有关键词归纳为 6 个研究领域：（1）面向环境监测大数据的管理、集成与共享机制研究，（2）信息政策的主要内容与演化综述研究，（3）支持决策的管理模型与系统研究，（4）科研数据的开放获取政策研究，（5）地理信息系统（GIS）作为地球科学数据管理工具的应用研究，（6）信息安全政策制定研究。

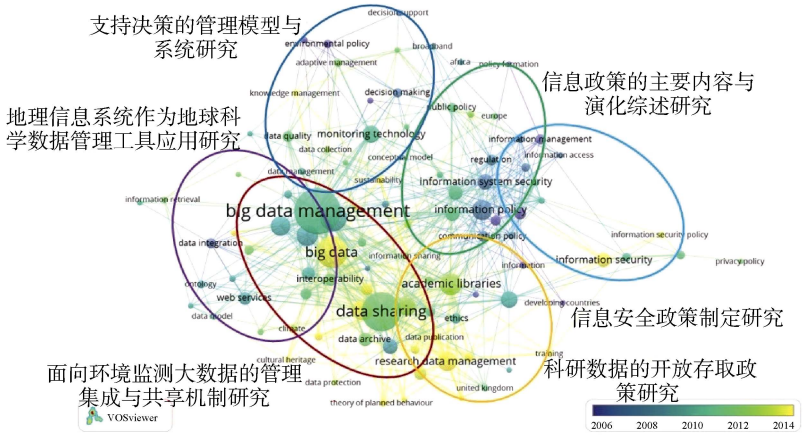

图 6 地球科学数据管理研究领域时间演变图

3.2.2 研究领域关联度分析

表 7 列出核心主题数的多少顺序列出了上述各研究领域的聚类参数。分析结果中的核心主题词平均被引频次代表包含此主题核心主题词的论文发表以来的平均被引频次；平均关联强度代表这个簇包含的核心主题词间联系的紧密程度，某个簇的平均关联强度越大代表核心主题词间共现强度越大、研究越集中，反之则代表共线强度相对较低、研究越分散。核心主题词的总关联强度代表此主题词在共现网络中的重要程度，越高则说明此主题词对于构建网络越重要。

表 7 研究领域聚类参数

编号	研究主题	核心主题数量	平均出现时间	平均被引频次	平均关联强度
1	面向环境监测大数据的管理、集成与共享机制研究	25	2011	7.52	23.32
2	信息政策的主要内容与演化综述研究	20	2008	12.6	15.35
3	支持决策的管理模型与系统研究	16	2010	15.47	10.06
4	科研数据的开放获取政策研究	14	2012	16.56	20.71
5	GIS 作为地球科学数据管理工具的应用研究	13	2010	8.32	13.54
6	信息安全政策制定研究	12	2008	17.82	8.58

“面向环境监测大数据的管理、集成与共享机制研究”的平均关联强度最高，是研究内容最集中的主题，与其他研究内容交叉相对较少；其次是“支持决策的管理模型与系统研究”，主要研究内容集中在决策支持的政策研究及相关系统的建设；“信息安全政策制定研究”的平均关联强度最低，是研究内容最发散的主题，与其他研究内容交叉相对较多，涉及地球科学数据获取、管理、集成多个环节，以及计算机科学、密码学等。

3.2.3　各领域研究进展分析

如表 7 所示，地球科学数据管理 6 个研究领域中，形成最早的是“支持决策的管理模型与系统研究”和“信息安全政策制定研究”，均出现在 2008 年。出现最晚的领域是“科研数据的开放获取政策研究”，见于 2012 年，相对而言是新兴研究主题。

（1）面向环境监测大数据的管理、集成与共享机制研究

国际地球物理年（IGY）（1957–1958）和国际生物学计划（IBP）（1964–1974）是现今生态环境大数据研究的雏形，被称为“大科学研究”，目的是获得较为可靠的大量观测数据，以研究地球各圈层和生态环境问题。这些研究最后演变成如今以长期定位观测为基础的生态系统研究网络，从而全面获取有关生态环境的观测数据。2008 年，Nature[28]等学术刊物相继出版专刊探讨大数据议题，标志着大数据研究得到世界范围内的关注和认可。从表 7 可知，该领域核心主题数为 25 个，其中核心关键词是“大数据管理”、“数据共享”、“监测技术”、“数据档案”等。在地球科学数据管理所有领域引用频次最高的 100 篇论文中，涉及生态环境大数据的文章有 15 篇，主要研究内容是针对环境科学领域各类监测系统获取的大数据进行管理制度、集成方法与共享机制的研究。大数据覆盖的领域包括各种大型巡天观测数据[29]、流域与大气污染监测数据[30]、农业资源与生产相关数据[31]、气象数据[32]和海洋数据[33]等。数据管理制度的研究与管理系统的开发紧密结合；数据集成方法的开发也涉及到网络基础设施的建设；数据的共享机制需要关注的内容包括版权、隐私问题和协作制度。本领域形成的平均年份是 2011 年，到目前是中外学术界持续的关注热点。

（2）信息政策的主要内容与演化综述研究

该领域核心主题数为 20 个，核心关键词是“信息政策”，“公共政策”，“政策信息”，和“模型”等，主要研究内容是分析与总结各国的信息政策制定与实践，多为综述论文[34]、公众的意见[35]等。最受关注的国家与地区是中国与欧洲，另外法国和英国也是主要的研究对象。政府层面的信息政策包含了知识产权政策、通讯政策、公众信息传播政策、信息获取政策等。信息政策概念的提出时间最早可追溯到上个世纪 90 年代[36]，是数据管理研究最早形成的议题。之后，逐步关注信息政策及其实施效果的研究[37–38]，系统互操作、开放获取系统、信息处理过程等方面的技术标准问题，以及制定完备的信息技术标准政策建议[39]等。

（3）支持决策的管理模型与系统研究

该领域核心主题数为 16 个，核心关键词是“决策制定”，“决策支持”，“管理模型”，“环境政策”等，主要研究内容是以决策支持为目标设计管理模型，核心问题包括数据的获取、数据质量、可持续性和知识管理[40]。另外，还重视开发相应管理平台等技术要点研究。此类决策支持研究主要应用在环境监测与保护相关的咨询问题[29]，另外也有电动汽车和医疗相关问题的案例[41]。

（4）科研信息的开放获取政策研究

该领域核心主题数为14个，核心关键词是“研究数据”，“研究数据管理”，“研究数据共享”，“数据存取”，“数据共享行为”等，平均被引频次16.56，仅次于“信息安全政策制定研究”。本领域的主要研究内容是高校、研究机构与政府产生或购买的科研数据的开放获取相关研究，具体问题包括管理支持、共享机制、实现途径、相关立法与服务许可。涉及的科研信息既包括学科研究直接产生的数据也包括通过各种方式购买的资源。

很多学者对科学数据共享行为和意图进行多方位的调查研究[42–46]。早在1985年Fienberg等就提出了科研数据的共享，1995年一些作者认为科研数据的强制共享政策是必要的[47]，2003年出版物共享引人关注，同年美国发布了《科学文献公共存取法案》（Public Access of Science Act，俗称Sabo法案）。该提案要求修改美国现行版权保护法律，免除对受公共资助的研究成果的版权保护。2011年左右原创科学数据出版与共享获得关注[48]。直到2013年2月，开放存取的支持者又提出了FRPAA（Federal Research Public Access Act）的强化版《公平获取科学和技术研究成果法案》（Fair Access to Science and Technology Research (FASTR) Act），该备忘录对科学数据开放存取的发展具有里程碑式意义[49]。此外，英国、丹麦等政府对于开放存取政策的制定相当重视。特别是英国，通过英国研究理事会（Research Councils UK，RCUK）初步形成了国家性的开放存取政策，并在实施中吸收各方意见，逐步完善。本领域形成的平均年份是2012年，到目前仍是中外学术界的研究热点。

（5）地理信息系统作为地球科学数据管理工具应用研究。

该领域核心主题数为13个，核心关键词是“地球空间数据”，“地学数据管理”，“图形数据库”，“拓扑学”，“数据整合”等，主要是以地理信息系统作为核心工具或研究对象展开的。研究的问题包括地理空间数据的集成与管理、多主体间的地理空间数据共享机制、相关管理系统的开发、地理服务的设计等。重点讨论的技术问题包括系统的互操作性、数据的标准化与集成、网络技术的开发、空间数据获取的基础设施建设、语义检索等。实践领域包括矿产资源探测、水资源管理、航道数据管理、生物多样性保护、建筑遗产勘察等领域。此外，在欧洲和中亚有一定实践研究。上述研究内容可以统称为“地理空间数据管理”。近十年来地理空间数据管理研究取得四大成就。第一，在数据、过程和应用层面推进GIS/BIM集成[50–54]，提高了地理空间数据管理水平。第二，将拓扑学作为地理空间数据管理的一个关键概念，解决了实体关系模型的构建[55–56]。第三，在3D/4D地理空间数据库建设方面取得了重大进展。如使用n维空间填充曲线进行查询的并行化已经被验证[57]。第四，地学人工智能（geoAI）方法用于地理空间数据管理，为包括地球传感器数据源的密集使用提供更有效的解决方案[58]。本领域形成的平均年份是2010年，具有广阔的发展前景。

（6）信息安全政策制定研究

该领域核心主题数为12个，核心关键词是“信息安全”，“安全政策”，“系统政策”，“研究框架”，“信息科学”，“公共管理”等，平均被引频次17.82，居6个研究领域之首。本聚类的研究主题是在复杂网络环境下的面向地理科学信息的网络安全策略制定[59–60]。具有多学科交叉研究特点，涉及的学科包括密码学、公共管理、公共政策制定等。涉及的数据信息持有者包括高校、政府、商业机构。此外，本领域研究在非洲的水资源管理和农业生产事务上有一定应用。

4　结论与展望

自数据管理概念提出至今，地球科学数据管理研究相关的论文数量及其影响力虽有波动但整体呈现上升态势。其中发文量前期增长较慢，2002 年开始发展较快。相关领域的发展大致经历三个阶段：1953–1974 年为萌芽阶段，1975–1997 年为成长阶段，1998 年后为形成与发展阶段。虽然美国于 1966 年率先确立了首个数据管理法规——《信息自由法》，但地球科学数据管理相关研究已在波兰早之十几年前就开始了。之后，地球科学数据管理相关研究热度虽增加较快，但也并非想象的那么“过热”。以此可以推测国际地球科学数据管理研究持续滞后于地球科学研究的发展，也远远低于地球科学整体研究影响力。我国应进一步加大地球科学数据管理研究力度，力争以最短的时间进入世界领先行列。

从发文量、发文重点机构、论文被引频次等多项综合指标来看，美国在地球科学数据管理研究领域处于领先地位，其次是英国。中国在该领域的研究正在兴起，发文量仅次于英国，发展提升空间较大。TOP25 重点机构中，拉夫堡大学、伊利诺伊大学芝加哥分校、美国国家海洋和大气管理局、武汉大学、肯塔基大学发文量较多。哥伦比亚大学、纽约州立大学奥尔巴尼分校、密歇根州立大学篇均被引较高。这说明研究热度较高的机构未必是影响力较大的机构。地球科学数据管理研究领域涉及的非地球科学学科，如信息科学与图书馆学、信息系统计算机科学等，具有较大的数据体量或信息技术优势的领域发展很快。这充分说明信息科学与图书馆学、信息系统计算机科学等已成为地球科学数据管理研究的主要理论和方法的基础。基于图谱分析，得出地球科学数据管理的 6 个研究领域分别是面向环境监测大数据的管理、集成与共享机制研究；信息政策的主要内容与演化综述；基于监测数据的决策支持研究；科研数据的开放获取政策研究；地理信息系统作为地球科学数据管理工具应用研究；信息安全政策制定研究。其中，地理空间数据管理已经发展成为一个跨学科的科学领域，已具备科学的方法、过程、算法和系统，能够从非结构化数据和结构化数据中提取知识、模式和结论。地理空间数据管理研究将会对地球科学大数据研究、数据管理决策模型等研究领域起到引擎驱动力的作用。未来地理空间数据管理的重点研究方向是（1）语义和几何学、拓扑可能成为支持地理空间数据建模和管理的关键概念；（2）数据流库和对象的“原位”地理计算直接应用于传感器将彻底改变地理信息科学和地理空间数据管理；（3）基于 geoAI 地理空间数据管理研究与应用将获得进一步的发展。

参考文献

[1] Guo, H. Big Earth Data: a new frontier in earth and information sciences [J]. *Big Earth Data*, 2017, 1(1/2): 4–20.
[2] 郭华东. 科学大数据——国家大数据战略的基石[J].中国科学院院刊, 2018, 33(8): 768–773.
[3] Boulton, G. The challenges of a big data earth [J]. *Big Earth Data*, 2018, 2: 1–7.
[4] 王卷乐, 杨雅萍, 诸云强等. “973”计划资源环境领域数据汇交进展与数据分析[J]. 地球科学进展, 2009, 24(8): 947–953.
[5] 司莉, 邢文明. 国外科学数据管理与共享政策调查及对我国的启示[J]. 情报资料工作, 2013(1): 61–66.
[6] 丁培, 国外大学科研数据管理政策研究[J]. 图书馆论坛, 2014(5): 103–110.
[7] 黎建辉, 虞路清. 国际科学数据库现状与发展趋势分析[J]. 科研信息化技术与应用, 2009(1): 6–13.

[8] 周小刚, 罗云峰. 美国国家大气研究中心优先研究领域新特点[J]. 地球科学进展, 2006(7), 751–756.
[9] White, R. M.Geophysical data management—why [J]. *Bulletin of the Amrican Meteorolgical Society*, 1969, 50(3): 143.
[10] NASA Distributed Active Archive Centres [EB/OL]. http: //gcmd.gsfc.nasa.gov/. 2005.
[11] NASA's Global Change Master Directory [EB/OL]. http: //gcmd.gsfc.nasa.gov/. 2007.
[12] The Canadian Earth Observation Network [EB/OL]. http: //www.geoconnections.org. 2005.
[13] 王卷乐, 石蕾, 王玉洁等. 科学数据汇聚的模式分析及对我国的发展建议[J]. 地球科学进展, 2020, 35(8): 839–847.
[14] Repository Finder [EB/OL]. https: //repositoryfinder.datacite.org/. 2020-12-01.
[15] Vantrump,G., Miesch, A. T. United States. geological survey rass-statpac system for management and statistical reduction of geochemical data [J]. *Computers & Geosciences*,1977, 3(3): 475–488.
[16] 李军, 陈崇成. 地球科学数据的元数据研究[J]. 地理研究, 1997(1): 31–38.
[17] 孙九林, 李爽. 地球科学数据共享与数据网格技术[J]. 地球科学, 2002(5): 539–543.
[18] 中国地质调查局发展研究中心数据处理室. 区域地球化学数据管理信息系统(GeoMDIS 2003)升级版问世——勘查地球化学家的实用工具[J]. 地质通报, 2003(7): 547–548.
[19] 杜云艳, 杨晓梅, 王敬贵. 中国海岸带及近海多源数据空间组合和运行的基础研究[J]. 海洋学报, 2003(5): 38–48, 57.
[20] 王卷乐, 游松财, 谢传节. 地学数据共享中的元数据标准结构分析与设计[J]. 地理与地理信息科学, 2005(1): 16–18, 37.
[21] 肖建华, 王厚之, 彭清山等. 地理时空大数据管理与应用云平台建设[J]. 测绘通报, 2016(4): 38–42.
[22] 刘闯, 郭华东, Paul 等. 发展中国家数据出版基础设施与共享政策研究[J]. 全球变化数据学报, 2017, 1(1): 3–11.
[23] 王卷乐, 王祎,卜坤等. 世界数据系统 CoreTrustSeal 数据中心认证实践——以 WDC 可再生资源与环境数据中心为例[J]. 农业大数据学报, 2019, 1(3): 71–81.
[24] 《地球科学大辞典》编委会, 地球科学大辞典/基础学科卷[M]. 北京: 地质出版社, 2006.
[25] Chen, C. M. CiteSpaceIII [DB/OL]. http:// cluster ischool.Drexeledu/cchen/citespace/download/. 2016.
[26] Centre for Science and Technology Studies, Leiden University. VOSviewer Version 1.6.4 [DB/OL]. http: //www.vosviewer.com/. 2016.
[27] 王卷乐, 林海, 冉盈盈等. 面向数据共享的地球系统科学数据分类探讨[J]. 地球科学进展, 2014, 29(2): 265–274.
[28] Frankel, F., Reid, R. Big data: distilling meaning from data [J]. *Nature*, 2008, 455(7209): 30–30.
[29] Ivezic, Z., Lupton, R. H., Schlegel, D. SDSS data management and photometric quality assessment [J]. *Astronmische Nachrichchten*, 2004, 325(6/8): 583–589.
[30] Chen, X. Y., Shao, S., Tian, Z. H. Impacts of air pollution and its spatial spillover effect on public health based on China's big data sample [J]. *Journal of Cleaner Production*, 2017, 142: 915–925.
[31] Steiner, J. L., Sadler, E. J., Chen, J. S. Sustaining the earth's watersheds-agricultural research data system: overview of development and challenges [J]. *Journal of Soil and Water Conservation*, 2008, 63(6): 569–576.
[32] Muller, C. L., Chapman, L., Grimmond, C. S. B. Toward a standardized metadata orotocol for urban meteorological networks [J]. *Bulletin of the American Meteorological Society*, 2013, 94(8): 1161–1185.
[33] Costello, M. J. Distinguishing marine habitat classification concepts for ecological data management [J]. *Marine Ecology Progress Series*, 2009, 397: 253–268.
[34] Amante, M. J, Correia, A. M. R., Wilson, D. Information policy in the EU: legislative framework in Portugal (1989–1992) [J]. *Cadernos BAD*, 1994(2): 9–28.
[35] Lemke, A. A., Wolf, W. A., Hebert-Beirne, J. Public and biobank participant attitudes toward genetic research participation and data sharing [J]. *Public Health Genomics*, 2010, 13(6): 368–377.
[36] Itermon, R., Relyea, H. C. Information Policy [M]. Encyclopedia of Library and Information Science (Volumn 48), Kent, Allen. ed. New York, MaroelDekker, 1991: 176–204.
[37] Shuler, J. A. Citizen-centered government: information policy possibilities of the 108th Congress [J]. *Journal of Academic Librarianship*, 2003, 29(2): 107–110.

[38] Hardwicke, T. E., Mathur, M. B., MacDonald, K. N. G., *et al*. Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal cognition [J]. *Royal Society Open Science*, 2010, 5(8): 180448.

[39] Moen, W. E. Interoperability of information access: Technical standards and policy considerations [J]. *Journal of Academic Librarianship*, 2000(2): 129–132.

[40] Michener, W. K. Ecological data sharing [J]. *Ecological Informatics*, 2015, 29: 33–44.

[41] Nakayama, T. Evidence-based healthcare and health informatics: derivations and extension of epidemiology [J]. *Journalof Epidemiology*, 2006, 16(3): 93–100.

[42] Cech, T. R., Eddy, S. R., Eisenberg, D., *et al*. Sharing publication-related data and materials: responsibilities of authorship in the life sciences [J]. *Plant Physiology*, 2003, 132(3): 19–24.

[43] Parr, C. S., Cummings, M. P. Data sharing in ecology and evolution [J]. *Trends in Ecology and Evolution*, 2005, 20 (7): 362–363

[44] Fienberg, S. E., Martin, M. E., Straf, M. L. Sharing Research Data [M]. Washington, D. C: National Academy Press, 1985.

[45] Constant, D., Kiesler, S., Sproull, L. What's mine is ours, or is it? A study of attitudes about information sharing [J]. *Information Systems Research*, 1994, 5: 400–421.

[46] Matzler, K., Renzl, B., Muller, J., *et al*. Personality traits and knowledge sharing [J]. *Journal of Economic Psychology*, 2008, 29: 301–313.

[47] McCain, K. Mandating sharing: journal policies in the natural sciences [J]. *Science Communication*, 1995, 16: 403–431.

[48] Piwowar, H. A. Who shares? Who doesn't? Factors associated with openly archiving raw research data [J]. *PLoS One*, 2011, 6(7): e18657.

[49] SPARC applauds White House for Landmark Directive Opening up Access to Scientific Research [EB/OL]. http: //www.sparc.arl.org/. 2013-08-28.

[50] Zhu, J., Wright, G., Wang, J., *et al.* A critical review of the integration of geographic information system and building information modelling at the data level [J]. *ISPRS International Journal of Geo-Information*, 2018, 7: 66.

[51] Sacks, R., Ma, L., Yosef, R., *et al*. Semantic enrichment for building information modeling: procedure for compiling inference rules and operators for complex geometry [J]. *Journal of Computing in Civil Engineering*, 2017, 31: 04017062.

[52] Irizarry, J., Karan, E. P., Jalaei, F. Integrating BIM and GIS to improve the visual monitoring of construction supply chain management [J]. *Automation in Construction*, 2013, 31: 241–254.

[53] Amirebrahimi, S., Rajabifard, A., Mendis, P., *et al*. BIM-GIS integration method in support of the assessment and 3D visualisation of flood damage to a building [J]. *Journal of Spatial Science*, 2016, 61: 317–350.

[54] Kang, T. W., Hong, C. H. A study on software architecture for effective BIM/GIS-based facility management data integration [J]. *Automation in Construction*, 2015, 54: 25–38.

[55] Ozel, F. Spatial databases and the analysis of dynamic processes in buildings [C]. In Proceedings of the Fifth Conference on Computer Aided Architectural Design Research in Asia, Singapore, 2000, 2: 97–106.

[56] Bradley, P. E., Paul, N. Using the relational model to capture topological information of spaces [J]. *The Computer Journal*, 2010, 53: 69–89.

[57] Guan, X., van Oosterom, P., Cheng, B. A parallel N-dimensional space-filling curve library and its application in massive point cloud management [J]. *ISPRS International Journal of Geo-Information*, 2018, 7: 327.

[58] VoPham, T., Hart, J. E., Laden, F., *et al*. Emerging trends in geospatial artificial intelligence (geoAI): potential applications for environmental epidemiology [J]. *Environmental Health*, 2018, 17: 40.

[59] Wang, L. G. Reference model for creating information security policy [C]. Proceedings of Information Technology and Environmental System Sciences, 2008, 2: 279–281.

[60] Tang, Y. L., Xu, G. A., Niu, Y. X., *et al*. Information security risk analysis model based on entropy [C]. Proceedings of Information Technology and Environmental System Sciences, 2008, 4: 1146–1150.