·研究论文·

我国开放数据研究进展与趋势(1996—2019 年)

王知津[1]　陈芊颖[1]　韩　峰[1]　吴东颖[2]　李巧英[1]　李圆方[1]
(1.南开大学商学院,天津,300071;　2.吉林省图书馆,长春,130028)

[摘　要]　针对我国开放数据的研究进展和趋势,运用文献计量学方法,以中国知网期刊全文数据库中 1996—2019 年的 1029 篇期刊论文为样本,借助计量可视化软件 CitespaceV,选取发表时间、作者、期刊、机构、关键词作为变量,进行分布特征、关键词共现、关键词聚类以及突现词的分析。研究发现,我国开放数据研究热点体现在政府政务、开放科学、其他应用等方面;研究前沿主要围绕政府政务及其他新兴领域。

[关键词]　开放数据　研究进展　计量分析　开放科学　政府政务

[中图分类号]G203　[文献标识码]A　[文章编号]2095-2171(2020)06-0047-13
DOI:10.13365/j.jirm.2020.06.047

The Progress and Trends of Open Data Research in China(1996-2019)

Wang Zhijin[1]　Chen Qianying[1]　Han Feng[1]　Wu Dongying[2]　Li Qiaoying[1]　Li Yuanfang[1]
(1.Business School,Nankai University, Tianjin 300071;　2.Jilin Province Library, Changchun 130028)

[**Abstract**]　This paper aims at the progress and trends of open data research in China. The 1029 journal papers from 1996 to 2019 in CNKI full-text database are took as a sample using bibliometric methods. The publication time, authors, journals, institutions, and keywords are took as variables. The distribution characteristics, keyword co-occurrence, keyword clustering, and burst terms are analyzed by the Citespace V. It is found that the hotspots of open data research in China are reflected in government affairs, open science, and other applications. And the research fronts are mainly in government affairs and other emerging fields.

[**Keywords**]　Open data　Research progress　Biblometric analysis　Open science　Government affairs

1　引言

随着互联网技术的发展,数据扮演着重要的角色,在数据整合的要求下,开放数据也成为信息时代的研究热点。开放数据是指没有任何版权、专利和其他机制的限制,能够被任何人无障碍、无限制地获取和使用的数据资源。英国开放知识基金会(Open Knowledge Foundation)认为其具有非歧视性、机器可读性、开放授权性[1]三个要素,对数据提供者提出了严格的要求;开放数据中心联盟(Open

[作者简介]　王知津,教授,博士生导师,研究方向为竞争情报与竞争战略、信息管理与信息系统;陈芊颖(通讯作者),硕士研究生,研究方向为信息管理与信息系统;韩峰,硕士研究生,研究方向为数据管护;吴东颖,助理馆员,研究方向为图书情报与数字图书馆、企业经济;李巧英,硕士研究生,研究方向为图书情报与数字图书馆、文献计量;李圆方,硕士研究生,研究方向为竞争情报与竞争战略、文献计量。

本文引用格式:王知津,陈芊颖,韩峰,等.我国开放数据研究进展与趋势(1996—2019 年)[J].信息资源管理学报,2020,10(6):47-59.

Data Center Alliance)视之为IT基础设施、云计算的应用模式和解决方案[2]。

开放数据包括开放政府数据、开放科学数据(或研究数据、科研数据)、开放机构数据(如开放企业数据)、开放个人数据等[3]。根据开放数据发达的国家的经验,开放数据在政府、互联网、图书馆等领域的信息组织中的实践已证明了其广阔的应用空间[3]。开放数据在开放、自由、共享的主题下,为信息互通、科学研究、学术交流等提供了重要支持。

我国开放数据研究起步稍晚,与发达国家相比,在政府数据开放水平、信息法律制度、组织管理体系以及技术架构等方面还存在一定的差距和不足[4]。我国政府数据开放实践仍存在一些缺陷,比如,数据量缺乏、数据主题分类不明确、支撑技术欠缺、制度保障不完善[5]等。

然而,我国政府对于开放数据的重视程度正在逐年加大,并积极开展大数据资源平台建设。2019年上海社会科学院信息研究所发布的《全球重要城市开放数据指数》报告中显示,在27个全球重要城市中,贵州贵阳公共数据可持续开放综合指数名列第四,仅次于首尔、纽约、芝加哥[6]。在开放数据政策研究方面,虽然我国不断完善开放数据的政策体系,但是,仍然缺少专门独立的数据开放法规,对数据开放的政策规定分散在多种不同类的政策中,没有形成系统化的开放数据政策体系[7]。许多地方政府基本处于各自独立的开放数据试验阶段,开放数据集的数量、质量和利用水平较低[8]。

本研究以中国知网全文期刊数据库为数据来源,以开放数据和数据开放研究的中文期刊文献作为研究对象,运用文献计量学的方法,借助Citespace软件(5.5.R2版本),对我国开放数据的相关文献进行可视化分析和讨论,旨在探索我国关于开放数据的研究进展,把握未来研究的发展趋势,为后续研究提供参考。

2 研究方法和数据来源

2.1 数据源与检索

本研究选取CNKI中国知识资源总库收录的主题为"开放数据"和"数据开放"的中文期刊论文作为文献计量的分析对象。检索策略为:以主题(SU)为检索字段,检索式为SU=('数据开放'+'开放数据'),数据库选择为期刊数据库,检索时间为"不限-不限",检索为精确检索,利用检索式可有效排除英文文献,也可筛除仅与"开放"和"数据"相关的论文。截至2019年12月31日,共检索出期刊论文2113篇。

2.2 数据清洗

由于主题词涉及开放数据和数据开放的文章种类繁多,得到的检索结果中仍存在大量与本研究所需文献无关的文章。为了保证分析结果的有效性和准确性,通过人工逐篇浏览的方式,对文献的题目、关键词与摘要进行综合分析考虑,最终筛除无效文章1084篇,具体类别如下:

(1)题目、关键词和摘要中均未出现数据开放或开放数据,主要内容并不是所需的开放数据或数据开放主题,此类文章共计471篇。

(2)题目、关键词或摘要中虽含有数据开放或开放数据,但只是一笔带过地提及到,文章主旨并不针对数据开放或开放数据,而主要围绕其他不相关的课题展开,与本文要求的主题不符,此类文章共计393篇。

(3)题目、关键词或摘要中含有数据开放或开放数据,但并不属于研究性论文,如选题计划、项目介绍、会议新闻、评论等内容。虽然其内容是围绕开放数据的语意,但对本文的研究主题无参考意义,此类文章共计220篇。

根据以上清洗标准,最终筛查出有效文章共计1029篇,它们与本文的研究主题相关。在这些文章中,最早发表时间为1996年,最晚发表时间为2019年,故本文将对1996—2019年发表的1029篇文章进行计量分析。

3 分布特征分析

3.1 时间分布

通过期刊论文发表的时间分布分析,可以看出研究主题的时间演变历程,包括研究进展、受关注程度以及未来发展趋势等。将筛查后的1029篇文献按照发表年份排列,绘制出柱状图,可以直观地反映出1996—2019年我国开放数据研究的动态变化,如图1所示。

如图1所示,1996—2019年间,CNKI中国

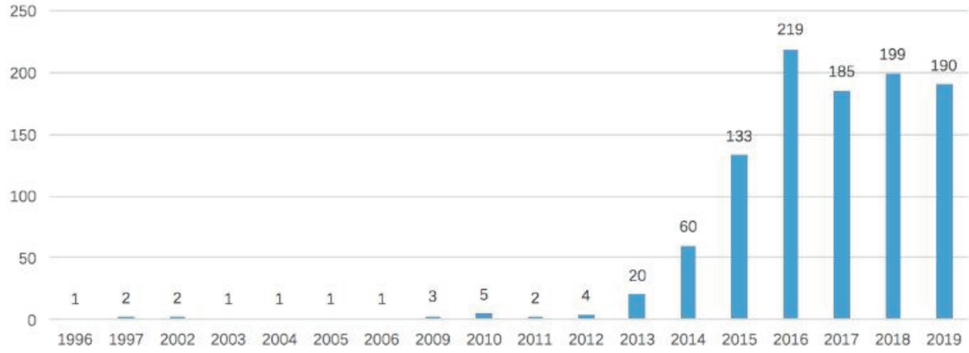

图 1　我国开放数据研究论文的时间分布

知识资源总库中收录的我国关于开放数据的论文数量总体呈上升趋势。自 2014 年后，每年的发文量均在 50 篇以上，说明我国对于开放数据领域关注度持续增长。根据发文量与发文时间的走势分布，可将我国开放数据研究分为三个阶段：

（1）酝酿萌芽期（1996—2012 年）。此阶段的特点为发文量少、波动小，每年发文量均未超过 5 篇。在这一阶段，开放数据最初只在一些发达国家被提出，直至 2009 年美国总统奥巴马签署《开放透明政府备忘录》，同年《开放政府令》获批准，从此，在全球范围内开放数据运动的浪潮迅速兴起[9]。我国对于开放数据的研究起步稍晚，即便有零星关注，也只是对于先行国家的介绍和借鉴。在这一阶段的后期，即 2012 年 6 月，我国开通了国内首个政府数据服务网[10]——上海市政府数据服务网。总体上说，2012 年前我国开放数据处于引进和酝酿阶段。

（2）快速增长期（2013—2016 年）。在这一阶段，开放数据受到我国越来越多学者的关注，发文量与时间成正比例关系增长，甚至翻倍递增，并在 2016 年达到该阶段的峰值。自 2012 年以后，北京、浙江、重庆、武汉、佛山、青岛等多地在上海之后推出开放数据门户网站。2014 年，我国加大了对开放数据的建设力度，成立了数据中心联盟（DCA）、开放数据中心委员会（ODCC）。2015 年首次提出“互联网＋”，并于当年国务院印发《促进大数据发展行动纲要》（国发〔2015〕50 号）。2016 年，《贵州省大数据发展应用促进条例》颁布。政府的一系列举措以及社会的关注共同推动了中国数据开放的进程，相应的研究成果也如雨后春笋般涌现出来。

（3）平稳发展期（2017 年至今）。此阶段的热度相较于 2016 年有小幅回落，但发文量基本稳定在高数量区域内。根据《2019 中国地方政府数据开放报告》中显示，截至 2019 年 4 月，我国已有 82 个地方政府推出政府数据开放平台，其中省级地方政府 13 个、副省级与地市级地方政府 69 个[11]。因此，在这一阶段，数据开放平台已经逐渐成为地方政府数字化建设的一个“标配”。按照《促进大数据发展行动纲要》中所要求的 2018 年底前建成国家政府数据统一开放平台，目前已经开放的数据领域包括教育科技、民生服务、道路交通、健康卫生、资源环境、文化休闲、机构团体、公共安全、经济发展、农业农村、社会保障、劳动就业、企业服务、城市建设、地图服务等。所以说，基于第二阶段热度骤升的基础，本阶段的研究趋于稳定和成熟。

3.2　作者分布

1029 篇文章涉及作者共计 1840 位，发文量在 6 篇及以上的作者共计 14 位，如表 1 所示。

表 1　我国发表 6 篇及以上开放数据研究论文的作者

序号	作者	发文量	序号	作者	发文量	序号	作者	发文量
1	马海群	33	6	周文泓	12	11	林岩	6
2	黄如花	29	7	夏义堃	12	12	王春迎	6
3	郑磊	20	8	温芳芳	9	13	赵龙文	6
4	陈美	18	9	迪莉娅	7	14	陈朝兵	6
5	翟军	14	10	周志峰	6			

由表 1 可知，发文量居于首位的是马海群，共发文 33 篇，主要研究方向是开放数据与数据安全政策、信息政策与法律、知识产权与信息管理、信息咨询等方面；其次是黄如花，共发文 29 篇，研究方向为信息组织与检索、信息服务、数字信息资源开放存取、信息素养教育；第三位是郑磊，共发文 20 篇，主要研究方向是数字治理与电子政务、政府数据开放、政府数据共享、政府数据治理、在线与移动公共服务、政府社会化媒体应用、电子政府准备度与发展水平评估。

根据发文作者共现图谱可以识别出同一个研究领域的核心作者和作者之间的合作关系。利用 Citespace 软件对这 1840 位作者进行共现分析，得到结果如图 2 所示。作者之间的连线数为 119 个，节点数为 180 个，网络密度为 0.0074。图中的节点越大，表明参与发文的次数越多。

通过综合分析发现，以马海群、黄如花、郑磊、翟军等为中心形成的作者群研究成果居多，但也依然存在很多高产的个人作者，如陈美等，作者分布总体呈聚焦点明显但群体分散的特点。

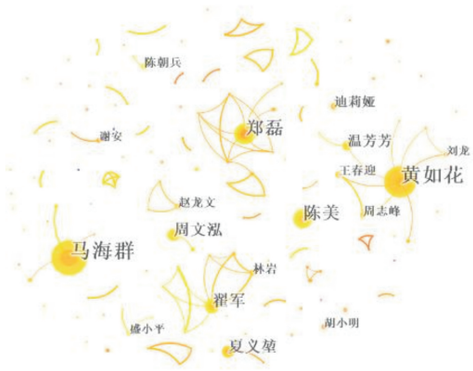

图 2　我国开放数据研究作者共现图谱

3.3　期刊分布

对于某一研究主题的文章，期刊所发表的文章数量越多，则表示该期刊对于该主题的影响越大，而刊载相关主题的期刊所在的领域对该主题的关注度也越大。

经统计，我国开放数据主题研究载文量 10 篇及以上的期刊如表 2 所示。

表 2　我国载文 10 篇及以上开放数据研究论文的期刊

期刊名	载文量	占比	所属领域	期刊名	载文量	占比	所属领域
电子政务	69	6.71%	电子政务	图书与情报	17	1.65%	图书情报
图书情报工作	50	4.86%	图书情报	图书馆理论与实践	16	1.56%	图书馆学
情报理论与实践	41	3.99%	情报学	情报资料工作	16	1.56%	情报学
情报杂志	40	3.89%	情报学	大数据	12	1.17%	大数据
图书馆	33	3.21%	图书馆学	中国行政管理	11	1.07%	政府管理
图书馆学研究	28	2.72%	图书馆学	计算机与网络	11	1.07%	电子科技技术
现代情报	25	2.43%	情报学	图书情报知识	10	0.97%	图书情报
情报科学	19	1.85%	情报学	图书馆杂志	10	0.97%	图书馆学
数字图书馆论坛	18	1.75%	图书馆学				

如表 2 所示，载文量 10 篇及以上的期刊共有 17 种，载文量共计 429 篇，占样本数量的 42.57%。这些期刊的研究领域多为图书情报、情报学、图书馆学、政府政务，还涉及到大数据、电子科学技术等。从整体上来看，期刊的研究领域较为集中，关注领域明显，有些领域之间存在交叉现象。

通过分析载文量较高的期刊类别发现，不同领域的期刊对我国开放数据研究的角度有所不同。图书情报类期刊主要从图书情报的综合角度，对比分析国内外开放数据的理论与实践，基于开放数据的国际化标准而展开研究；图书馆学类期刊主要从图书馆的角度出发，重点研究开放数据的实际应用，如信息服务、数据平台等；情报学类期刊主要围绕开放数据管理问题而展开，如数据质量、元数据标准等；政府政务类期刊主要站在政府的立场，围绕开放数据在政务上的应用、战略规划、开放平台建设、政策制定等而展开。此外，大数据类期刊多数将大数据与数据开放结合起来进行探讨，电子科学技术类期刊侧重于开放数据的应用技术问题。

3.4 机构分布

1996—2019 年之间，共有 26 家机构发表开放数据研究论文超过 5 篇，经整理如表 3 所示。发文量最多的为武汉大学信息管理学院 68 篇、武汉大学信息资源研究中心 39 篇、黑龙江大学信息管理学院 37 篇、黑龙江大学信息资源管理研究中心 30 篇。

表 3 我国发文 5 篇及以上开放数据研究论文的机构

序号	机构	发文量	序号	机构	发文量
1	武汉大学信息管理学院	68	14	南京大学信息管理学院	9
2	武汉大学信息资源研究中心	39	15	中山大学资讯管理学院	9
3	黑龙江大学信息管理学院	37	16	大连海事大学航运经济与管理学院	8
4	黑龙江大学信息资源管理研究中心	30	17	华南理工大学经济与贸易学院	8
5	湖北工业大学经济与管理学院	18	18	北京师范大学政府管理学院	8
6	四川大学公共管理学院	17	19	上海大学图书情报档案系	8
7	中国科学院文献情报中心	17	20	西南财经大学公共管理学院	7
8	中国人民大学信息资源管理学院	13	21	大连海事大学交通运输管理学院	7
9	复旦大学国际关系与公共事务学院	11	22	华南师范大学经济与管理学院	7
10	中国科学院大学	11	23	北京大学信息管理系	7
11	复旦大学国际关系与公共事务学院数字与移动治理实验室	10	24	上海理工大学管理学院	7
12	华中师范大学信息管理学院	10	25	武汉大学法学院	6
13	安徽大学管理学院	9	26	中国科学院国家科学图书馆	6

结合各机构发文时间，绘制发文机构时间线视图，如图 3 所示。时间线视图主要侧重于勾画聚类的起始时间和历史跨度。将时间切片设置为三年，选出发文量较多的机构。图中圆圈对应的时间代表该机构首次发表论文的年份，在后来的年份中，该机构发表的论文越多，则圆圈越大。由图 3 可见，从开放数据研究机构来看，武汉大学信息管理学院的聚类时间最早(2008 年)，发表论文最多，并且时间跨度最大，而武汉大学信息资源研究中心次之，这表明，在开放数据研究方面，武汉大学走在全国的前列。

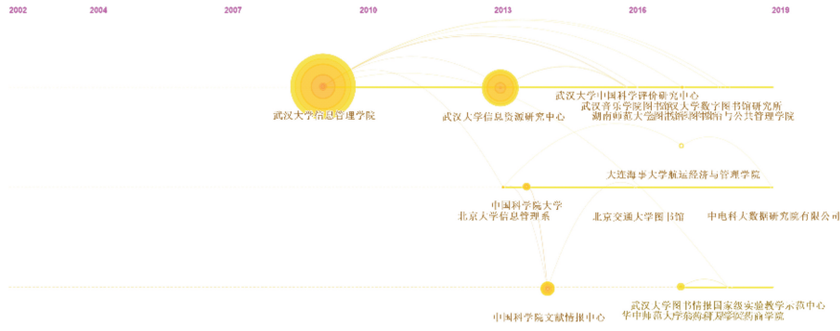

图 3 我国开放数据研究机构的时间线视图

4 关键词共现分析

在学术期刊的全部内容中，关键词反映的是全文的主旨内容。作为文章核心内容的集中概括，关键词的分布和出现频次可以有效地反映出该领域的主题分布特点。在文献计量学中，利用关键词共现的算法，形成一个虚拟的关键词网络，可以明确该文献集所代表的学科各主体之间的关系，对于研究某个主题的成熟度、知识结构和研究规模[12]具有重要意义。在虚拟的关键词网络中，关键词是其中的一个一个节点，关键词的共现则体现为节点之间的直接连线。两个关键词在同一篇文章中出现的频次越多，则表明这两个关键词之间的紧密度越高。

本文利用 Citespace 软件，对涉及我国开放数据的 1029 篇论文中的关键词进行共现分析，形成节点 200 个、连线 877 条的关键词共现图谱，如图 4 所示。

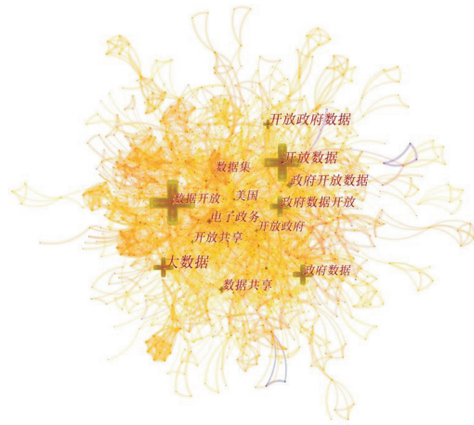

图 4　我国开放数据研究论文关键词共现图谱

本文绘制的关键词共现图谱的时间段为 1996—2019 年，时间切片设置为 1 年。根据被引频次 C(citation)、两篇文献的共被引频次 CC(cocitation)和共被引系数 CCV(cosinecoefficient)三个层次设置阈值(C,CC,CCV)，为得到更加科学的共现图谱，经过反复试算和比较，得出阈值设置结果分别为(1,1,5)、(1,1,5)、(2,2,20)。按照整个时间跨度的前、中、后三个时间分区分别设定，其余时间分区的阈值利用线性插值算法来确定。为了评判绘图效果，Citespace 提供了模块值(Modularity，简称 Q 值)和平均轮廓值(Mean Silhouette，简称 S 值)两个指标，分别判断网络结构是否显著和聚类的清晰度。一般而言，Q 值在区间[0,1]内，Q≥0.3 则说明划分出来的聚类结构显著；Q 值在 0.4～0.8 时，聚类效果比较好。S 值在－1～1 之间，若在 0.5 以上，聚类一般认为是合理的；当 S 值＝0.7 时，聚类是高度令人信服的。在本文的共现图谱中，Q 值＝0.6876，S 值＝0.7219，说明该图谱聚类结构显著，且可信度较高。

共现图谱中的节点表示关键词，节点越大，关键词出现的频次越高，表明研究者对该关键词的关注度越高。由图 4 可见，出现频次较高的关键词除了开放数据和数据开放外，还有大数据、政府数据开放、政府数据、数据共享等。将出现 15 次以上的高频关键词按照频次由高到低排序，得到高频关键词列表(表 4)。结果显示，除了排在前三位的数据开放、开放数据和大数据都比较宽泛外，排在第 4—7 位的都与政府数据开放有关，此外还有开放政府、电子政务、政府信息公开、地方政府和政府等高频关键词。由此可见，目前我国对于开放数据的研究主要集中在政府数据开放上。

表 4　我国开放数据研究论文高频关键词

序号	关键词	频次	序号	关键词	频次	序号	关键词	频次
1	数据开放	294	12	科学数据	32	23	政策	18
2	开放数据	252	13	美国	31	24	开放获取	18
3	大数据	141	14	信息公开	31	25	关联数据	18
4	政府数据	140	15	政府信息公开	25	26	数据安全	17
5	政府数据开放	138	16	地方政府	24	27	数据资源	17
6	开放政府数据	71	17	智慧城市	23	28	数据管理	17
7	政府开放数据	50	18	英国	23	29	政府	17
8	数据共享	45	19	开放平台	21	30	公共服务	16
9	开放政府	44	20	数据集	20	31	隐私保护	16
10	电子政务	43	21	图书馆	19	32	元数据	15
11	开放共享	36	22	数据开放平台	19	33	大数据产业	15

中心度是网络分析中度量节点重要性和权威性的指标。在关键词共现图谱中，中心度的值越高，表明该关键词越重要。由于本文关键词的基数较大，中心度的值精确到小数点后两位，许多关键词的中心度小于 0.01，在结果中约等于 0，但不代表实际中心度为 0。整理 1996—2019 年中心度较高的关键词，得到表 5。

表 5　我国开放数据研究论文高频关键词中心度

年份	中心度	关键词	年份	中心度	关键词	年份	中心度	关键词
2002	0.04	科学数据	2013	0.07	政府数据	2014	0.03	英国
	0.07	美国		0.08	政府数据开放		0.05	开放平台
2003	0.02	元数据		0.08	开放政府		0.06	数据集
2009	0.03	政策		0.05	政府信息公开		0.03	数据安全
	0.03	开放获取		0.04	智慧城市		0.03	数据资源
2010	0.07	电子政务		0.05	图书馆		0.04	政府
	0.03	关联数据		0.02	公共服务		0.04	大数据产业
2011	0.15	开放政府数据	2014	0.12	政府开放数据	2015	0.01	数据开放平台
	0.06	信息公开		0.1	数据共享	2016	0.02	地方政府
	0.02	数据管理		0.07	开放共享		0.02	隐私保护

由表 5 可知，关键词共现主要分三个阶段：

(1)初步发展阶段(1996—2003 年)。这一阶段，由于开放数据研究在我国刚刚兴起，因此只有三个关键词呈现出明显的中心度，分别为美国、科学数据和元数据。其中，美国的中心度最大。结合图 4 可见，该关键词的节点也相对较大，这表明在这一时期，我国对于美国开放数据的研究热度很高。这是由于美国开放数据运动开始的较早，处于国际领先地位，我国在开放数据初步发展阶段的研究与实践大多参照美国经验。

(2)快速发展阶段(2009—2014 年)。开放数据研究得到迅速发展。该阶段出现的关键词较为丰富且多样。主要关键词围绕政府政务、开放数据的应用、数据管理等。例如，中心度较高的电子政务、开放政府数据、政府数据等都是围绕政府主题展开研究，同期开放数据的概念引入中国，政府部门早期起到了推动性的积极作用。另外，智慧城市、图书馆等多围绕开放数据在实践中的应用，数据安全、数据集等大多围绕数据管理而展开。

(3)平稳发展阶段(2015—2016 年)。开放数据研究的热度逐渐趋于平静，2015 年的关键词为数据开放平台，中心度为 0.01，并且在快速发展阶段也出现了相似的关键词，说明开放数据平台的研究仍受到专家学者的关注；2016 年出现的关键词为地方政府和隐私保护，中心度均为 0.02，说明开放数据研究的热点在地方政府领域，同时隐私保护也受到了重视。

5　关键词聚类分析

在关键词聚类中，两个主题在同一篇文献中出现的次数越多，则两个关键词之间的距离越近，按照高频词之间的距离远近划分为不同的研究子领域，形成了一个一个的类团。利用 Citespace 的关键词聚类功能，绘制我国开放数据关键词聚类知识图谱，如图 5 所示。

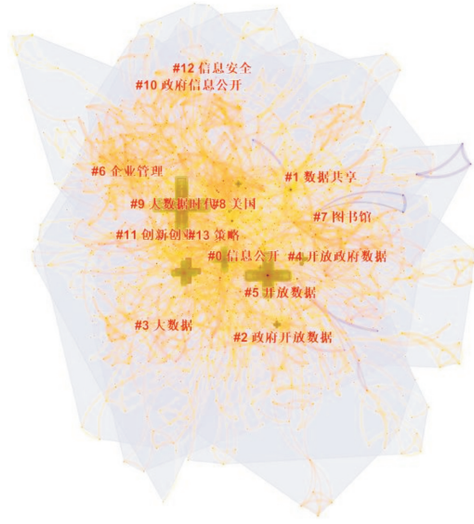

图 5　我国开放数据研究论文关键词聚类知识图谱

由图 5 可见，1996—2019 年自动形成了许多聚类，从＃0 开始代表节点数最多的聚类，随着聚类编号的增加，节点数变少。根据 Citespace 生成的聚类知识图谱导出各个聚类的详细信息，筛选出节点数在 50 以上的聚类＃0 至＃14 共 15 个聚类，并列出聚类内的高频关键词，形成表 6。

从表 6 可以看出，聚类的出现时间点主要在 2014 年、2015 年、2016 年三个年份，轮廓值越接近于 1，说明独立性越强。将 15 个聚类按照研究领域划分，可分为 5 个领域，即政府政务、开放科学、大数据、政策推行、其他应用。

表 6　我国开放数据研究论文关键词聚类

聚类号	标签词	节点数	轮廓值	出现年份	聚类内关键词
0	信息公开	143	0.793	2015	电子政务、互联网＋、政府治理、政府网站、信息安全、公共管理、公共安全图像信息、在线服务
1	数据共享	122	0.849	2016	司法统计、司法大数据、数据堡垒、科学数据共享、科学数据管理、科学数据政策
2	政府开放数据	110	0.838	2016	层次分析法、评价体系、数据质量控制、模糊综合法、政府技术能力、政府开放数据价值
3	大数据	107	0.762	2015	开放平台、教育数据、教育信息化2.0、实施路径、数据中心、数据分析、开放政策
4	开放政府数据	105	0.833	2016	数据质量、元数据、元数据标准、关联开放数据
5	开放数据	103	0.822	2014	伦敦、智慧城市、创作共用、智慧交通、思维形式、智慧档案馆、个人信息保护
6	企业管理	87	0.886	2015	企业、开放共享、资产化、开放数据、制度设计、决策、产业化、服务网
7	图书馆	80	0.897	2014	数字人文、开放数据服务、服务创新、信息生态链、高校图书馆、数据素养教育
8	美国	73	0.873	2014	政务大数据、网站评价、共享平台、科学数据、data.gov、效率评价、开放型政府
9	大数据时代	70	0.865	2015	云计算、腾讯公司、电子政务建设、创新、政府公信力、电子政务云
10	政府信息公开	56	0.886	2015	政务、准备度、经济、政治逻辑、综合档案馆、交通运输服务、评估指标体系、地理空间
11	创新创业	56	0.864	2016	经济价值、众创空间、生态系统、开发利用、政务数据开放、政策工具、开放数据、政府数据开放平台

(1)政府政务:聚类＃2 政府开放数据、＃4 开放政府数据、＃10 政府信息公开。这三个聚类都涉及政府政务,有些聚类中包含的关键词含义有重叠,说明这些聚类的研究内容有所交叉,但各聚类的研究重点和方向不同。

聚类＃2 政府开放数据的轮廓值为 0.838,共有关键词 110 个;聚类＃4 开放政府数据的轮廓值为 0.833,共有关键词 105 个。这两个聚类的出现时间均为 2016 年,即我国开放数据快速发展阶段刚刚过去,迎来了平稳发展阶段。政府开放数据是近年来电子政务和信息管理领域的研究热点。随着政府信息化水平不断提升,政府部门生成、收集、保存了大量与公众生产生活息息相关的数据。开放政府数据是实现大数据战略的重要前提。政府向社会开放数据,供社会利用,将创造巨大的公共价值,推动经济增长和社会发展,提升国家整体竞争力[13]。

聚类＃2 政府开放数据包括的主要关键词有层次分析法、评价体系、数据质量控制、模糊综合法、政府技术能力、政府开放数据价值等。在前期较为坚实的研究基础上,开始转向利用一系列评价方法和评价体系,对政府开放数据的政策、平台、成熟度、质量、技术能力、价值等进行评价分析,这是我国开放数据较为成熟的标志。

聚类＃4 开放政府数据包括的主要关键词有数据质量、元数据、元数据标准、关联开放数据等。数据质量是政府开放数据的基本保证,受到关注是必然的;元数据标准是政府开放数据平台的重要基础,用户发现、理解和使用开放数据离不开高质量元数据的支持[14],提供翔实的元数据,有利于用户根据自身需求对数据进行开发利用,从而提高数据的适用性[13];关联数据是政府开放数据中的关键技术,政府采用关联数据的标准发布数据,可以提高政府数据的透明度,提升政府数据的利用率[15]。对政府开放数据的评价离不开对开放政府数据本身的研究,两者相互依赖、相辅相成,共同完善政府开放数据。

聚类＃10 政府信息公开,轮廓值为 0.886,2015 年出现,共有关键词 56 个。政府

信息公开与政府数据开放的概念相关但内涵不同。政府信息公开主要停留在政府法规、流程、权力等方面，更多的是规章、制度等信息层面。而真正意义上的政府数据开放主要是指原始数据的开放。政府信息公开构建了开放政府这一理念，而政府开放数据在此基础上通过原始数据的开放让政府更加透明[16]。在开放政府的推进过程中，对数据“开放”的特定要求逐渐演变出新的制度体系，这种体系结合政府信息公开，完善了政府数据开放的结构与内容。

根据该聚类的出现时间和关键词等，不难发现，目前对于该领域的数据开放研究大多集中在通过科学而前沿的技术方法，加强政府数据质量、平台、政策等建设，力求构建一个高质高效的政府开放数据平台。

(2)开放科学：聚类＃0 信息公开、＃1 数据共享、＃5 开放数据。

聚类＃0 信息公开，轮廓值为 0.793，2015 年出现，共有关键词 143 个，主要包括电子政务、互联网＋、政府治理、政府网站、信息安全、公共管理、公共安全图像信息、在线服务等，关键词呈现出信息公开的多元化发展趋势。与聚类＃10 政府信息公开不同的是，除了政府的信息公开以外，还包括互联网＋传统行业的信息公开、公共管理、信息开放的安全问题等，信息公开的范围进一步扩大。在信息公开与开放数据中，公开质量、开放程度、数据利用、隐私安全等问题目前尚不成熟，需要进一步改善和提升，有可能成为今后的研究课题。

聚类＃1 数据共享，轮廓值为 0.849，2016 年出现，共有关键词 122 个，主要包括司法统计、司法大数据、数据堡垒、科学数据共享、科学数据管理、科学数据政策等，围绕司法数据和科学数据展开。数据共享机制可以实现司法数据的自由流动，有利于司法信息化、司法现代化、司法科学化[17]的形成，为司法现代化提供良好的数据和技术支持。科学数据共享是数据共享的重要方面，包括科学数据的管理、政策与实践等。科学数据开放共享的主要目标是研究加速、支持新发现、促进合作、提升研究责任、提高研究效率与创新能力[18]。科学数据共享对于完善科学技术制度措施、促进科研成果开放获取、为社会创造新的知识和科研价值，将会发挥重要作用。

聚类＃5 开放数据，轮廓值为 0.822，2014 年出现，共有关键词 103 个，主要包括伦敦、智慧城市、创作共用、智慧交通、思维形式、智慧档案馆、个人信息保护等。其中的高频词与聚类＃2 政府开放数据、＃6 开放政府数据的研究热点并无明显重合，说明该聚类更多地关注于政府开放数据以外的内容，比如智慧城市、智慧交通、智慧档案馆等，而智慧城市又通常与智慧交通、智慧档案馆等领域交叉。这些领域相当于一个一个的海量数据库，研究和分析其中数据的可用性、质量、成本、管理等诸多问题，并共同利用信息化技术和创新性思维的概念，能够实现优化服务的目的，从而打破城市发展、交通、档案馆等实体的局限性，契合科技时代的发展需求。

就开放科学而言，近年来的研究热点主要是基于实务与应用产生的开放数据研究，鼓励非科学专业人士参与问题探索、数据收集与分析。可以预测，随着开放数据的概念深入人心，信息技术日渐成熟，开放数据越来越基于公众需求而展开，在实践应用上的研究将更加深入。

(3)大数据：聚类＃3 大数据、＃9 大数据时代。

聚类＃3 大数据和聚类＃9 大数据时代出现的时间均为 2015 年，即我国开放数据研究刚刚进入平稳发展阶段。进入大数据时代后，开放数据显得越来越重要，作为大数据的重要部分，已经成为一种新的资源。

聚类＃3 大数据的轮廓值为 0.762，主要包括开放平台、教育数据、教育信息化 2.0、开放政策、实施路径等关键词，这说明教育大数据成为我国开放数据的研究热点。在教育领域，将不涉及个人隐私、国家安全机密的教育数据按照规范化的方式公开，能够实现教育数据的创新应用与价值增值，推动教育事业高质量发展[19]。2014 年，教育部印发《教育信息化 2.0 行动计划》，力求构建“互联网＋教育的大平台”。

聚类＃9 大数据时代的轮廓值为 0.865，

主要包含电子政务建设、政府公信力、电子政务云等关键词，这说明电子政务和电子政府不仅是聚类＃2 政府开放数据和＃4 开放政府数据的重要研究课题，而且也成为大数据和大数据时代开放数据的研究热点。毫无疑问，大数据时代电子政务中的开放数据研究对于提升政府数据公开化水平，构建更先进的政府管理和服务模式，高效率发挥政府职能，具有重要意义。

除了教育和电子政务外，还出现了一些与健康医疗有关的关键词。我国的医疗开放数据主要围绕医疗费用和医疗质量等方面展开[20]。健康医疗大数据将为临床诊疗、药物研发、卫生监测、公众健康、医药卫生政策制定和制度执行等带来创造性变化，全面提升健康医疗领域的治理能力和水平，创造极大的价值[21]。

(4)政策推行：聚类＃8 美国

美国、英国、澳大利亚、加拿大等国家的开放数据运动早于中国，在政策和实践中具有很大的参考价值。在我国最初的开放数据研究中，大多侧重于这些国家的法律、政策介绍和评述，因此出现了聚类＃8 美国。

聚类＃8 美国，轮廓值为 0.873，2014 年出现，共有关键词 73 个，主要包括政务大数据、网站评价、共享平台、科学数据、data.gov、效率评价、开放型政府等。美国是世界上最大的信息产业国，在信息公开、信息获取、信息开发、信息隐私保护、信息安全和信息产权保护等方面积累了丰富的经验[22]。与美国的全方位、多层次、多形式[23]的数据开放共享合作模式相比，我国还存在一定的差距。我国对于开放数据的理论和实践研究上起步较晚，在数据开放制度上还没有探索出自己的道路，开放数据的运动尚未真正形成，因此，在许多政策法规和实际应用中需要借鉴美国等国家的经验。自 2014 年开始，我国开放数据的发文量成倍增长，主要集中于法律法规、先进策略和具体实践。对于美国等开放数据的持续关注，将为我国的立法和政策制定提供参考和借鉴。

(5)其他应用：聚类＃6 企业管理、＃7 图书馆、＃11 创新创业。

聚类＃6 企业管理，轮廓值为 0.886，2015 年出现，共有关键词 87 个，主要包括企业、开放共享、资产化、开放数据、制度设计、决策、产业化、服务网等。这说明企业数据的资产化、产业化、制度设计、决策方向成为开放数据在企业中的研究热点。在企业生产经营活动中所产生的有经济价值的数据归属权属于企业本身，通过开放数据可以为企业带来巨大的潜在效益，创造新的产品和服务。重视开放数据对于经济增长的推动作用，把开放数据提高到企业创新主体来认识，将成为今后的研究课题。

聚类＃7 图书馆，轮廓值为 0.897，2014 年出现，共有关键词 80 个，主要包括数字人文、开放数据服务、服务创新、信息生态链、高校图书馆、数据素养教育等。图书馆是数据基础设施的构建者，开放数据的倡导者[24]。随着网络技术的发展，将信息生态链数据、数字人文、数据素养教育等概念引入图书馆，实现服务创新是近年来的研究热点之一。同时，开放数据环境也改变了高校图书馆的学术交流速度与模式[25]。

聚类＃11 创新创业，轮廓值为 0.864，2016 年出现，共有关键词 56 个，主要包括经济价值、众创空间、生态系统、开发利用、政务数据开放、政策工具、开放数据、政府数据开放平台等。开放数据的创新创业研究热点表现为各领域的多元化发展，涉及到数据开放环境下各领域的创新创业服务途径、对于生态系统的思考、政策建议等。在大数据时代，数据的价值需要得到了挖掘，数据的开放共享为创新创业提供了高质量的信息资源，使数据具有更高的经济价值。政府作为数据的产生者和拥有者，推动数据开放共享与开发利用，既有利于提升公共服务水平，也能挖掘数据的价值[26]。

通过关键词聚类分析可以发现我国开放数据的研究前沿和未来趋势，首先，持续重点关注政府政务数据开放，包括数据质量及评价与控制、元数据标准、关联开放数据、政府治理、公共管理与安全；其次，开放科学数据逐渐受到重视，包括科学数据共享、科学数据管理、科学数据政策；第三，开放数据应用领域的拓展，包括司法、教育、交通、企业、图书馆数字人文、创新创业以及数据素养。

6 突现关键词分析

通过突现词可以探究我国开放数据的研究前沿与未来趋势。根据 Citespace 软件的算法将突现词出现的时间进行排列得到图 6。

图 6 我国开放数据研究论文的突现关键词

图 6 中 Keywords 表示突现强度最高的关键词；Year 表示样本数据的起始年；Strength 表示突现词的强度，数值越大表明可信度越高，同时热点度也越高；Begin 表示突现词最早出现的年份；End 表示突现词最晚出现的年份。而“1996－2019”表示时间轴，每个节点表示一年时间，红色节点表示突现词持续的时间。经过分析发现，随着时间发展，突现词不尽相同。

(1)2009 年及以前。未出现突现词，根据上文的时间分布特征和关键词聚类可知，2009 年前，每年关于开放数据的发文量均未超过 5 篇，这是由于 2009 年以前数据开放运动刚刚开始兴起，数据开放还未达到萌芽阶段，极少受到国内各领域的关注。

(2)2010—2015。受到世界各地开放数据运动的影响，我国开放数据进入增长期阶段，这一阶段的期刊发文量逐年递增，出现的突现词有电子政务、关联数据、信息公开、政务、企业管理。

2010 年出现的突现词是电子政务、关联数据，均与电子信息化有关。开放数据的兴起促进了电子政务的发展，电子政务成为突现词；关联数据成为突现词是因为将数据关联起来有利于开放数据之间的相互参考、借鉴创新和开放利用。

2011 年出现的突现词是信息公开。信息公开与数据开放不同，信息公开主要是政府部门对公民知情权的保障，更多的是停留在政府法规、流程、权力等方面，即规章、制度等信息层面。数据开放是在知情的基础上，让公众获得和利用数据。真正意义上的数据开放主要是指原始数据的开放。数据开放继承了开放政府的理念，拓展了开放政府的内涵，是向大数据时代的自然延伸。二者的关系为继承但不取代[17]，不过，它们都是由政府作为中间纽带。

2014 年出现的突现词是政务。数据开放能够为政务公开、业务协同、辅助决策、公共服务等提供信息支持。2014 年我国在电子政务方面的数据开放实践广度和深度都有所提高，虽然还未建立国家级政府数据开放门户网站，但是北京、上海、青岛等地方政府都开始着手规划建设并完善各自的政务门户网站。

2015 年出现的突现词是企业管理。这说明我国对于开放数据的研究开始不仅仅专注于政府政务，也将开放数据应用到商业活动中。基于开放数据所形成的产业是信息资源产业的一个新领域，在对数据内容进行采集、加工、传播、利用等过程中产生数据产品或数据服务。以开放数据为基础进行的商业活动大致有这样几类：数据整理、数据分析、数据转换、数据应用软件开发、数据平台等。开放数据为新商业生态圈的形成起到了促进作用。

(3)2016－2018 年。该阶段处于开放数据期刊发文量的平稳时期，出现的两个突现词为“互联网＋”和地方政府。

2016 年出现的突现词是“互联网＋”。在 2015 年十二届全国人大三次会议上，李克强总理在政府工作报告中首次提出“互联网＋”行动计划，以便充分发挥互联网在各领域中的优化和集成作用。与“互联网＋”相关的开放数据研

究主要围绕各个传统行业依托移动互联网、大数据、云计算等技术开展的实际应用。例如，在“互联网＋”环境下，企业开放数据平台的应用、智慧民生公共服务模式的研究、数字政府的构建、高校信息化建设模式的探索、档案工作要点研究、图书馆开放数据服务研究等。

2018 年出现的突现词是地方政府。结合 2010 年出现的电子政务、2015 年出现的政务，可以看出，我国对于开放数据研究一直保持在政府、政务这一类的大方向上。基于早些年的理论和实践，信息技术、法律政策和社会环境发展到达了一定水平，国内数据开放也积累了一些经验。政府数据开放主要体现在这样几个方面：各地政府依据信息化管理体制进行政府数据开放平台的建设和管理，根据各地方特色建立政府数据共享和内部公开推进机制，完善面向社会的政府数据开放机制和管理制度，加强政府数据开放平台建设。

通过突现关键词分析可以发现，我国开放数据研究热点的变化规律和发展趋势。2010 年前，我国开放数据研究刚刚起步，尚未形成研究热点；自 2010 年起，逐渐形成了一些研究热点，首先是电子政务（2010 年）、关联数据（2010 年）和信息公开（2011 年），其次是政务（2015 年）、企业管理（2015 年）和“互联网＋”（2016 年），最后是地方政府（2018 年）。从热度回落情况来看，最早回落的是信息公开（2013 年）和关联数据（2014 年），其次是电子政务（2015 年）、政务（2015 年）以及企业管理（2016 年）和“互联网＋”（2016 年），而热度未减的是地方政府（2019 年）。可以预见，政府政务开放数据仍然是未来的研究重点和研究方向。同时，开放数据的应用研究正在逐渐地扩展到其他领域，呈现出多元化的发展趋势，所以其他领域的开放数据应用研究今后也将受到关注。

7　结论

本文运用文献计量学方法，以中国知网期刊全文数据库为数据源，借助可视化软件 Citespace V，对 1996—2019 年我国开放数据研究论文的分布特征、研究热点和研究趋势等进行可视化分析和客观描述。研究发现和结论如下：

（1）论文分布特征。①时间分布：我国开放数据研究大体上可分为三个阶段：第一阶段（1996—2012 年），发文量较少，每年发文量均少于 5 篇，相关研究受到国内较少学者的关注；第二阶段（2013—2016 年），发文量逐年递增甚至成倍增长，开放数据受到国内大批学者的关注；第三阶段（2017 年至今），发文量在到达峰值之后有所回落，但文章数量依然维持在高水平上，日趋平稳和成熟。②作者分布：发文 6 篇及以上的高产作者共有 14 位，其研究领域有所重叠，研究方向主要为开放数据、信息管理、政府数据。作者群形成的成果居多，但也不乏许多高产的个人作者。③期刊分布：载文量在 10 篇及以上的期刊共有 17 种，期刊研究领域主要为图书馆学、情报学、政府政务。另外，也涉及到大数据和电子科技领域。从整体上看，研究焦点大多集中在图书馆学和情报学领域。④机构分布：共有 26 家机构的发文量超过 5 篇。武汉大学信息管理学院的发文量与持续时间都居于国内首位，其次是武汉大学信息资源研究中心。根据最早发文机构所属领域可知，我国最早关注开放数据的领域为图书馆学和信息资源管理领域。

（2）通过关键词共现分析发现，我国开放数据研究的中心性总体上分成三个阶段：第一阶段（1996—2003 年），中心度较大的关键词相对较少，中心度较高的有美国、科学数据和元数据；第二阶段（2009—2014 年），开放数据研究发展迅速，关键词较为丰富多样，中心度较大的关键词是电子政务、开放政府数据、政府数据；第三阶段（2015—2016 年），开放数据研究趋于平稳，中心度较大的关键词是数据开放平台、地方政府和隐私保护。

（3）通过关键词聚类分析发现，我国开放数据的研究前沿和未来趋势首先是持续重点关注政府政务数据开放，包括数据质量及评价与控制、元数据标准、关联开放数据、政府治理、公共管理与安全；其次，开放科学数据逐渐受到重视，包括科学数据共享、科学数据管理、科学数据政策；第三，开放数据应用领域的拓展，包括司法、教育、交通、企业、图书馆数字人

文、创新创业以及数据素养。

(4)通过突现词分析发现,我国开放数据研究热点具有一些变化规律。2010年前,尚未形成研究热点;自2010年起,逐渐形成了一些研究热点,最早出现的是电子政务、关联数据和信息公开,其次是政务、企业管理和“互联网+”,最后是地方政府。从热度回落情况看,回落最早的是信息公开和关联数据,其次是电子政务、政务以及企业管理和“互联网+”,而热度仍然不减的是地方政府。由此可见,我国开放数据研究对于政府政务方面的探讨热度始终不减,无论是政策借鉴、理论研究,还是技术应用等,都处于研究热点上,并且一直延续到现在甚至未来。这表明,政府政务开放数据是开放数据研究的重中之重。与此同时,关联数据、企业管理和“互联网+”也始终是开放数据研究中不可忽视的领域,仍然可以作为未来的研究课题。特别是“互联网+”,由于它可以跟任何一个传统行业相结合,因此为开放数据研究拓展了广阔的应用空间。

参考文献

[1] McKinsey Global Institute. Open data: Unlocking innovation and performance with liquid information[EB/OL]. [2020-02-15]. https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information.
[2] 谭健.开放数据及其应用现状[J].图书与情报,2011(4):42-47.
[3] 盛小平,杨智勇.开放科学、开放共享、开放数据三者关系解析[J].图书情报工作,2019,63(17):15-22.
[4] 夏义堃.国际比较视野下我国开放政府数据的现状、问题与对策[J].图书情报工作,2016,60(7):34-40.
[5] 濮静蓉,刘桂锋,钱锦琳.国内外开放数据的研究进展与述评[J].图书与情报,2017(6):124-132.
[6] 方亚丽.贵阳开放数据指数为何能排全球第四[J].当代贵州,2019(40):38-39.
[7] 闫倩,马海群.我国开放数据政策与数据安全政策的协同探究[J].图书馆理论与实践,2018(5):1-6.
[8] 刘海房,莫世鸿,范冰冰.开放数据最新进展及趋势[J].情报杂志,2016,35(9):163-167.
[9] 顾铁军,夏媛,徐柯伟.上海市政府从信息公开走向数据开放的可持续发展探究——基于49家政府部门网站和上海政府数据服务网的实践调研[J].电子政务,2015(9):14-21.
[10] 李卫东,余奕昊,徐晓林.智慧城市数据开放机制研究——以上海市政府数据服务网为例[J].企业经济,2018,37(6):163-172.
[11] 陈晶晶,陈康清.《2019中国地方政府数据开放报告》暨“中国开放数林指数”在贵阳数博会发布![EB/OL].[2020-02-15].http://gz.people.com.cn/n2/2019/0524/c391492-32976531.html.
[12] 王程韡.“大数据”是“大趋势”吗:基于关键词共现方法的反事实分析[J].科学学与科学技术管理,2015,36(1):3-11.
[13] 郑磊.开放政府数据研究:概念辨析、关键因素及其互动关系[J].中国行政管理,2015(11):13-18.
[14] 于梦月,翟军,林岩.我国地方政府开放数据的核心元数据研究[J].情报杂志,2016,35(12):98-104.
[15] 钱国富.基于关联数据的政府数据发布[J].图书情报工作,2012,56(5):123-127.
[16] 王万华.论政府数据开放与政府信息公开的关系[J].财经法学,2019(1):13-24.
[17] 江国华,何盼盼.数据共享与中国司法现代化[J].中国高校社会科学,2017(1):80-88,158.
[18] 张晓青,盛小平.国外科学数据开放共享政策述评[J].图书馆论坛,2018,38(8):147-154.
[19] 杨现民,周宝,郭利明,等.教育信息化2.0时代教育数据开放的战略价值与实施路径[J].现代远程教育研究,2018(5):10-21.
[20] 刘宁,陈敏.医疗数据开放方法及策略研究[J].中国医院管理,2015,35(9):37-39.
[21] 代涛.健康领域如何掘金大数据[N].健康报,2015-09-28(006).
[22] 冉从敬,刘洁,陈一.Web2.0环境下美国开放政府计划实践进展评述[J].情报资料工作,2013(6):89-95.
[23] 黄如花,陈闯.美国政府数据开放共享的合作模式[J].图书情报工作,2016,60(19):6-14.
[24] 金钰.开放数据与数字图书馆转型[J].实事求是,2019(6):103-107.
[25] 吕春晖.开放数据背景下高校图书馆移动信息技术优化研究[J].情报科学,2020,38(3):124-128.
[26] 周志峰.创新创业视域下促进政府开放数据开发利用的对策分析[J].情报杂志,2017,36(6):141-147.

(收稿日期:2020-04-13)