doi:10.3772/j.issn.2095-915x.2023.04.008

开放科学（资源服务）标识码（OSID）

基于 ITGInsight 的国内外科学数据共享研究对比分析

孟佳琪[1] 支凤稳[1,2] 郑彦宁[2]

1. 河北大学管理学院 保定 071002；
2. 中国科学技术信息研究所 北京 100038

摘要：[目的 / 意义] 对比分析国内外科学数据共享领域的研究现状和演化趋势，从而揭示两者之间的异同，为后续更深入的研究提供有益参考。[方法 / 过程] 以万方数据库、CNKI 和 Web of Science 核心合集作为数据来源，检索并获取国内外相关文献，并借助 ITGInsight 可视化软件，从作者合著、机构耦合、关键词同现和主题词演化四个维度进行挖掘和对比分析。[局限] 数据样本筛选和技术工具选择上可能具有一定的主观性。[结果 / 结论] 国外研究发展迅速，作者合作关系松散，关键词聚类模糊，主题分布广泛且学科交叉性强，偏向共享数据利用。国内作者之间合作密切，关键词聚类明显，主题词具有跨学科、跨领域的特点，偏向数据管理。

关键词：科学数据；共享；ITGInsight；对比

中图分类号：G350

A Comparative Analysis of Domestic and International Studies on Scientific Data Sharing: Based on ITGInsight

MENG Jiaqi[1] ZHI Fengwen[1,2] ZHENG Yanning[2]

1. School of Management, Hebei University, Baoding 071002, China;
2. Institute of Science and Technology Information of China, Beijing 100038, China

Abstract: [Objective/Significance] Comparatively analyzing the research status and evolution trend in the field of domestic and international scientific data sharing, so as to reveal the similarities and differences between the two, and provide useful reference for further in-depth research. [Methods/Processes] Taking Wanfang database, CNKI and Web of Science core collection as

基金项目 国家科技基础条件平台中心课题“面向数据密集型科研的科学数据管理应用模式与技术研究”（2022WT20）；河北省高等学校人文社会科学研究项目“元宇宙时代科学数据共享模式及其应用研究”（BJS2022027）；河北大学研究生创新资助项目“政策工具视角下京津冀智库人才政策演化研究”（HBU2024SS023）。

作者简介 孟佳琪（1999-），硕士研究生，研究方向为科学数据。支凤稳（1987-），博士后，硕士生导师，研究方向为竞争情报、科学数据；郑彦宁（1965-），博士，研究馆员，博士生导师，研究方向为竞争情报理论与方法，E-mail：ynzheng@istic.ac.cn。

引用格式 孟佳琪 , 支凤稳 , 郑彦宁 . 基于 ITGInsight 的国内外科学数据共享研究对比分析 [J]. 情报工程 , 2023, 9(4): 89-107.

doi:10.3772/j.issn.2095-915x.2023.04.008

data sources, this paper retrieves and obtains relevant literature at home and abroad, and carries out mining and comparative analysis with the help of ITGInsight visualization software in four demensions, including author collaboration, organization coupling, keyword presence and subject word evolution. [Limitations] There may be some subjectivity in data sample screening and technical tool selection. [Results/Conclusions] Foreign research develops rapidly, the author's cooperative relationship is loose, the keyword cluster is fuzzy, the topics are widely distributed and the subject intersection is strong, and tend to share the data utilization. Domestic authors have close cooperation, obvious clustering of keywords, and the subject words have the characteristics of interdisciplinary and interdisciplinary, and are biased to data management.
Keyword: Scientific data; Sharing; ITGInsight; Comparison

引言

科学数据主要包括在自然科学、工程技术科学等领域，通过基础研究、应用研究、试验开发等产生的数据，以及通过观测监测、考察调查、检验检测等方式取得并用于科学研究活动的原始数据及其衍生数据[1]。在当今大数据时代，科学数据已经成为各国重要的战略性科技资源[2]，对其有效的管理与开放共享也直接关系着各国的资源利用率和国际竞争地位，作为推进各国科技创新、提高国家核心竞争力、促进社会经济发展的重要举措，科学数据共享的重要地位日益凸显。科研数据共享的价值和意义最早可追溯到1985年，美国科学院（National Academy of Sciences）指出科研数据共享能够强化开放科学需求，并对原始的结论加以验证和改进，进而帮助研究者们克服狭隘的观点和主观的态度[1]。

经过学者们的接续探索，有关科学数据共享的研究成果不断涌现。目前，国内主要聚焦于共享行为的影响因素[3]、共享模型与运行机制[4]、共享政策[5-6]、共享对策及建议[7-8]等方面。伴随着研究成果的积累，其研究视角和内容也变得更加丰富与新颖。如将双路径模型（ELM）与技术接受模型（TAM）相结合，探索数据使用者视角下的共享行为的影响因素[9]；打破医学数据共享的常规思想，提出构建医学数据区块链共享的管理体系[10]；为推动农业科学数据共享进程，构建基于联盟链的农业科学数据管理模型AgriDSM[11]。

国外相关研究得益于更加丰富的信息资源和更加先进的科学技术，研究前沿和热点与国内不尽相同。如基于认知文化和集体行动理论，采用混合方法设计将调查数据与定性数据相结合来克服数据共享带来的障碍[12]；提出构建一种数据共享平台，为制定更加科学合理的政策提供信息，同时监控数据共享实践，并引导队列和数据基础设施本身的资金优先级[13]；通过科学家们互相共享交换未经审查的数据，同时采用问卷调查的方法去了解数据共享的优劣性，以此来揭示数据共享的问题与前景[14]。国内外研究一直在不断完善和发展演化中，从不同维度对比分析二者的研究现状及演化过程，可以更好地了解我国在科学数据共享领域中的优势与不足，

有助于为我国把握研究前沿、开展创新性的研究提供新思路。尽管已有学者对科学数据共享相关研究进行了梳理，采用的方法与工具也多种多样，如文献计量学方法[15]、Citespace 软件[16]、联机分析处理（OLAP）方法[17]、理论分析[18]等，但鲜有学者从多个维度对国内外相关研究进行可视化对比。此外，已有研究表明，ITGInsight 在分析数据量、清洗功能、安全性、兼容等方面更具特色和优势，已得到学者们的青睐[19]。本研究旨在探索国内外科学数据共享领域的作者合著情况、机构耦合情况、研究热点与主题分布情况，同时识别出核心作者群、核心机构群，以及各时期的研究演化情况。而 ITGInsight 在功能上拥有更加突出的优势，更能清晰直观展示各方面的现状及态势，如作者合著方面以不同颜色来区分各个作者合著群，且还可以显示不同作者作为一作、二作、三作等不同的发文量。因此，本研究利用 ITGInsight 可视化分析软件，探索国内外在科学数据共享领域的研究热点、关联网络，以厘清其研究现状和演化发展态势，作者及其机构、关联词和主题词的演化分布。借鉴国外相关研究的优势和经验，为促进国内科学数据共享理论研究与实践发展提供有益的帮助。

1　文献来源及研究方法

1.1　文献检索与年度分布

本文以万方数据库、CNKI 为主要中文文献来源，以 Web of Science 数据库为主要外文文献来源，利用检索式“主题 :(科学数据共享) and 关键词 :(科学数据 or 科研数据 or 数据共享)”和“(TS=(scientific data sharing)) OR TS=(Scientific research data sharing)) AND AB=(data sharing)”分别进行检索，时间截止到 2022 年 12 月 9 日。对于中文文献，先进行去重，再剔除无作者的、以“序”“前言”“简介”“卷首语”为题的和专业相关度不高的文献。而对于外文文献，先按照相关度从高到低排序，再从 Web of Science 核心合集中剔除重复、关联度低的文献，且语种设为“English”。经过筛选，初步确定中文文献共 3095 篇、外文文献共 4642 篇，作为本文的研究对象。

对第一轮数据清洗后得到的词表进行二次清洗，得到最终用于可视化分析的有效文献，其中中文 3065 篇，外文 4562 篇。文章对 2001—2022 年期间，国内外有关科学数据共享的发文量做了统计，以此对国内外在该领域下的研究趋势进行分析与预测，见图 1。

从图 1 可以看出，无论是国内还是国外，有关科学数据共享的发文量总体呈上升趋势，尤其是国外在近几年呈现出较为明显的“指数式增长”特征。

国外在此期间的发文量总体可分为两个阶段，即缓慢增长阶段和快速增长阶段。在缓慢增长的这十年（2001—2011 年）中，OECD（Organization for Economic Cooperation and Development，经济合作与发展组织）成员国以及中国等 30 多个国家于 2004 年 1 月签署了《开放获取公共资助的科学数据宣言》[20]，成为开启科学数据共享时代的导火索。随后以英美、

doi:10.3772/j.issn.2095-915x.2023.04.008

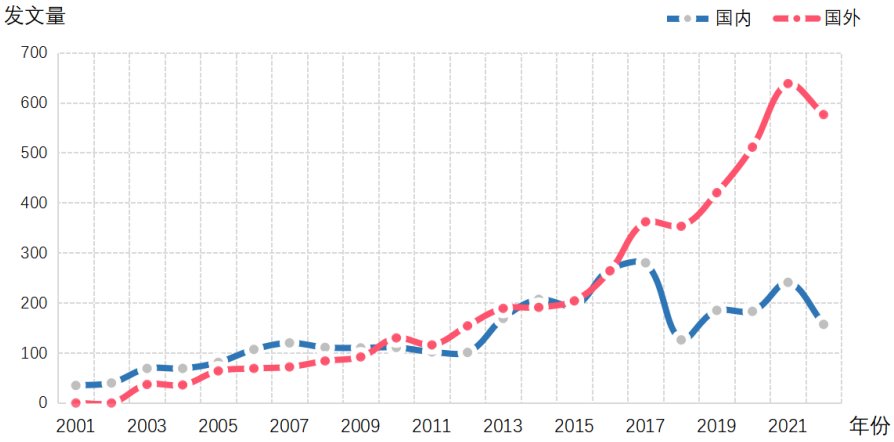

图 1 2001-2022 国内外科学数据共享年度发文量

澳大利亚为代表的各政府部门、科研机构和高校等制订的一系列科学数据共享政策，如《开放获取公共资助科学数据的原则和指南》（2007）[21] 等，进一步推动了科学数据研究的进程。自 2012 年开始进入快速增长阶段，随着各国的政策法规逐步趋于成熟，以 2015 年发布的《科学出版物与研究数据开放存取指南》（第二版）为代表，开启了开放科学数据试点的实施工程，科学数据研究进入快速发展阶段。

我国的发文量分别在 2012 年和 2018 年出现两个拐点，共经历三个阶段。2001—2012 年是缓慢上升期，我国于 2000 年开始主导推进科学数据的管理工作；随着“实施科学数据共享工程”的建议被提出，“科学数据共享工程”于 2002 年正式启动，在此期间，我国科学数据共享工作一直处于初步发展的稳定期。到 2012 年以后，进入快速上升期，国家科学技术部于 2014 年召开第一届“中国科学数据大会”，国务院办公厅于 2018 年出台首个国家层面的《科学数据管理办法》，有效促进了研究成果的产出。2018 年至今属于起伏发展期，以国家层面的政策为依据，各地方也相继制定与完善适合本地实际的数据共享政策，相关研究成果整体呈上升趋势，但在 2022 年有所下降，这可能与成果发表的滞后性有关。

1.2 研究方法及过程

文章基于文献计量、类比分析、定量定性分析等研究方法，利用 ITGInsight 软件，对国内外科学数据共享领域的作者、机构、关键词及主题词进行预处理和数据分析，并以可视化的方式输出网络图、聚类图和演化图，从而对该领域的发展现状、不同特征以及未来发展的趋势做出综合性评述和总结。

ITGInsight 是一款高级的科技文本挖掘与可视化分析工具，主要针对专利、论文、报告、报刊等科技文本进行数据挖掘和图谱建立 [22]。在本文中，作者合著、机构耦合、关键词同现以及主题演化部分均以 ITGInsight 分析工具作为技术支持，并辅以 Excel 对数据集进行初步数据统计。首先，分别对国内外所得的数据集

进行清洗，将时间设为 2001—2022 年期间，从清洗后的数据中提取作者、机构等研究子对象（提取条件为 n≥2，n 代表研究子对象的频数），形成词表；其次，利用得到的词表对原数据集进行数据分析；最后，分别对排名前 50 的作者、机构、关键词、主题词构建图谱，并对图谱进行解读。

2 基于 ITGInsight 的国内外可视化分析

2.1 作者合著

对不同格式、不同写法但属于同一作者的姓名进行合并，利用 ITGInsight 软件分别提取国内外在科学数据共享领域发文量中排名前 50 的作者，构建作者合著网络图，并形成聚类关系图。图中节点数字代表该作者的发文总量，节点大小与之成正比，各节点之间的连线代表相连两作者有合著关系，连线的粗细代表与该作者合著次数的多少，相同颜色的节点集合代表一个作者合著群。

2.1.1 中文文献作者

从图 2 可以看出，发表中文文献的大部分都是国内作者，在科学数据共享领域研究中的前 50 位作者中共有 6 个合著群，1 个作者合著对，其余没有合作关系的 8 位作者分别作为独立个体存在。比较直观的是，中文文献作者之间的合著关系是比较紧密的，且在合著群内部的联系更加紧密。在这 6 个作者合著群中，较核心的合著群体是以王卷乐、诸云强为代表的 7 人作者团体和以王健、赵华为代表的 8 人作者团体；其中，除分别以钱庆和王松为代表的 4 人组成的合著群是独立存在之外，其余 4 个合著群之间也有合著关系。从作者个人来说，比较高产的核心作者有王卷乐、诸云强、王健、钱庆，所代表机构是中国科学院地理科学与资源研究所、江苏省地理信息资源开发与利用协同创新中心、中国农业科学院农业信息研究所、中国医学科学院医学信息研究所。值得注意的是，在单人作者中关健的发文量就已达到了 26 篇，是该领域中较为优秀的研究者。

2.1.2 外文文献作者

相比中文文献，外文文献作者之间的合著关系就略显稀疏，相对的单人作者数量较多。从图 3 可以看出，发表外文文献的作者既有国内学者也有国外学者，且数量不一，在前 50 位作者中共有 4 个作者合著群，2 个作者合著对，其余没有合作关系的 9 位作者作为独立个体存在。4 个作者合著群的发文量不相上下，其中以 Chen, X、Wang, G 为代表的 10 人作者团体、以 Zhao, Y、Chen, Y 为代表的 11 人作者团体和以 Zhang, Z、Foster, I 为代表的 10 人团体之间也有合著关系；而以 Alfonso, F 为代表的 6 人作者团体与其他 3 个合著群没有合著关系。以作者个人来说，每位作者的发文量趋于一个平均水平，不存在真正意义上的高产作者。此外，考虑到发表外文文献的国内作者及机构可能会影响国外总体的合著程度，尝试将这些作者排除在外，发现 4 个合著群中有一半以上的人都是国内作者。因此，若排除相关数据，则更加凸显出国外作者之间合作关系的稀疏，并不会影响最终的结论。

doi:10.3772/j.issn.2095-915x.2023.04.008

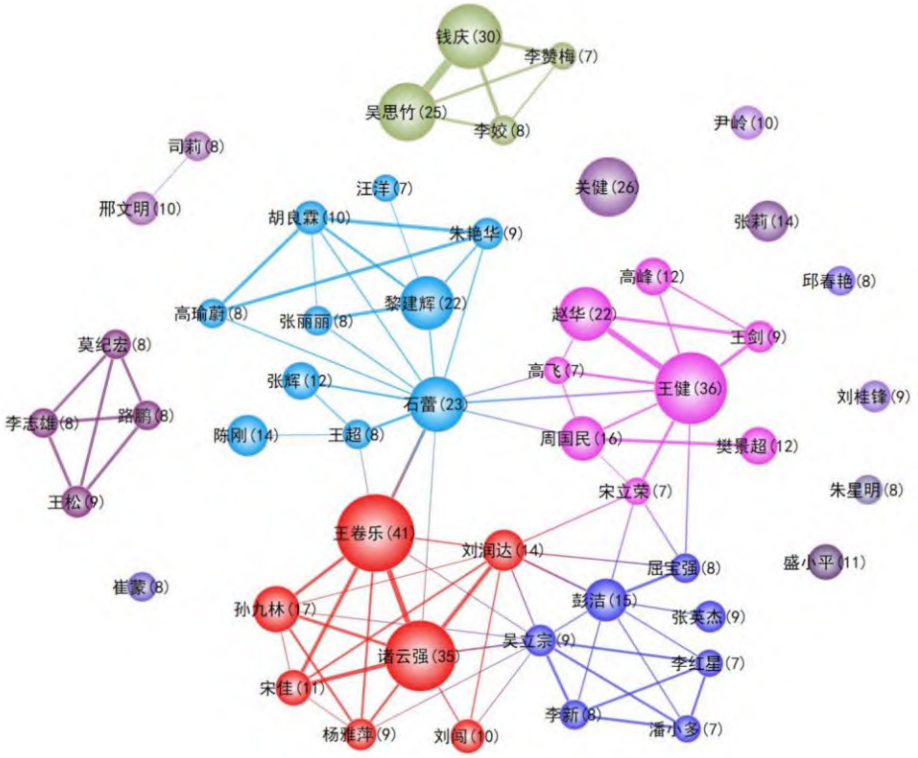

图 2 国内作者合著网络

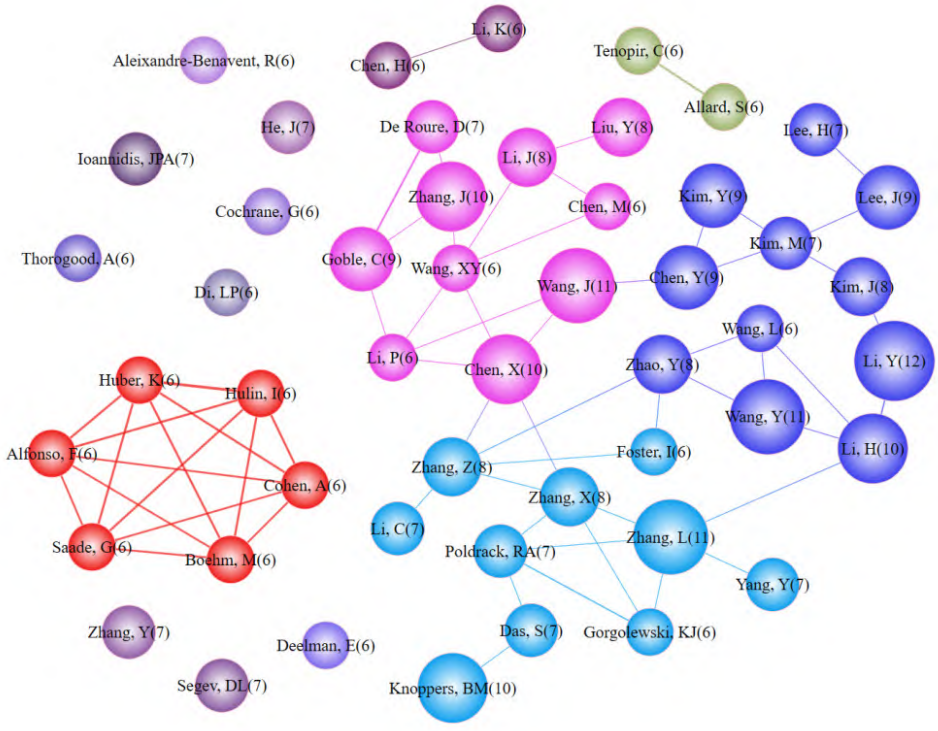

图 3 国外作者合著网络

2.2 机构耦合

利用 ITGInsight 软件分别提取国内外在科学数据共享领域发表文献数量排名前 50 的代表机构，构建机构耦合网络，并形成聚类关系图。其中，每个节点代表一个机构，节点数字代表机构发文量，节点大小与之成正比，节点之间若有连线则代表机构之间有合著关系，不同颜色的节点代表不同的机构个体。

2.2.1 国内机构

从图 4 可以看出，在排名前 50 的国内代表机构中，大多以独立的个体存在，基本不存在合著关系。据统计，国内在科学数据共享领域拥有研究成果的机构共有 1996 所，发表文献数量在 2 篇以上（包含 2 篇）的机构就有 440 所。其中发表文献最多的机构是中国科学院地理科学与资源研究所，进一步分析发现，该机构倾向于其他机构少有涉足的地球科学方面的研究，也说明科学数据共享在地理科学、资源科学领域更具有可研究的价值和前景。其次，武汉大学、中信所、中科院等机构也在科学数据共享领域取得了不错成就。总的来看，这些代表机构的发文量从 8 到 86 篇不等，上下幅度大且多都集中于 30 篇以下，平均水平在 16.8（840/50），总体偏下。

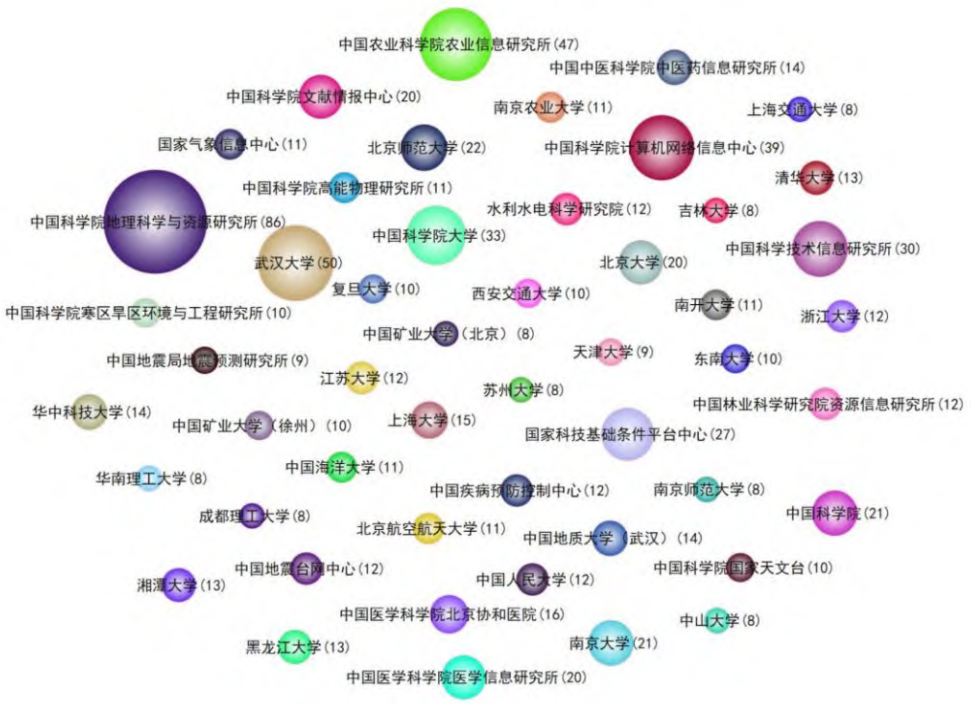

图 4 国内机构耦合网络

2.2.2 国外机构

如图 5 所示，国外机构的耦合程度较国内来说并没有明显差别。据统计，国外在科学数据共享领域拥有研究成果的机构有 7533 所，发表文献数量在 2 篇以上（包含 2 篇）的机构就有 1785 所。其中发文量较多的代表机构有

doi:10.3772/j.issn.2095-915x.2023.04.008

Stanford Univ、Univ Oxford、Harvard Univ、Univ Calif San Diego，这四个机构的文献量均在 65 篇以上。进一步分析发现，这 4 个机构的研究主题集中于 Data Sharing、Computer Science、Medical Informatics、Bioinformatics 等，这说明数据共享、计算科学、医学情报和生物信息学等领域的学者更青睐科学数据共享的研究。总的来看，各机构的发文量从 29 到 100 篇不等，上下幅度中等且多都集中于 30 篇以上，平均水平在 42.36（2118/50），高于国内。

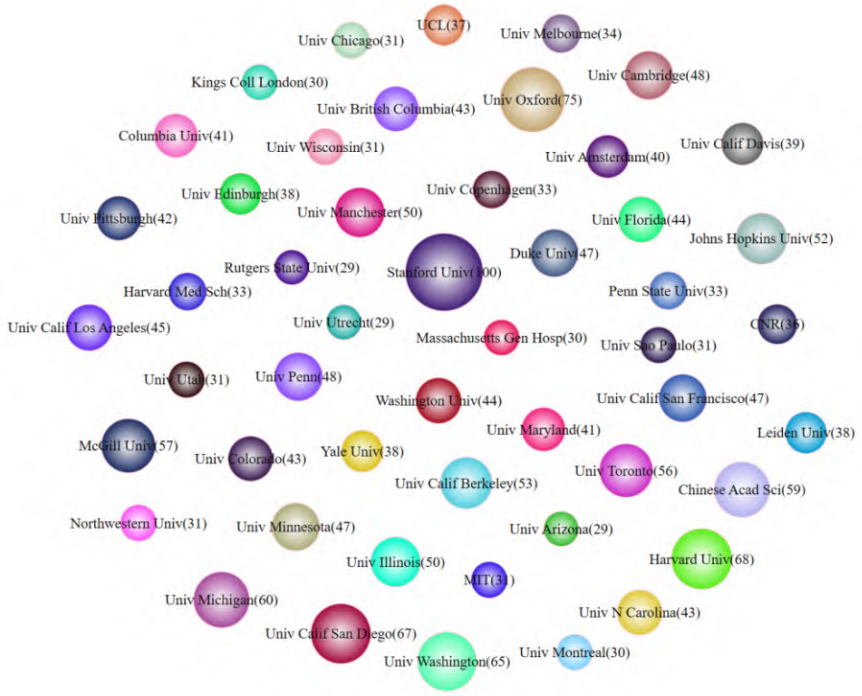

图 5 国外机构耦合网络

2.3 关键词同现

关键词同现分析可以帮助我们更加明确绘制概念、想法和问题之间关系 [23]，确定该文献集合所代表的学科主题之间的关系，从而揭示该学科的研究水平及学科结构，概述该学科的研究热点，分析其发展过程及趋势 [24]。为此，本文利用 ITGInsight 软件分别提取出现频次排名前 50 的关键词，构建关键词同现网络，形成聚类关系图。其中，节点数字代表该关键词的词频，节点大小与之成正比，各节点之间的连线代表关键词之间的同现关系，连线的粗细代表同现次数的多少，相同颜色的节点集合代表一个关键词网络。

2.3.1 中文关键词

以中文文献的关键词代表国内的关键词数据，图 6 所示的是国内出现频次排名前 50 的关键词，各节点之间的关联程度较为紧密，经过聚类后，这些关键词被分为了三大类。

第一类是以“数据共享”“科学数据”为中心词，其关联词或衍生词以“开放数据”“科研数据”“农业科学数据”“科技资源”“元

数据”“数据管理”“开放科学”“开放共享”等为代表。这类关键词大多研究的是各类科研、科技数据的共享以及如何被共享的问题，如早期浙江省为推动科技资源开放共享而实施的创新券政策[25]；基于我国现有的科学数据共享协议提出的魏公村科学数据双许可证（草案）[26]，同样促进了农业科学数据的开放共享。

第二类是以“科学数据共享”“大数据”“数据库”为中心词，其关联词或衍生词以“信息化”“云计算”“共享服务”“信息技术”“信息系统”等技术支持类词为代表。这类关键词主要研究科学数据共享的技术、平台、系统、机制等，一般是解决其如何构建、如何应用、如何开发等问题。如国家微生物科学数据中心的建设[27]，极大程度上使海量的微生物数据资源得到了有效的规范整合和开放共享；基于云计算技术面向服务的体系架构（SOA）思想的提出，有助于实现科学数据的资源聚合[28]。

第三类是以“项目信息门户”“工程施工信息管理”“施工管理信息系统”“地理信息系统”为主的信息管理类关键词，顾名思义，这类关键词与信息管理有关，主要面向信息资源管理和共享服务，旨在解决数据管理、数据应用、数据治理等问题。如盛小平等[29]曾从数据管理和数据治理两个层次出发去探索二者之间的差异与联系，从而为制定科学数据开放共享政策提供帮助。

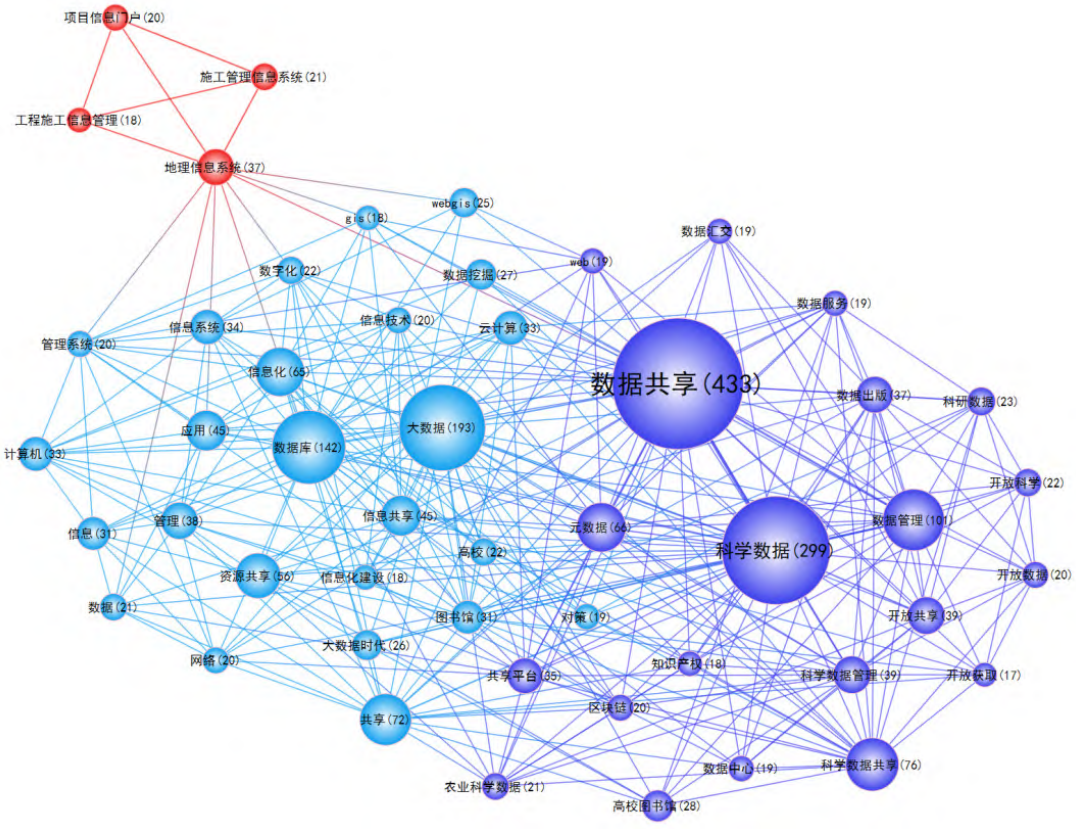

图 6 国内关键词同现网络

doi:10.3772/j.issn.2095-915x.2023.04.008

2.3.2 外文关键词

以英文文献的关键词代表国外的关键词数据，图 7 所示的是国外出现频次排名前 50 的关键词，各节点之间的关联程度紧密，经过聚类后，这些关键词被分为了四大类，类别与国内相比略有差异。

第一类是以“Data Sharing”“Open Science”“Reproducibility”为中心词的数据类关键词，其下还有一些子类，如以“Data Management”“Open Access”“Collaboration”“Metadata”为代表的数据管理子类、以“Ontology”“Transparency”“Interoperability”为代表的数据特性子类、以“Bioinformatics”“Climate Change”为代表的生物科学子类等。与国内涉足的地球科学领域相比，国外所涉足较多的是生物科学领域，主要研究生物科学、医学界等如何实现科学数据共享的问题，如在临床试验数据共享领域，Gudi, Nachiket 团队 [30] 基于现有数据共享政策，提出建立一个中立的体制来监督数据信息的共享的建议。

第二类是以“COVID-19”“SARS-Cov-2”“Public Health”为中心词的时事类关键词，尤其是新冠疫情爆发之后，这类关键词的数量呈爆发式增长，此外还包含了以“Social Media”“Twitter”为代表的社交媒介子类等。这类关键词主要出现在如何实现新冠疫情有关数据共享的研究中，如 San Torcuato, Maider 团队 [31] 持续跟踪 2020 年 1 月至 2021 年 3 月有关 COVID-19 的出版物和主题演变，目的就是为了调查探究与 COVID-19 相关的研究交流、论文等数据共享的程度。

第三类是以“Machine Learning”“Big Data”“Cloud Computing”为中心词，其关联词或衍生词主要是以“Artificial Intelligence”“-Data Mining”“Blockchain”为代表的技术支持类关键词，此外，还包括以“Privacy”为主的信息隐私安全子类和以“Cancer”为主的医疗健康子类。其中，信息隐私安全是近年来科学数据共享领域下比较火热的研究议题，主要是用于解决数据共享过程中涉及到的信息泄露、信息隐私等问题，基于区块链的协作科学实验信任架构 [32]，有助于在保证互操作性、隐私性、可追溯性和信任度的基础上实现科学数据共享。

第四类是以“Citizen Science”“Crowdsourcing”“Biodiversity”为主的生物科学类关键词。新冠疫情爆发之后，各国学者呼吁相关科学数据公开与共享，并针对数据管理、数据共享、数据利用与数据治理方法与技术进行了系列的探索，产出了丰富的研究成果。可见，科学数据共享也是新冠疫情背景之下促进各国学术交流与科研产出的重要桥梁。

2.4 主题词演化

作为学科新兴趋势探测方法之一，高频主题的演化分析有助于了解领域主题产生、消亡、增强、减弱、聚合和裂变的过程 [33]。对高频主题词汇总分析，不仅可以识别研究热点，还能为后续的相关研究提供方向和依据。本文利用 ITGInsight 软件分别提取国内外在科学数据共享领域中出现频次排名前 50 的主题词，构建主题词演化网络图，同时列出排名前 20 的主题词，便于直观了解。其中，节点数字代表该主题词的词频，节点大小与之成正比，各节点之间的连线代表主题词之间的演化关系，相同颜色节点的连线代表同一主题词的演化路径。

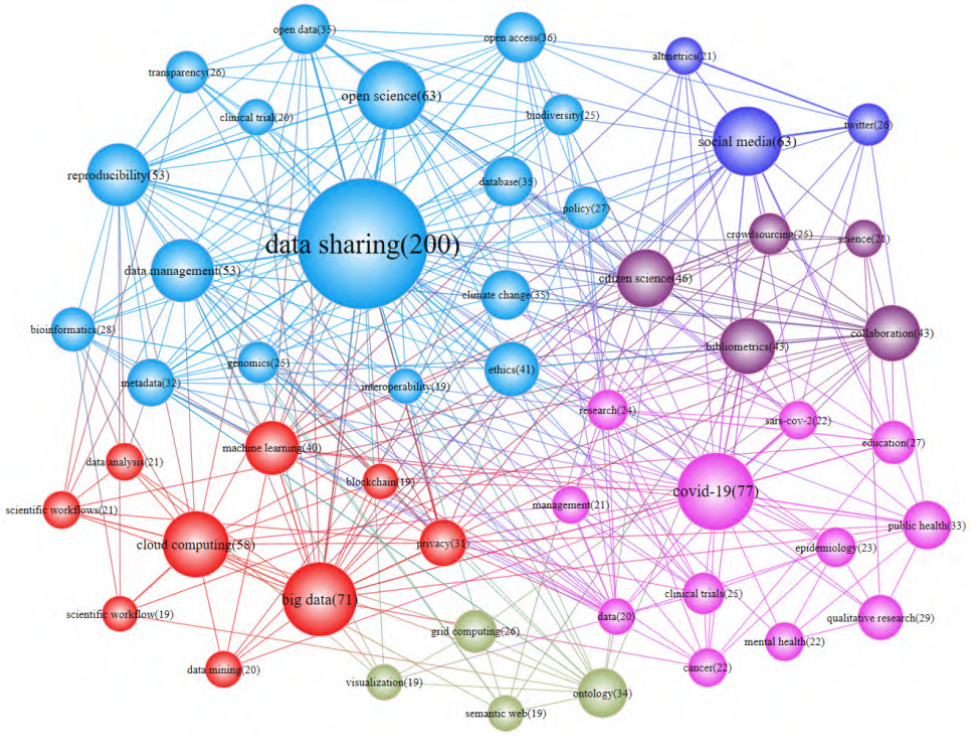

图 7 国外关键词同现网络

2.4.1 中文主题词

以中文文献的主题词代表国内的主题词数据，从图 8 可以看出，2001—2012 年，国内学者主要对图书馆、信息化、数据库等进行研究，以数据统一发布和数据共享为目的构建较为理想的服务体系框架 [34]，使在技术层面实现数据共享成为了可能。从 2013 年开始，科学数据共享经历了为期五年的高产研究阶段，研究主题多样化，但始终离不开资源共享和图书馆建设两大主题。相比前一阶段，演化出来的新主题有工程施工、数据管理、信息系统、地理信息、科研人员等；进一步分析得出，2014 年开始出现的工程施工，很大程度上与当年发布的试点共享工程的政策有关。下一阶段自 2018 年开始，由于《科学数据管理办法》出台，使得科学数据共享领域研究有了新的进展，研究主题开始涉及数据安全、标准规范、数据保护等。随着大数据时代的到来以及新一代信息技术的飞速发展，数据共享已不再是一种奢求，数据安全与隐私保护问题成为国内学者研究的重点，而该领域下新主题的出现也许与此有关。此外，有关地球科学的研究也在此阶段兴起，近几年全球气候变暖、生态环境恶化等，使得国内学者更重视地球科学研究，特别是进入 21 世纪以来，更加强调用先进技术去认识、理解和保护人类赖以生存的地球 [35]。

为了更加清晰地了解国内有关科学数据共享的高频主题词，进一步分析出该领域下的研究前沿及热点，进而推测未来研究发展趋势，将排名前 20 的高频主题词列举如下，见表 1。

doi:10.3772/j.issn.2095-915x.2023.04.008

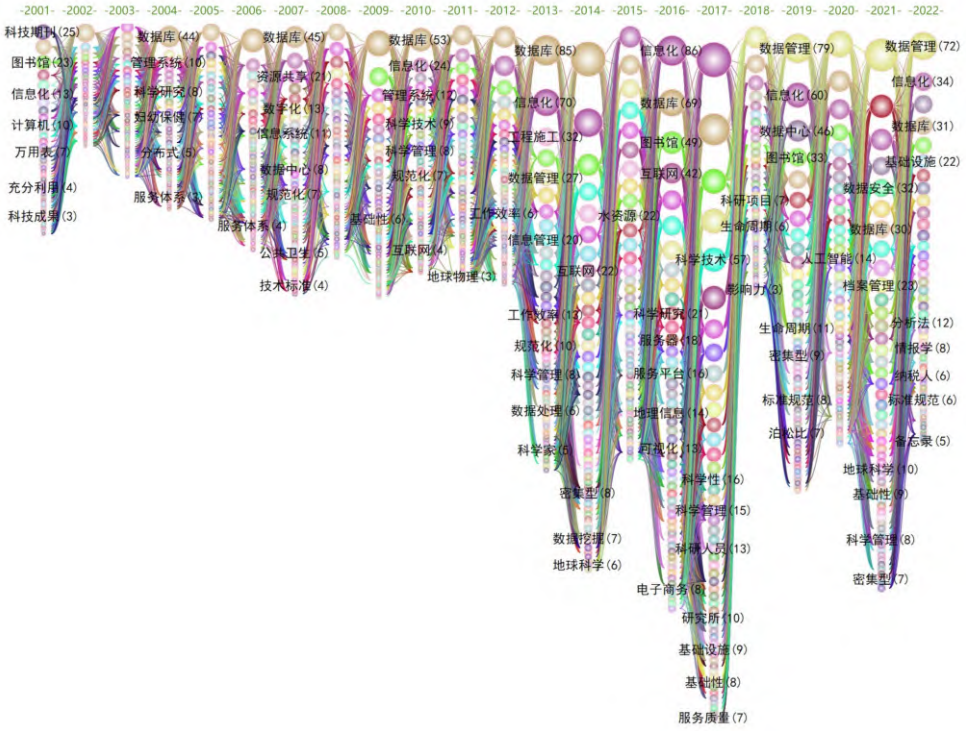

图 8 国内主题词演化网络

表 1 国内排名前 20 的高频主题词及词频

序号	国内主题词	词频	序号	国内主题词	词频
1	数据库	2949	11	管理系统	472
2	信息化	2451	12	科学技术	462
3	数据管理	2310	13	计算机	460
4	图书馆	1973	14	水文学	400
5	数据中心	951	15	仪器设备	397
6	科技期刊	785	16	工程施工	397
7	数字化	757	17	地球科学	389
8	互联网	641	18	国土资源	385
9	资源共享	634	19	档案管理	382
10	科学研究	513	20	实验室	380

可见，在科学数据共享领域，国内研究最常见的热门主题主要集中在数据库、信息化、数据管理、资源共享、图书馆、科学技术等。显而易见，构建资源整合的数据库是实现科学数据共享的重要前提，是提供共享数据来源的平台保障，在逐渐信息化的同时，还要对数据进行管理，以保证数据资源能够顺利共享与利用。而地球科学、数据安全、数据保护等是目前比较前沿的热门主题，上述分析也可以证实这一点。

2.4.2 外文主题词

以英文文献的主题词代表国外的主题词数据，从图 9 可以看出，国外相关研究主题在整体上呈现出逐年延伸与扩展的演化趋势，可以将其分为三个阶段：缓慢起步阶段、稳定增长阶段和快速发展阶段。2001—2010 年，研究主题词由前期单一且比较分散的“scientific discipline”“life science”“data set”到后期转变为“scientific datum”“raw datum”“information retrieval”等，研究重点倾向于共享数据的获取，如联邦政府通过强制公开科研数据 [36]、鼓励科学发现和教育 [37] 来实现科学数据共享。自 2011 年开始，主题词的演化开始呈快速增长趋势，2012 年的“scientific inquiry”、2013 年的“data share website”、2016 年的“na tech event”以及 2017 年的“data collection”等都是新兴代表，在这一阶段，学者尝试利用先进的理论、科学技术、工具与政策等去实现生物医学领域的科学数据共享，如提出共享生物医学大数据的政策框架 [38]、突破科技实现共享人类样本和患者数据 [39] 等。直到 2019 年新冠疫情之后，随之而起的热门主题是“covid-19 pandemic”“-scientific research”“social medium”“public health”“biologica research”“surveillance system”“data management”等。该阶段的研究一方面是为了处理随疫情而来的大量新鲜数据，另一方面是为了解决受疫情影响带来的数据共享滞缓及其恢复的问题。如提倡及时共享试验数据 [40]，获得一手的医疗临床数据，以期满足应对 COVID-19 挑战的需求，更快地实现对

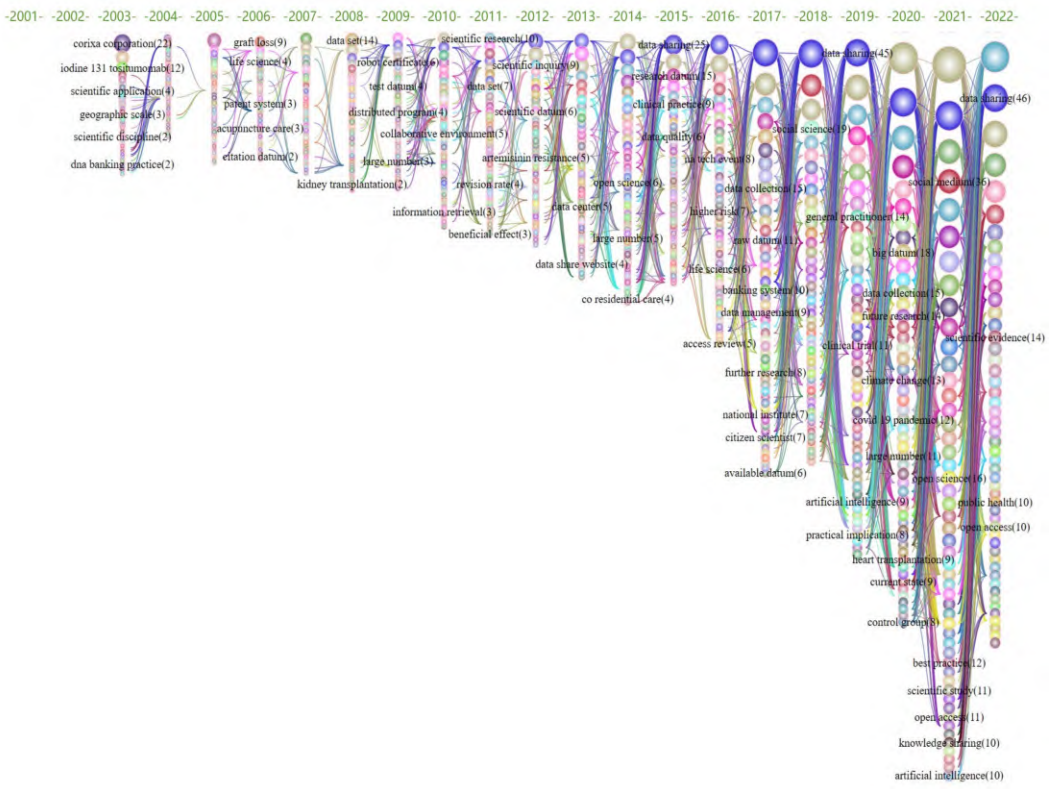

图 9 国外主题词演化网络

doi:10.3772/j.issn.2095-915x.2023.04.008

COVID-19 乃至所有疾病的科学理解；有团队研究发现抗击新冠肺炎疫情最有效的方法之一就是国际信息共享，但这种国际共享需要在一定条件的基础上[41]；科学界在抗击 COVID-19 方面也取得了几项重要进展，并在全球注册了 2500 多项临床试验，这些数据有期待被共享[42]。这一阶段整体上处于快速发展之中，尽管 2022 年的主题词有所减少（或许与研究成果发表的时滞性有关）。随着公众逐渐认识到数据共享的重要性及共享实践在多领域的推进，科学数据共享的研究必将迎来更大的突破。

同样将国外有关科学数据共享排名前 20 的高频主题词列举如下，以便进一步分析出该领域下的研究前沿及热点，进而推测未来研究发展趋势，见表 2。

表 2 国外排名前 20 的高频主题词及词频

序号	国外主题词	词频	序号	国外主题词	词频
1	research question	904	11	scientific data	273
2	data sharing	784	12	open data	240
3	social medium	573	13	data set	227
4	corixa corporation	484	14	general practitioner	222
5	scientific community	461	15	nonsmoking mother	211
6	primary care	359	16	scientific literature	209
7	climate change	353	17	land plot	200
8	scientific research	344	18	data reuse	196
9	social network	323	19	citizen science	187
10	research data	299	20	scientific workflow	184

在科学数据共享领域，国外热门主题主要集中于研究问题、数据共享、社交媒体、corixa 公司（医药研发公司）、科学共同体、基础医疗全科医生、气候变化、开放数据、数据重用等。相比国内，国外在医学领域、生物科学领域的科学数据共享研究成果更为突出，更加倾向于人类的生命健康科学研究。事实上，相比人文社会科学，自然科学尤其是医学、生物学领域的数据共享实践也更为广泛与深入。

3 国内外多维对比分析

通过对国内外科学数据共享研究的可视化分析发现，国内外相关研究在作者合著、机构耦合、关键词同现和主题词演化方面的表现不尽相同。接下来，本文将从这四个维度进行进一步的对比分析，深入挖掘国内外在各维度下的异同，进而为我国未来相关研究提供新思考。

3.1 作者合著维

受领域范围、教育水平等因素的影响，外文文献的独立作者居多，各作者之间的合作关系也较为松散。而国内作者则更倾向于合作发文，故多人作者团体居多。从科研人员在科学数据共享中扮演的角色及其作用这一角度来看，作为科学数据开放共享的核心力量，科研人员拥有着作为生产者、传播者、管理者和利用者的不同角色定位[43]，这些角色也可能是影响他们合作关系的因素之一。其次从领域分布来看，同一个机构或组织内部的人更有可能拥有这种合作关系，而那些不同组织的人可能会因为平时交集不深而失去这种合作的机会。

表 3 国内外多维对比结果

维度	国内（中文文献）	国外（外文文献）
作者合著	作者合著关系紧密 一般以多人团体为主 发文量高低不等，分布不均匀	作者合著关系松散 独立作者居多 每位作者发文量趋于一个平均水平
机构耦合	机构之间关联程度极低 一般为独立机构，多集中于高校 发文量高低不等，分布不均 平均发文水平偏低	机构之间关联程度极低 一般为独立机构，多集中于高校 发文量高低不等，分布较均匀 平均发文水平高于国内
关键词同现	关键词聚类明显 各词之间同现率偏高 更倾向于数据管理	关键词聚类模糊 各词之间同现率偏低 更倾向于数据利用
主题演化	“S”型演化趋势 跨领域、跨学科 以实现数据共享为主轴 新领域：地球科学、信息安全等	“J”型演化趋势 跨领域、跨学科且学科交叉明显 以共享数据应用为主轴 新领域：生物、医学等

3.2 机构耦合维

虽然国内和国外的机构耦合度并没有太大差别，但各机构的研究基础与研究方向也会影响其发文量和发文的水平。如国外机构立足社会需求，选择能够适应当前发展的研究话题，并进行实证与仿真，可能会在某一时间段产出大量的研究成果。而国内则更倾向于针对近几年的热点研究理论分析与总结归纳，研究方向与方法有待于丰富。此外，可能因为国内文献机构主要来自于中国，而国外来自于国际范围内的多个国家，所以国内发文量相对偏低。

3.3 关键词同现维

关键词同现极大程度上反映了研究重点和研究方向，国内的关键词有较高的同现率且词之间聚类明显，说明国内相关研究的重点清晰、方向明确。进一步分析发现，其在总体上更偏向于数据管理。相比国内，国外的关键词同现率偏低，且各词之间的聚类较为模糊，说明国外在科学数据共享领域的研究范围跨度广、方向多，进一步分析发现，其在总体上更偏向于

doi:10.3772/j.issn.2095-915x.2023.04.008

数据利用一类。

3.4 主题词演化维

由图 8、图 9 可以看出，国内外的主题词演化均有明显的路径特点，国内研究呈现跨领域、跨学科的特征，且主要围绕如何高效实现数据共享这一目的开展，总体呈“S”型演化趋势，近阶段主题词多集中于地球科学、数据管理、数据安全等。国外研究的跨领域、跨学科性更明显，且学科之间具有较强的交叉性，研究主要围绕如何实现共享数据的利用这一目的进行，总体呈“J”型演化趋势，近阶段主题词多集中于生物、医学等。

4 总结与展望

文章基于 ITGInsight 可视化分析软件，利用文献计量和对比分析相结合的方法，从作者合著、机构耦合、关键词同现和研究主题四个维度对国内外科学数据共享的相关研究进行深度挖掘，揭示了国内外研究现状，并总结了国内外研究的不同特征，对了解和预测国内该领域的研究重点及未来发展方向具有一定的参考意义。主要研究发现如下：

（1）国外研究发展迅速，作者合作关系松散，机构耦合程度低，关键词聚类模糊，主题分布广泛且学科交叉性强，偏向共享数据利用。

（2）国内研究发展平缓，作者之间合作密切，机构耦合程度低，关键词聚类明显，主题词具有跨学科、跨领域的特点，偏向数据管理。我国仍面临诸如对科学数据共享的重视度不够、科学数据主权流失、科学数据共享技术不成熟、科学数据共享范围不广等问题 [44]，这可能与国内外政策、经济、技术等差异有关。

为促进我国科学数据共享研究更好地发展，以更快地走向国际，引出新的议题思考与探索，本文尝试就未来研究发展和方向提出以下几点建议：

（1）政策保障。科学数据共享工作的开展离不开政策法律的引导和推动 [45]，有关部门应该进一步制定与完善相关政策，如共享策略、共享原则以及相关的数据安全与保护等政策，这也是科学数据共享中所面临的主要障碍之一。

（2）资金支持。资本可得性障碍对科学数据共享效果具有反向的影响 [37]，建立合理的数据共享激励机制，肯定数据生产的贡献，为其提供必要的经费支持，提升其数据共享行为。

（3）加强多方合作关系，促进学科之间的交流与协作。首先，对外要将科学数据主权牢牢掌握在自己手中 [17]，加强国际合作交流，了解国外研究话题和研究热点，拓宽科学数据共享范围；其次，对内注重跨学科、跨机构、跨地域之间的交流与合作，主动汲取其他学科的理论优势和方法技术，不断拓展自身领域的新天地；最后，在保证科学数据主权的基础上，注重在跨国数据流动方面和各机构之间的合作，同时也要充分调动科学数据多方利益相关者的共享积极性，尤其是我国研究者的积极性。

（4）促进新兴技术应用，强化数据管理。加强对科学数据的管理，为科学数据的广泛获取与开发利用提供支持 [47]，要利用好现代信息技术，尤其是新兴的区块链、数据安全加密等

技术，解决科学数据共享过程中涉及到的数据安全性、数据严密性等问题，这也是目前国内研究比较热门的一个话题。此外，科学数据共享正面临着向人文社会科学领域、向微观数据管理、向多学科交叉融合的趋势发展[48]，数据的利用和价值正逐渐受到重视。随着 AI、元宇宙以及 Chat-GPT 等新兴概念和技术的涌入，未来将持续出现一些新特征、新模式、新方法，值得研究人员继续探索和挖掘。

本研究还存在一些不足之处，有待进一步关注与完善：

（1）文章所选用的文献数据是经过人工筛选、去重，可能会存在误删、漏删和数据交叉或重叠等情况，尽管作者已尽量保证数据的准确性。

（2）文章利用 ITGInsight 软件对文献数据进行可视化分析，研究结果很大程度上依赖于软件本身，具有一定局限性，后续可尝试借助其他方法和工具进行综合对比分析，从而使研究结果更加全面可信。

（3）文章关注作者、机构、关键词和主题词四个维度，未曾涉及其他维度和各维度之间的关联，未来可以考虑对不同维度进行交叉分析。

参 考 文 献

[1] 国务院办公厅 . 国务院办公厅关于印发科学数据管理办法的通知 [EB/OL]. [2023-05-16]. http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm?ivk_sa=1024320u.

[2] 周文能 , 刘云 , 王刚波 . 国内外科学数据管理与共享政策分析及对国家自然科学基金的启示 [J]. 中国科学基金 , 2023, 37(1): 150-160.

[3] 王丹丹 , 刘清华 , 王晓梅 . 科学数据共享行为影响因素的元分析 [J]. 图书馆学研究 , 2021(22): 74-84.

[4] 冯媛 . 科学数据开放共享的价值共创模型及运行机制研究 [J]. 图书馆 , 2022, 50(9): 29-37.

[5] 孙苗 , 姜晓轶 , 王子珂 . 海洋科学数据共享政策法规与标准规范研究及启示 [J]. 科技导报 , 2022, 40(10): 22-29.

[6] 戚筠 , 何琳 . 国内外数据知识库的科学数据开放共享政策对比研究 [J]. 图书馆学研究 , 2023, 529(2): 42-53+81.

[7] 张潇月 , 顾立平 , 胡良霖 . 国内外开放科研数据重用困境解决措施述评 [J]. 图书馆 , 2021, 49(3): 80-89.

[8] 普丽娜 , 殷晓 , 谢文娴 . 上海科学数据管理和共享需求分析及对策 [J]. 情报工程 , 2021, 7(6): 88-100.

[9] 支凤稳 , 张萌 , 赵梦凡 , 等 . 双路径视角下科学数据共享行为的影响因素研究 [J]. 信息资源管理学报 , 2021, 11(6): 40-50.

[10] 张新凤 . 区块链视域下医学图书馆科学数据共享机制研究 [J]. 图书馆工作与研究 , 2022, 44(9): 13-18+28.

[11] 陆丽娜 , 尹居峰 , 于啸 , 等 . 基于联盟链的农业科学数据共享模型构建研究 [J]. 图书情报工作 , 2022, 66(17): 1-9.

[12] PRIEGO LP, WAREHAM J, ROMASANTA AKS. The puzzle of sharing scientific data[J].Industry and Innovation, 2022, 29(2): 219-250.

[13] DEVRIENDT T, SHABANI M, LEKADIR K, et al. Data sharing platforms: instruments to inform and shape science policy on data sharing?[J]. Scientometrics, 2022, 127(6): 3007-3019.

[14] FIIALKA S, FIGOL N, FISENKO T, et al. Sharing unreviewed research data: Problems and prospects[J]. Amazonia Investiga, 2022, 11(55): 40-49.

[15] 马慧萍 . 2010—2019 年国内图书馆科学数据共享

doi:10.3772/j.issn.2095-915x.2023.04.008

研究综述 [J]. 图书馆学研究 , 2020, 42(8): 19-26.

[16] 白云朴 , 李果 . 科学数据共享研究的演化路径分析 [J]. 情报杂志 , 2022, 41(8): 138-148.

[17] 孙雨潇 , 李艳丽 , 李峰 , 等 . 国内外科学数据共享现状研究与发展建议 [J]. 农业大数据学报 , 2022, 4(2): 88-98.

[18] 支凤稳 , 张萌 , 郑彦宁 . 科学数据共享研究的多视角分析与整合框架构建 [J]. 中国科技资源导刊 , 2023(2): 1-9, 33.

[19] 汪雪锋 , 于苗苗 , 韦华楠 , 等 . 中国学者在顶级期刊发文的历史变迁与特征演化 [J]. 情报工程 , 2021, 7(3): 3-17.

[20] 温芳芳 . 国外科学数据开放共享政策研究 [J]. 图书馆学研究 , 2017, 404(9): 91-101.

[21] PILAT, DIRK AND YUKIKO FUKASAKU. OECD Principles and Guidelines for Access to Research Data from Public Funding[J]. Data Sci. J., 2007(6): 4-11.

[22] 刘玉琴 , 刘晶 , 张勇斌 . 中国图书情报领域专利研究的计量分析 [J]. 情报工程 , 2018, 4(6): 87-97.

[23] 魏玉梅 , 滕广青 . 网络视域下领域重要关键词提取方法的比较研究 [J]. 情报资料工作 , 2020, 41(3): 97-104.

[24] 范少萍 , 李迎迎 , 张志强 . 国内外共词分析研究的文献计量分析 [J]. 情报杂志 , 2013, 32(9): 104-109.

[25] 何世伟 , 葛慧丽 , 严伟 , 等 . 浙江省实施创新券政策推动科技资源开放共享的实证研究 —— 以科学仪器设备为例 [J]. 中国科技资源导刊 , 2019, 51(3): 24-28.

[26] 陈异凡 , 闫燊 , 杨亚超 , 等 . 我国农业科学数据共享协议 [J]. 大数据 , 2022, 8(1): 46-59.

[27] 范国梅 , 孙清岚 , 史文聿 , 等 . 国家微生物科学数据中心数据资源服务与应用 [J]. 微生物学报 , 2021, 61(12): 3761-3773.

[28] 张新兴 . 基于云计算的科学数据资源聚合系统研究 [J]. 图书馆学研究 , 2017, 39(21): 60-64+101.

[29] 盛小平 , 宋大成 . 数据管理与数据治理的比较分析及其对制定科学数据开放共享政策的启示 [J]. 图书情报工作 , 2020, 64(22): 4-10.

[30] GUDI N, KAMATH P, CHAKRABORTY T, et al. Regulatory Frameworks for Clinical Trial Data Sharing: Scoping Review[J]. Journal of Medical Internet Research, 2022, 24(5): 1-16.

[31] MAIDER ST, NURIA BP, OLATZ A, et al. Tracking Openness and Topic Evolution of COVID-19 Publications January 2020-March 2021: Comprehensive Bibliometric and Topic Modeling Analysis[J]. Journal of Medical Internet Research, 2022, 24(10): 1-15.

[32] COELHO R, BRAGA R, DAVID J M N, et al. A Blockchain-Based Architecture for Trust in Collaborative Scientific Experimentation[J]. Journal of Grid Computing, 2022, 20(35): 1-31.

[33] 刘晶 . 中国高校科技成果研究的文献计量学分析 [J]. 情报工程 , 2019, 5(6): 98-108.

[34] 陆冬云 , 张和珍 , 何险峰 , 等 . 科学数据库建设框架 — 统一发布及数据共享方案 [J]. 计算机与应用化学 , 2004, 21(1): 103-106.

[35] 王卷乐 , 王玉洁 , 张敏 , 等 . 2020 年地球数据科学与共享热点回眸 [J]. 科技导报 , 2021, 39(1): 105-114.

[36] GARDNER W. Compelled Disclosure of Scientific Research Data[J]. The Information Society, 2004, 20(2): 141-146.

[37] CHRISTINE F N. Enabling Public Data Sharing: Encouraging Scientific Discovery and Education[J]. Methods in Molecular Biology, 2009, 569(10): 25-32.

[38] TOGA AW, DINOV ID. Sharing big biomedical data[J]. Journal of Big Data, 2015, 2(7): 1-12.

[39] BROES S, LACOMBE D, VERLINDEN M, et al. Sharing human samples and patient data: Opening Pandora's box[J]. Journal of Cancer Policy, 2017,

12(10): 65-69.

[40] LI R, WOOD J, BASKARAN A, et al. Timely access to trial data in the context of a pandemic: the time is now[J]. BMJ Open, 2020, 10(10): 1-5.

[41] RIOS RS, ZHENG KI, ZHENG MH. Data sharing during COVID-19 pandemic: what to take away[J]. Expert Review of Gastroenterology & Hepatology, 2020, 14(12): 1125-1130.

[42] DRON L, DILLMAN A, ZORATTI MJ, et al. Clinical Trial Data Sharing for COVID-19–Related Research[J]. Journal of Medical Internet Research, 2021, 23(3): 1-5.

[43] 盛小平 , 王毅 . 利益相关者在科学数据开放共享中的责任与作用 —— 基于国际组织科学数据开放共享政策的分析 [J]. 图书情报工作 , 2019, 63(17): 31-39.

[44] 杨晶 , 康琪 , 李哲 . 推动科学数据开放共享的思考及启示 [J]. 全球科技经济瞭望 , 2019, 34(10): 37-43.

[45] 江慧慧 , 赵丽梅 . 科学数据共享障碍及消解措施分析 [J]. 图书馆研究 , 2022, 52(3): 36-42.

[46] 华小琴 , 司莉 , 李亭 . 我国科学数据共享中障碍因素分析及其启示 [J]. 图书馆工作与研究 , 2019, 41(11): 18-26.

[47] 盛小平 , 袁圆 . 国内外科学数据开放共享影响因素研究综述 [J]. 情报理论与实践 , 2021, 44(8): 173-179.

[48] 屈宝强 , 彭洁 , 刘蔚 , 等 . 科学数据共享及其发展趋势 [J]. 情报学进展 , 2020(13): 381-420.