

对大数据背景下科学数据开放共享的研究与思考

王瑞丹 高孟绪 石 蕾 王 超 徐 波
(国家科技基础条件平台中心, 北京 100038)

摘要: 科学数据是支撑国家科技创新和经济社会发展的重要基础性、战略性资源, 在大数据时代, 推动科学数据开放共享对于加强科学数据利用、提升科学数据应用价值具有重要意义。在分析我国科学数据开放共享现状的基础上, 结合典型科学数据领域的资源情况, 总结大数据时代科学数据的特征, 针对在大数据背景下加强科学数据开放共享的问题, 从数据汇交、数据标识和数据出版等方面提出了思考与建议。

关键词: 科学数据; 开放共享; 大数据; 数据汇交; 数据出版

中图分类号: G311

文献标识码: A

DOI: 10.3772/j.issn.1674-1544.2020.01.001

Research and Thoughts on the Opening and Sharing of Scientific Data under Background of Big Data

WANG Ruidan, GAO Mengxu, SHI Lei, WANG Chao, XU Bo

(National Science and Technology Infrastructure Center, Beijing 100038)

Abstract: Scientific data is an important basic and strategic resource to support national scientific and technological innovation as well as economic and social development. In the era of big data, promoting the opening and sharing of scientific data is of great significance to strengthen the utilization of scientific data and enhance the application value of scientific data. This paper analyzes the current situation of the opening and sharing of scientific data in China, summarizes its characteristics in the era of big data combined with the resources in the field of typical scientific data, and puts forward some thoughts and suggestions on how to strengthen the opening and sharing of scientific data under the background of big data from the aspects of data collection, data identification and data publishing.

Keywords: scientific data, open and share, big data, data collection, data publication

在大数据时代, 数据密集型的科学研究已成为新的科研范式。科学数据作为大数据时代最基本、最活跃的一类科技资源, 具有客观性、多结构性、分散性、时效性、共享性、易传递性和再创造性等多种特点^[1], 贯穿于科技创新活动的全过程, 其在科学研究过程中的重要作用日益凸

显, 是科技创新的重要产出和动力源泉^[2-3]。当今许多前沿科学领域取得的重大突破和发现越来越依赖于对海量科学数据的全面收集、深入分析挖掘与利用, 科学研究水平也越来越取决于对科学数据的积累, 以及将数据转换为信息和知识的能力^[4-5], 在这一过程中对科学数据开放共享的

作者简介: 王瑞丹 (1965—) 女, 国家科技基础条件平台中心研究员, 研究方向: 科技资源管理; 高孟绪 (1982—), 男, 国家科技基础条件平台中心副研究员, 博士, 研究方向: 科技资源管理与共享 (通信作者); 石蕾 (1982—), 女, 国家科技基础条件平台中心研究员, 研究方向: 科技资源管理; 王超 (1987—), 男, 国家科技基础条件平台中心助理研究员, 博士, 研究方向: 科技资源管理与共享; 徐波 (1988—), 男, 国家科技基础条件平台中心助理研究员, 研究方向: 科技资源管理。

收稿日期: 2020 年 1 月 12 日。

需求也越来越高。大数据背景下,推动科学数据开放共享,既有利于提升科学数据的质量,促进海量科学数据的持续、有序积累,也有利于避免产生相同科学数据的重复性劳动,促进海量科学数据的流通与有效利用。鉴于此,本文将重点研究探讨大数据背景下科学数据的开放共享的相关问题,对进一步推进我国科学数据开放共享提出相应的思考与建议。

1 我国科学数据开放共享的现状

自2001年我国正式启动“科学数据共享工程”项目以来,先后在气象领域开展了气象数据共享试点,启动实施了国家科技基础条件平台建设专项。经过近年来的发展,我国在科学数据共享与管理领域已取得巨大进步,各项工作也逐渐与国际接轨^[1]。从科学数据开放共享相关制度的完善,到科学数据的汇交、管理,我国科学数据开放共享的生态环境有了很大的发展,科学数据开放共享的成效不断显现。

1.1 科学数据开放共享的相关政策制度逐步完善

为推动科学数据开放共享,我国高度重视科学数据管理和开放共享,制定发布了相关的政策和标准,进一步完善了科学数据开放共享的政策法规体系。2007年修订的《中华人民共和国科学技术进步法》指出,利用国家财政资金设立的科研机构应当建立共享机制,促进科技资源的有效利用。科技部发布的《国家重点基础研究发展计划(973计划)管理办法》要求项目承担单位建立规范、健全的项目和成果数据库。此外,还有科技部颁布的《国家高技术研究发展计划(863计划)管理办法》、自然科学基金委员会颁布的《国家自然科学基金管理条例》等也对相关科研成果的开放共享提出了明确的要求。这些政策的制定和出台在一定程度上推动了科研项目产生科学数据的开放共享^[5]。

2018年3月,国务院办公厅印发了《科学数据管理办法》,这是我国首个在国家层面出台的 science 数据管理办法,加强和规范了科学数据管理、安全保障和开放共享,为科学数据工作提

供了行动纲领。该管理办法强调了科学数据的共享利用,尤其是明确提出由政府预算资金资助形成的科学数据应当按照“开放为常态、不开放为例外”的原则,及时面向社会和相关部门进行开放^[6]。

科学数据的标准化建设也取得了重要的进展。截至2019年年底,依托全国科技平台标准化技术委员会(SAC/TC486)已立项26项国家标准,发布实施了《科技资源标识》等14项国家标准,初步建立了我国科技平台标准体系。此外,全国信息技术标委会等也制定发布了《信息技术 科学数据引用》《信息技术 数据质量评价指标》等多个数据相关标准。这些国家标准的制定实施,规范了包括科学数据在内的科技资源管理,为促进科学数据的有效共享利用提供了重要标准。

1.2 科学数据汇交逐步成为数据积累的重要途径

科学数据汇交作为科研项目管理的重要内容,科技部、自然科学基金委员会、中国科学院等多个部门通过多方探索已经汇交整合了一大批科学数据资源。2008年科技部出台了《国家重点基础研究发展计划(973计划)资源环境领域项目数据汇交暂行办法》,明确了数据汇交的标准规范和工作流程,并持续开展了973计划资源环境领域相关项目的科学数据汇交工作,目前已汇交了科学数据集近2000个,整合各类型数据资源量超过2TB。1999年开始启动了科技基础性工作专项,科学数据是该专项支持产生的最主要成果^[7]。截至2019年年底,已完成科学数据汇交项目331个(占需汇交项目的81.3%),汇交科学数据总量达53TB。自然科学基金委员会建设了“科学基金共享服务网”,汇聚项目产生的科学数据和相关信息,目前已涵盖国家自然科学基金委员会8个学部的面上项目、重点项目、重大项目、青年项目等产出的成果信息,并面向科研用户开放。中国科学院积极开展科研信息化专项,通过院内科学数据的汇交整合与管理,目前已经整合了全院50家单位的科学数据库,共享数据量达3.2PB,约是之前数据量的5倍,涉及地学、

生物、材料、化学、物理等学科领域。

1.3 科学数据中心成为推动科学数据开放共享的重要载体

科学数据中心是科学数据的汇集、管理、开放共享和保存的重要基础设施，建设国家科学数据中心是科学数据管理的重要手段^[8-9]。为促进科学数据的汇交整合和开放共享，我国加强了国家科学数据中心的建设。自2004年起，科技部、财政部联合启动国家科技基础条件平台建设专项，支持在地球系统、人口健康、农业、林业、气象、海洋等8个领域建成了国家科技资源共享服务平台，初步形成了一批具有一定资源优势的科学数据中心。2019年6月，科技部、财政部在原有科学数据类国家平台的基础上进一步优化调整形成了20个国家科学数据中心^[10]，作为促进科学数据开放共享的重要载体，并强调对于科学数据的开发应用与分析挖掘利用。同时，我国一直积极参与国际科学理事会（ISC）下属的世界数据系统（WDS）和国际数据委员会（CODATA）的相关工作，积极组织国内各有关部门、研究机构参加WDS和CODATA的各类学术活动，推动和协调各学科领域的科学数据工作。截至目前，在WDS现有的81个数据中心正式成员中，我国数据中心已有9个^[11]，有力地提升了我国相关领域科学数据的开放共享与利用水平。

1.4 科学数据开放共享取得显著成效

目前，我国科学数据的开放共享在服务科研活动、推动科研创新方面的成效显著，促进了科学研究的蓬勃发展。如国家地球系统科学数据中心通过建成的跨多学科、多时空尺度的地球系统科学数据库群，覆盖学科领域同类资源约75%，且全部对外开放共享，无偿为8000多项国家重大科研项目、国家重大工程等提供数据服务约18PB，累计节约国家经费上亿元。此外，围绕“一带一路”“生态文明建设”“区域协同发展”“精准脱贫”“创业创新”等国家战略需求、经济及社会热点应用等，通过推动不同学科和领域科学数据的开放共享，为国家战略决策、科学

研究和工程建设提供了有力的数据支撑。如国家气象科学数据中心通过推动京津冀地区气象与环境、农业等行业融合，产生的直接或间接经济效益超过1.3亿元，并为冬奥会筹备等提供了支撑保障。同时，科学数据在自然灾害防护、边检防疫、医疗健康、农业发展等社会民生服务方面作出了应有的贡献。如在安徽省阜南县利用医疗数据资源开展的远程医疗试点，自2017年以来，该县向外转诊率下降了10.7%，当地确诊入院治疗率增加了24%^[11]。通过科学数据的开放共享既节约了大量相关资源，也提高了科学数据的利用效率。

在大数据背景下，我国科学数据开放共享工作取得显著进展和成效，但仍然存在一些问题和不足^[12]，主要体现在以下几个方面：一是科学数据汇交与共享的理念仍需加强，科学数据开放共享的生态还有待改善，集中统一汇聚保存大量的科学数据缺乏合理的机制和渠道，这在一定程度上制约了我国科学数据的规范化持续积累；二是已经开放共享科学数据的质量参差不齐，直接影响了科学数据的有效利用程度和水平，在一定程度上阻碍了对这些科学数据的深入分析与挖掘利用，科学数据质量的规范化评价还有待增强；三是对科学数据开放共享中的知识产权缺乏切实可行的有效保护措施、创新形式和激励机制，影响了科学数据生产者开展科学数据开放共享的积极性；四是针对开放共享科学数据的有效分析和深度挖掘能力不足，尤其是新技术新手段的应用。

2 大数据时代科学数据的特征与开放共享

大数据时代，随着多源、异构海量科学数据的持续产生、积累，催生了对海量数据管理、分析与挖掘的巨大需求，数据密集型科学也由传统的基于假设驱动向基于数据进行科学探索的转变。在大数据背景下，需要强化对科学数据的质量要求，确保开放共享科学数据的规范性、准确性、真实性和时效性，保证科学数据的可发现、可访问、可互操作和可重用，促进不同类型、学科领域科学数据的交叉融合与综合应用。近年

来,生态系统、对地观测、生命与健康、天文观测等学科领域通过持续的观测、监测等方式产生并积累了海量科学数据,对数据的开放共享与分析挖掘的需求更为迫切^[13-14]。

面对不同领域的海量科学数据,只有打破数据封闭的传统模式,形成开放共享的数据思维方式,才有可能有效解决大数据背景下的诸多科学问题,真正发挥大数据的价值。以生态系统为例,长期以来,我国在生态系统的结构、功能等方面进行了持续观测、实验和科学研究,积累了大量的生态系统相关科学数据。在大数据背景下,新一代生态观测体系呈现了以物联网技术为支撑的多尺度、多要素、多过程等重要特征。同时,长期定位观测数据、相关遥感数据以及实验模拟数据等的快速增长,有力地促进了生态数据分析挖掘技术的发展,开展多维、多源的生态数据整合挖掘、分析模拟与预测成为大数据时代生态学研究的重要手段。而通过深入应用数据众包、数据出版等新颖的数据获取、处理与知识产权保护方式,也有助于推动生态领域科学数据的高效开放共享与分析利用^[13]。

大数据时代的科学数据开放共享离不开数据处理、分析挖掘与信息共享技术的强有力支持。如在对地观测领域里,遥感技术、通讯技术及计算机技术的快速发展,带来了对地观测领域科学数据的爆炸式增长,尤其是高分辨率对地观测技术的日益发展,通过机载和卫星传感器等多种途径获取的遥感数据量呈指数递增,推动对地观测领域进入大数据时代。对地观测领域科学数据应用范围的不断扩大,促使其数据共享方式从单纯的数据共享转变为服务共享、信息共享,并逐渐步入以数据交叉融合、深度挖掘与综合应用为主要特征的智能化阶段。这一数据共享方式的变化,要求提高对地观测数据的数据流通与分析挖掘效率,打通从对地观测数据接收、分析处理到信息提取、共享分发的全流程,而大数据时代云计算、人工智能以及物联网等技术的发展也有力促进了对地观测数据的开放共享^[14]。

3 加强科学数据开放共享的思考与建议

在大数据背景下,开放共享具有价值多样性的科学数据能够为更多的科学研究提供丰富的基础研究资源,扩展科学研究的范围和视角。充分地开放共享和重复使用科学数据不仅能节省数据采集生产与获取、加工处理等方面的投入,加快科学发现的进程,而且在保持科学研究结果的可重复性、保持科学事业自我修正能力等方面具有不可替代的作用。长期以来,我国已经开展了大量工作积极推进科学数据开放共享并取得成效。为了持续推动科学数据的开放共享,提出以下几点思考与建议。

(1) 加强科技计划项目数据的规范化汇交。科技计划项目形成的科学数据,应按国家科学数据汇交政策的要求,以及各领域、行业、部门细化落实的实施办法,按时、保质、保量地汇交到管理方,确保数据的及时性、真实性和可靠性。同时,按照汇交要求的标准进行规范加工处理,确保数据的可操作和可重复利用。科学数据汇交的内容既要包括科技计划项目执行过程中产生的科学数据实体以及对科学数据的描述信息,也要包括科学数据辅助工具软件等,以保证数据能够有效处理、加工和分析。此外,还要优先汇交国家重点研发计划项目产生的数据。同时,要推动并鼓励其他各级各类科技计划(专项、基金等)项目数据的汇交工作。为加强科学数据汇交工作,还要建立健全科学数据汇交责任单位和责任人的绩效管理和评价机制,制定奖惩办法,充分调动法人单位和科研人员的积极性。要充分发挥国家科学数据中心在大数据时代的重要作用,持续提升国家科学数据中心的能力和水平,完善国家层面对科学数据的存储管理和服务保障体系,强化科学数据汇交保障能力。

(2) 推动科学数据质量评价。科学数据质量的评价是科学数据共享管理的重要内容之一,既是科学数据质量保障的关键环节,也是科学数据质量提升的重要保障手段。科学数据质量的评价主要包括数据生产、管理、传播以及再利用全过程的评价。要加强科学数据质量评价新标准新方

法研究，制定通用的科学数据质量评价指标体系，提升科学数据质量评价的标准化与规范化程度。同时还要积极探索开放式同行评审和先出版后评审等新的评价方法，鼓励科学数据中心和数据出版机构积极开展探索实践，推动新型科学数据质量评价方法发展，以更好地满足大数据背景下科学数据开放共享的新需求。

(3) 加快推进科学数据资源标识工作。近年来，标识制度广泛应用于期刊出版、物联网等实体资源管理领域，并呈现出向科学数据等数字资源快速延伸的趋势。我国已于2017年发布实施了国家标准《科技资源标识》，为开展和完善科技资源标识体系建设奠定了重要基础，对加强科学数据管理、保护知识产权并推动数据开放共享具有重要作用。对科学数据进行标识相当于为每一个科学数据赋予独一无二的身份证、标识符。科学数据的标识将有助于实现科学数据资源的永久可定位、可追溯、可引用、可统计与可评价，促进科学数据开放共享，也有助于规范科学数据引用，保障科学数据生产者权益，构建科技资源标识为核心的科学数据管理与开放共享生态系统。在已有科技资源标识国家标准的基础上，要建设我国自主知识产权的标识体系标准规范，支持跨学科、跨层级、跨系统的科技资源标识，并积极推动在国家科学数据中心先期试点建设，逐步实现与国际相关标识体系的对接，通过提供高水平、专业化的数据标识服务提升科学数据开放共享的水平。

(4) 积极推动科学数据出版。科学数据出版是大数据时代开展科学数据开放共享的一种创新模式，符合新时代数据密集型科学研究范式的特征。近年来，《全球变化数据学报》和《中国科学数据》等数据期刊的创办与实践，通过发表数据论文的方式，为科学数据共享积累了大量经验。此外，数据存储库能够为数据集提供相对集中存储和访问的平台，并支持数据质量的标准化控制和完整的全生命周期管理，受到科研人员的广泛关注。而在大数据背景下数据管理与开放共享的关键是存储库的可信性和可持续性。针对科学数据出版的新需求和新模式，要不断强化国家

科学数据中心对科学数据资源的汇聚集能力，推动其成为重要的科学数据存储库。还要借鉴国际数据仓储库的做法，探索科学数据出版的新模式、新方法，通过采用联合数据存缴政策等与知名期刊合作，共同提升数据出版的质量，提升科学数据开放共享的效率和价值。

(5) 探索科学数据开放共享新技术、新手段的应用。在大数据时代，在科学数据开放共享的过程中，必然也将伴随出现数据权属、质量、安全等问题。而这些问题成为制约科学数据流通的重要瓶颈，如知识产权保护对开展科学数据共享积极性的影响、异构海量科学数据传输处理的安全性等。区块链、安全多方计算等新技术的应用，有助于促进数据共享流通过程中的可追溯、可监控，从而减少了知识产权可能遭遇侵害的风险。其中，相比传统中心化的数据共享模式，基于区块链技术的数据共享，通过分布式账本、数据隐私安全、数据确权等机制有助于解决不同数据机构间的数据共享，如医疗数据已开始尝试利用区块链存储、管理和共享个人健康数据。此外，分布式边缘计算技术等将有可能发挥更大作用，通过区块链一体化能够实现数据的快速采集、处理和分析^[15-16]，从而大幅提升科学数据开放共享的效率。

4 结语

本文在总结科学数据开放共享现状的基础上，结合典型领域的资源情况，探讨了大数据时代背景下科学数据共享开放。笔者认为，在大数据背景下，加强科学数据的开放共享、推进科学数据的综合利用既是大数据时代科学数据工作面临的重要机遇，也是一种重要挑战。长期以来，我国已经开展了大量工作积极推进科学数据开放共享并取得成效。在大数据时代，要继续借鉴发达国家在科学数据开放共享方面的成功经验和做法，进一步深化科学数据的管理与开放共享。要持续加强科学数据管理，从规范科学数据汇交管理、提升科学数据质量、推动科学数据标识和出版以及加强新技术应用等方面多管齐下，进一步

(下转第26页)

- 况的研究[J]. 高校实验室工作研究, 2002(4): 62-65.
- [8] 黄新华. 高校“211工程”与设备管理探讨[J]. 中国现代医学杂志, 2002(9): 109-111.
- [9] 张子石. 基于CiteSpace的网络学习知识图谱分析[J]. 中国电化教育, 2015(8): 77-84.
- [10] 段春雨, 蔡建东. 国际泛在学习领域知识图谱研究[J]. 现代远程教育研究, 2016(1): 85-95.
- [11] 李宜祥, 张竞宇. 加强教育引导营造仪器设备共享平台的良好氛围[J]. 实验室研究与探索, 2012, 31(12): 215-218.
- [12] 赵玉茹, 冯建跃, 赵月琴. 浙江省高校大型仪器设备开放共享现状调研与思考[J]. 实验技术与管理, 2017, 34(10): 255-258, 266.
- [13] 徐红岩, 曾令宇, 陆召军. 构建地方医药院校大型科研仪器开放共享平台的探索与实践[J]. 实验室研究与探索, 2017, 36(11): 252-254.
- [14] 王志峰, 胡茂志, 陈文飞, 等. 地方高校测试中心仪器设备开放共享探究[J]. 实验室研究与探索, 2016, 35(11): 269-272.

(上接第5页)

加强大数据时代我国科学数据的开放共享能力建设, 为提升我国科技创新水平、服务国民经济社会发展提供强有力的支撑。

参考文献

- [1] 国家科技基础条件平台中心. 国家科学数据资源发展报告: 2018[M]. 北京: 科学技术文献出版社, 2019: 6.
- [2] 孙九林, 林海. 地球系统研究与科学数据[M]. 北京: 科学出版社, 2009.
- [3] 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. DOI: 10.16418/j.issn.1000-3045.2018.08.001.
- [4] British Academy. Data management and use: Governance in the 21st century[R]. London: The Royal Society, 2017.
- [5] 王瑞丹, 杨静, 高孟绪, 等. 加强和规范我国科学数据管理的思考[J]. 中国科技资源导刊, 2018, 50(2): 1-5. DOI: 10.3772/j.issn.1674-1544.2018.02.001.
- [6] 国务院办公厅关于印发科学数据管理办法的通知[EB/OL]. (2018-03-17)[2020-01-08]. http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm.
- [7] 杨杰, 宋佳, 诸云强, 等. 科技基础性工作专项数据汇交共享平台建设[J]. 中国科技资源导刊, 2017(5): 52-59. DOI: 10.3772/j.issn.1674-1544.2017.05.007.
- [8] 王卷乐, 王明明, 石蕾, 等. 科学数据管理态势及其对我国地球科学领域的启示[J]. 地球科学进展, 2019, 34(3): 306-315. DOI: 10.11867/j.issn.1001-8166.2019.03.0306.
- [9] 高孟绪, 王瑞丹, 王超, 等. 关于国家科学数据中心建设与发展的思考[J]. 农业大数据学报, 2019(3): 19-25. DOI: 10.19788/j.issn.2096-6369.190302.
- [10] 科技部财政部关于发布国家科技资源共享服务平台优化调整名单的通知[EB/OL]. (2019-06-05)[2020-01-08]. http://www.most.gov.cn/mostinfo/fo/xinxifenlei/fgzcgfxwj/gfxwj2019/201906/t20190610_147031.htm.
- [11] ICSU - World Data System Regular Members[EB/OL]. [2020-01-08]. <http://www.icsu-wds.org/community/membership/regular-members>.
- [12] 杨行, 屈宝强, 赫运涛, 等. 世界主要国家科学数据资源共享和管理的对比分析和启示[J]. 中国科技资源导刊, 2016, 48(6): 18-25. DOI: 10.3772/j.issn.1674-1544.2016.06.003.
- [13] 于贵瑞, 何洪林, 周玉科. 大数据背景下的生态系统观测与研究[J]. 中国科学院院刊, 2018, 33(8): 832-837. DOI: 10.16418/j.issn.1000-3045.2018.08.010.
- [14] 何国金, 王力哲, 马艳, 等. 对地观测大数据处理: 挑战与思考[J]. 科学通报, 2015, 60(5): 470-480. DOI: 10.1360/n972014-00907.
- [15] 张丽丽, 温亮明, 石蕾, 等. 国内外科学数据管理与开放共享的最新进展[J]. 中国科学院院刊, 2018, 33(8): 774-782. DOI: 10.16418/j.issn.1000-3045.2018.08.002.
- [16] 薛腾飞, 傅群超, 王枫, 等. 基于区块链的医疗数据共享模型研究[J]. 自动化学报, 2017, 43(9): 1555-1562. DOI: 10.16383/j.aas.2017.c160661.