

大数据时代加强科学数据管理的思考与对策

苏 靖

(国家科技基础条件平台中心,北京 100038)

摘 要:随着大数据时代的来临,科研范式发生了革命性变革,进入了以数据密集型为特征的第四范式。在这一背景下,加强科学数据管理具有重要意义。首先对大数据时代科学数据管理的新特征和新趋势进行梳理分析,再对我国科学数据管理现状进行刻画,最后从处理好数据治理与价值实现之间的关系、处理好公益性与市场化的关系、处理好开放共享与安全保密之间的关系、处理好国内共享与国际合作的关系 4 个方面对加强和规范我国科学数据管理提出对策和建议。

关键词:大数据时代;科学数据;数据管理

中图分类号:F062.5

文献标识码:A

文章编号:1005-0566(2022)09-0050-05

Consideration and suggestion on strengthening scientific data management in the era of big data

SU Jing

(National Science and Technology Infrastructure Center, Beijing 100038, China)

Abstract: With the advent of the big data era, the scientific research paradigm has undergone a revolutionary change, and has entered the fourth paradigm characterized by data intensiveness. In this context, it is of great significance to strengthen scientific data management. This paper summarizes and analyzes the new characteristics and new trends of scientific data management in the big data era, describes the current situation of scientific data management in China, countermeasures and suggestions for strengthening and standardizing the management of scientific data in China has given based on handling the relationship between data governance and value realization, public welfare and marketization, open sharing and confidentiality, domestic sharing and international cooperation.

Key words: big data era; scientific data; management

在科学发展的进程中,数据作为对科学研究的记录,一直是科学研究的重要基础。大数据时代,对大量数据的收集、管理以及将数据转换为信息和知识的能力在一定程度上决定了科学发现的价值,带来了以数据密集型为特征的科研范式重大变革^[1]。当前,我国正处于实施创新驱动发展战略和建设世界科技强国的关键时期,加强和规范对科学数据的管理是提升我国科技创新能力和保障国家安全的重要方式和手段。2018 年,国务

院办公厅^[2]印发《科学数据管理办法》,开创了科学数据管理新局面,有效促进了科学数据的管理与共享应用。本文试图对大数据时代我国科学数据管理现状进行梳理总结,并提出思考与对策建议。

一、大数据时代科学数据管理的新特征和新趋势

1. 海量数据快速积累,催生新的科研范式

源于越来越多大科学装置的建设 and 重大科学实验的开展,以及科学传感器和传感网络的广泛

收稿日期:2022-07-28 修回日期:2022-08-20

作者简介:苏靖(1968—),男,山西朔州人,国家科技基础条件平台中心主任,研究员,博士。

应用,源源不断产生的 EB 量级科学数据将科学研究快速推进到一个前所未有的大数据时代。数据的飞速积累增强了人类详细表征和描述复杂性科学问题的能力,并使分析高度耦合复杂系统的动态行为成为可能,进而催生了大批科学发现。随着大数据在各领域的广泛出现,科学研究的模式发生明显变化,继实验科学、理论科学、计算科学之后,出现了“数据密集型科学研究”的第四科研范式。这一范式强调数据作为科学方法的特征,不仅是新的科研方法,也是重要的思维转变^[3]。

2. 大数据技术快速发展带来新的技术挑战

大数据的爆炸式增长给数据从存储到分析的各方面技术都带来了挑战^[3]。在数据存储方面,传统的静态存储方案已不能应对数据的动态演化所带来的挑战,需要低成本、分布式、可横向扩展的存储架构,以满足持续增长的大量数据的存储需求^[4]。在数据传输方面,在大数据时代科研活动的巨大需求和网络技术创新双重动力的推动下,下一代互联网、光网络、网络虚拟化等技术迅速应用到科研网络中,满足科研活动中大数据传输的差异化需求,服务全球科技创新。在数据分析处理方面,基于分布式平台的大数据处理算法和低内存消耗的快速算法成为主要研究方向,多元异构数据的挖掘也成为重要挑战。

3. 科学数据已成为国家战略资源

科学数据作为一类重要的数据,已经成为国际科学研究和发展的战略性、基础性资源,成为国际竞争的重要领域。在大数据时代,一个国家的科学研究在国际上的地位,很大程度由其在科学数据积累方面的优势以及从数据获得信息和知识的能力决定^[5]。美国、欧洲等发达国家和地区高度重视大数据,已经将对数据的挖掘复用能力提高到了国家战略的高度,投入巨大的人力物力支撑其快速发展,通过立法确立国家科学数据管理基本原则,同时也将保障数据安全纳入国家安全体系。各国普遍将科学数据中心建设作为加强科学数据管理和促进数据应用的重要载体,并通过多种机制推动科学数据向数据中心汇聚^[6]。多年来,发达国家积极依靠权威科学数据中心,对全球科学数据资源进行整合和汇集,并逐渐形成通用的科学数据收集、管理和存储解决方案。

4. 开放获取和开放科学快速发展与科学数据管理相互促进

大数据时代,互联网的不断发展和数字技术的日益强大使得数据呈指数级增长,全球发展呈网络化趋势,科学交流不断扩大,科学社区也日渐全球化,科研生产出现了新的模式,促使开放科学不断发展。当前,气候、资源、健康、安全等问题已经成为全球性问题,国际社会相互依赖程度达到前所未有的高度。重大科学发现、尖端科技创新也越来越需要深度的国际合作。据统计,全球超过 1/3 的研究论文与国际合作直接相关^[7]。“开放获取”和“开放科学”为解决科技与人类社会发展的全球性问题和降低全球科研成本提供了可行的解决方案。2021 年,联合国教科文组织发布《开放科学建议书》将开放研究数据作为重要内容之一,开放科学迈入全球共识的新阶段,也影响着研究工作和科学组织的运作方式,数据共享、开放获取、研究透明的开放科学在全球范围内快速发展。

二、我国科学数据管理的新进展

1. 《科学数据管理办法》印发,开创科学数据管理新局面

2018 年,国务院办公厅印发《科学数据管理办法》(简称“《办法》”),这是我国第一个从国家层面出台的科学数据管理办法,为我国科学数据工作确立了行动纲领,在推动我国科学数据管理走向规范化、提升科学数据应用水平、促进科技创新和经济社会发展、提升国家安全支撑保障能力等方面具有重要意义。《办法》明确了科学数据参与方的职责分工,强化了科学数据所在法人单位的主体责任,提出按照“分级分类管理,确保安全可控”的原则,明确主管部门和法人单位依法确定科学数据的密级及开放条件。《办法》还对科学数据使用者和生产者的行为进行规范和约束,将严惩数据造假等行为。《办法》要求科技计划项目产生的科学数据进行强制性汇交,并通过科学数据中心进行规范管理和长期保存。同时,《办法》提出了法人单位要在岗位设置、绩效收入、职称评定等方面建立激励机制,将科学数据工作情况作为对有关人员考核内容的重要方面。

《办法》发布后,多个部门、省市先后发布了实施细则。实施细则既贯彻了国家总体要求又清晰

呈现出部门/地方科学数据管理的基础和特点,形成了科技主管部门牵头、责任部门承担、科学数据中心实施的多级管理架构,规范科学数据全生命周期管理,并高度重视科学数据安全^[8],部分省市立足需求细化管理体系,提出设立“科学数据管理委员会”对科学数据进行宏观管理和综合协调。

2.《国家科技资源共享服务平台管理办法》出台,优化整合形成首批 20 个国家科学数据中心

2018 年,科技部、财政部^[9]联合印发《国家科技资源共享服务平台管理办法》,明确国家科学数据中心是国家科技资源共享服务平台的一类,属于基础支撑与条件保障类国家科技创新基地。2019 年,在原有国家科技资源共享服务平台工作基础上,科技部、财政部共同组织开展优化整合,形成 20 个国家科学数据中心,主要分布在地学、生命科学、基础科学等领域。科学数据中心主要承担各自领域科学数据的整合汇交,负责对本领域科学数据进行分级分类、加工整理和分析挖掘,在保障科学数据安全的前提下,依法依规推动科学数据开放共享,并主动加强与国内外科学数据有关方面交流与合作。截至 2021 年年底,各领域国家科学数据中心累计整合科学数据资源总量超过 70 PB,建成了若干具有较强国际影响力的科学数据库。多个省分别设置科学数据中心,并多按层级结构划分,如“省级总中心—领域行业中心—市级中心”三级数据中心体系,一些地方提出建设科研机构科学数据中心的构想,目前尚在进一步探索中。

3. 加大数据资源整合力度,重点推动科技计划项目数据汇交

按照《科学数据管理办法》要求,“政府预算资金资助的各级科技计划(专项、基金等)项目所形成的科学数据,应由项目牵头单位汇交到相关科学数据中心,各级计划(专项、基金等)项目管理部门应建立先汇交科学数据、再验收科技计划(专项、基金等)项目的机制”。科技部、国家自然科学基金委员会、中国科学院等多个部门加大数据资源整合力度,重点推动科技计划项目数据汇交。2018 年,科技部印发《国家重点研发计划项目综合绩效评价工作规范(试行)》,明确要在综合绩效评价过程中完成科学数据汇交。2021 年颁布的《科技计划形成的科学数据汇交 技术与管理规范》等

3 项国家标准,成为开展科技计划项目科学数据汇交重要的参考标准。至目前,已有超过 3 000 个国家重点研发计划项目向国家科学数据中心进行数据汇交,形成各类科学数据库(集)5 万余个,汇交数据总量超过 4 PB。目前科学数据汇交作为科研项目管理的重要内容,已经成为积累我国科技创新活动一手数据的重要渠道,将有力地支撑新一轮创新研发活动。

4. 建立健全科学数据标准体系,有序推进标准化工作

长期以来,我国多渠道推动科学数据标准规范的研究制定工作,各部门、科研机构和各领域标准化技术委员会组织研究制定了多项与科学数据相关的标准规范。全国科技平台标准化技术委员会一直大力推动科学数据标准制修订工作,组织研究提出了科技平台标准体系框架,并开展了包括科学数据相关标准在内的一系列国家标准制修订工作。国家科技基础平台中心组织研究制定了《科技平台 资源核心元数据》《科技平台 服务核心元数据》《科技资源标识》《科技计划形成的科学数据汇交技术与管理规范》等 20 余项国家标准,为进一步规范化开展科学数据管理与应用提供的重要支撑。

5. 推动开展科学数据标识和科学数据出版工作

推动科学数据标识体系建设具有重要意义,既可保障我国科学数据等资源安全,也可支撑知识产权保护,促进国内科学数据开放,同时也是提升我国科学数据资源应用服务能力的重要支撑,构建基于标识的科技资源生态系统有助于推动数据关联分析、知识图谱等科技服务,可有效促进科技资源互联互通与综合应用。目前,《GB/T 32843-2016 科技资源标识》国家标准已由国家标准化管理委员会正式发布实施,落地细则正在制定中,根据国家标准建设的国家科技资源标识体系正在有序推进之中。目前,通过中国科技资源共享网科技资源标识系统已注册各类科技资源 280 余万项。

在科学数据出版方面,国内数据出版和数据传播工作处于快速发展阶段^[10],其中科学数据期刊建设发展尤其迅猛,创办了《中国科学数据》《全球变化数据学报》科学数据出版期刊。2018 年,《全球变化数据学报》的“全球变化科学研究数据

出版系统”获得联合国信息社会世界峰会奖(电子科学组冠军奖)。此外,国内医学、地学、纳米科学领域的部分期刊专门设置了数据论文专刊,对相关科学数据集进行描述和发布。

三、加强我国科学数据管理的对策与建议

经过多年努力,我国科学数据管理和开放共享取得了长足发展,对科学数据管理的认识逐步加深,在推动科学数据开放共享、科学数据中心建设、科技计划科学数据汇交、科学数据标识及出版等方面已取得了初步成效,但也应该看到在科学数据管理方面还有很多工作需要深入开展,以便于大数据时代抢占发展先机,笔者认为应重点从以下4个方面关系入手,以进一步提高我国科学数据管理与应用水平。

1. 处理好数据治理与价值实现的关系

近年来,我国各类科学数据快速积累,成为我国科技创新的重要财富。加强需求牵引和场景驱动,聚焦国家重大科技创新需求和民生发展需要,促进科学数据资源建设与数字技术深度融合,实现海量数据资源优势 and 丰富应用场景的有效结合,赋能科研范式转变升级,是当前我国科学数据治理的主要目标,也是科学数据支撑价值发现的重要方向。尤其在大数据背景下,不能就数据治理谈数据治理,就数据应用谈数据应用,必须实现两者间的充分对接与融合发展。

在科学数据治理方面,首先要聚焦科学数据安全、开放应用、服务评价等关键问题,加强科学数据管理制度体系与标准规范体系研究。分领域组织完善科学数据安全管理、数据分级分类管理、科学数据质量控制、科研项目科学数据管理等技术规范。要充分利用大数据、区块链、人工智能等技术赋能科学数据高效治理,支持智能化科学数据应用服务平台建设和软件工具研发,提升科学数据在线提交与治理支撑能力。

在科学数据支撑价值实现方面,要加强科学数据应用服务的需求牵引和场景驱动,充分发挥科学数据中心及相关科研机构海量科学数据优势,促进数据资源和数据技术深度融合,增强数据支撑保障能力。聚焦新冠病毒变异预测预警、全球变化与应对研究、生态系统碳汇能力分析与提升、重大设施装备安全等重大科学问题、战略前沿领域和国家安全

保障等方面的典型应用场景,组织跨科学数据中心的多元数据融合分析与数据资源产品建设。

2. 处理好公益性与市场化的关系

2020年3月,中共中央、国务院^[11]发布《关于构建更加完善的要素市场化配置体制机制的意见》(简称“《意见》”),数据首次被正式纳入生产要素范围,与土地、劳动力、资本、技术并列作为五类要素之一。《意见》明确提出要“加快培育数据要素市场”,推进政府数据开放共享,提升社会数据资源价值,加强数据资源整合和安全保护。科学数据作为一类重要的数据,对科技创新和经济社会发展都具有非常重要的意义,处理科学数据管理与应用中公益性和市场化关系,要把握好以下3点。

一要建立健康、有序的科学数据资源市场运行基本准则。建立用于规范科学数据资源市场活动的如交易定价、交易行为的准则,建立符合科学数据特点的数据使用权而非所有权的交易机制。建立良性的市场竞争机制,打造科学数据产品与服务品牌。建立包括数据生产服务机构内部控制、政府监管、行业自律、社会监督在内的科学数据资源市场化运行的监督体系。

二要支持对科学数据确权。建立健全涉及科学数据产权界定、开放共享、隐私保护、数据安全等方面可操作性的法律法规。通过数据标识、数据出版等配套措施,明确数据生产者、管理者、服务者等各方在数据工作中的贡献。鼓励科研人员对科学数据进行分析挖掘,形成有价值、可推广的科学数据产品,在依法加强安全保障和隐私保护的前提下,提升科学数据作为一类数据要素的作用。

三要在确保科学数据公益服务的同时完善市场化增值收费机制。加强科学数据的分级分类管理,对于政府决策、公共安全、国防建设、环境保护、防灾减灾、公益性科学研究等涉及国家和公共利益需要的,法人单位应当无偿提供;确需收费的,应按照规定程序和非营利原则制定合理的收费标准,向社会公布并接受监督。

3. 处理好开放共享与安全保密之间的关系

开放共享与安全保密是科学数据管理的两个方面,只有正确处理二者之间的关系,才能最大程度发挥科学数据的效益。科学数据管理应严格按照“分级分类管理,确保安全可控”的基本原则,建立

完善数据分级分类制度,在保证数据安全并合法合规的前提下,根据数据的重要性和敏感程度确定共享范围、权限和方式,促进数据开放共享和利用^[12]。

在数据管理过程中,应建立规范的数据接收、审核和等级注册机制,数据提供方应确保数据真实、完整、安全、有效、可用,来源明确,界限清晰。

在数据共享过程中,应把数据交换过程中的安全管理要求放在首位,加强数据分级分类管理,突出重点数据管护,并建立数据安全应急处置预案。对于涉及国家秘密的科学数据的全生命周期管理都要严格按照国家有关保密规定执行。对于涉及个人隐私或敏感信息的科学数据,应对数据进行去隐私、脱敏等处理后方可对外共享。

4. 处理好国内共享与国际合作的关系

大数据时代,世界各国信息交流愈发频繁,数据在国家之间流动传递成为关系各国政治、经济、社会的核心议题^[13],科学数据作为一类重要的数据,也与国家发展息息相关,如何在安全和发展中取得平衡,是各国政府在数据管理和开放应用过程中重点关注的方面^[14]。有必要在大数据发展的大背景下,在确保数据安全的前提下实现自由流动,提出制定和参与国际规则的策略方案,构建我国科学数据资源战略和数据治理能力。

科学数据的国际合作既是合作问题也是安全问题,《科学数据管理办法》也明确提出科学数据开放共享必须要以安全可控为前提,强化国家关键数据资源保护能力。要按照国家有关法律法规要求,依法确定科学数据安全等级及开放条件,严格做好科学数据保密工作,建立数据共享和对外交流的安全审查机制。

在确保数据安全的基础上,我国也应该按照科学数据特点探索推动建立科学数据跨境流动新机制。一是分级分类建立完善数据跨境流动监管制度,根据数据风险加强分级分类管理,建立灵活多样的管理制度体系。二是建立科学数据在跨境流动过程中的互信合作机制^[15],建立数据流动的协议和标准,通过开展双边合作、多边合作等方式促进数据互联互通。三是探索通过市场化手段解决科学数据跨境流动中的问题。充分挖掘数据作为生产要素的价值,在数据安全法、个人信息保护法等相关制度的管理下,探索基于市场的科学数

据国际合作新机制。

四、结语

大数据时代,科学数据已经成为开展科学研究和创新发现重要的基础性战略资源,加强科学数据开放共享,保障科学数据安全,成为国际社会普遍共识。处理好科学数据管理中的突出问题和各方关系,加强国家科学数据中心能力建设,完善数据标准体系、规范科学数据标识等既是保障数据安全、促进数据共享、加强数据应用的重要基础和保障,也是推动科学数据支撑科技创新和经济社会发展的重要举措。

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 2016[M]. 北京:科学技术文献出版社,2016,2.
- [2] 国务院办公厅. 科学数据管理办法[Z]. 2018.
- [3] CODATA 中国全国委员会. 大数据时代的科研活动[M]. 北京:科学出版社,2014,7.
- [4] 蒋昌俊,丁志军,王俊丽,等. 面互联网金融行业的大数据资源服务平台[J]. 科学通报, 2014(36):3547-3553.
- [5] 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(3):306-315.
- [6] 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.
- [7] 陈传夫,李秋实. 数据开放获取使科学惠及更广——中国开放科学与科学数据开放获取的进展与前瞻[J]. 信息资源管理学报 2020,10(1):4-13.
- [8] 高瑜蔚,石蕾,朱艳华,等. 《科学数据管理办法》实施细则比较研究——以正式发布的 11 份细则为例[J]. 中国科技资源导刊,2019,21(3):1-10,17.
- [9] 科技部,财政部. 国家科技资源共享服务平台管理办法[Z]. 2018.
- [10] 秦顺,邢文明. 开放共享安全:我国科学数据共享进入新时代——对《科学数据管理办法》的解读[J]. 图书馆,2019(6):36-42.
- [11] 中共中央,国务院. 关于构建更加完善的要素市场化配置机制体制的意见[Z]. 2020.
- [12] 闫桂勋,刘蓓,程浩,等. 数据共享安全框架研究[J]. 信息安全研究,2019,5(4):39-47.
- [13] 王勇. 我国首部设计数据跨境流动地方法规评析[J]. 检察风云,2019(22):28-29.
- [14] 张荣楠. 跨境数据流动:全球态势与中国对策[J]. 开放导报,2020(2):44-50.
- [15] 顾伟. 警惕数据跨境流动监管的本地化依赖于管辖冲突[J]. 信息安全与通信保密,2018,300(12):27-32.

(本文责编:海 洋)