

推动科学数据开放共享的思考及启示

杨 晶, 康 琪, 李 哲

(中国科学技术发展战略研究院, 北京 100038)

摘 要: 在数字转型背景下, 科学数据已成为科学研究和创新发展的基础性战略资源, 加快推动科学数据开放共享对促进创新体系演化具有积极而重要的意义。本研究梳理总结了我国科学数据的发展现状与问题, 表明我国已在科学数据资源的积累和管理方面取得了一定成效, 为科学研究提供了重要支撑, 但在科学数据的存储积累、产权归属、范围边界、国际合作等方面仍存在一些需要解决的问题。建议加强对科学数据的保存、积累和分析挖掘, 加强对科学数据权利进行合理界定和有效保护, 逐步推动科学数据的分级分类开放, 提升科学数据开放共享的国际合作实践。

关键词: 科学数据; 开放共享; 数字转型; 数据管理; 开放数据

中图分类号: G311 **文献标识码:** A **DOI:** 10.3772/j.issn.1009-8623.2019.10.006

科学数据主要指在自然科学、工程技术科学等领域, 通过基础研究、应用研究、试验开发产生的数据及通过观测监测、考察调查、检验检测等方式取得并可用于科学研究活动的原始数据及其衍生数据^[1]。对于科学研究和国家创新发展来说, 科学数据是一种基础性战略资源, 开放共享是实现其自身价值与意义的根本途径。

科学数据开放共享是适应科学研究模式转变的必然要求。由于信息技术的影响, 关于科学的一切几乎都在变化中。科学研究越来越依赖大量、系统、高可信度的数据, 进而发展出与实验科学、理论推演、计算仿真这3种科研范式相辅相成的第4种科研范式——“数据密集型科学”。这一范式不仅意味着新的科研方法, 而且意味着重要的思维转变, 其目标是拥有一个所有科学数据都在线且能够彼此交互操作的世界^[2]。因此, 科学数据资源被视为一种科技基础设施的观念逐步得到认可, 国家层面的科研竞争力将更多取决于数据优势以及将数据转化为信息和知识的能力^[3]。在保障隐私和国家安全的

前提下, 最大限度地促进科学数据的流动性和可获取性至关重要。

科学数据开放共享是提高创新治理能力的重要手段。当前, 国际形势错综复杂, 经济下行压力加大, 抓住以信息化为标志的新技术革命和产业变革的机遇, 加强科学数据开放, 实现数据平台中各类创新主体之间的高效互动, 成为提升国家创新体系效率的关键。只有打通政府与各个创新主体之间的数据壁垒, 连接数据孤岛, 促进全社会的共同参与, 推动科学数据有效聚合、可用性开发和重复应用, 才能有效地发掘科学数据的潜在价值。

1 中国科学数据的发展现状

随着科技创新投入和水平不断增强, 中国已经发展成为国际上推动科学数据资源建设与发展的重要参与者, 越来越重视提升科学数据的管理和保存能力, 推动科学数据的开放共享。

1.1 科学数据的级别与类型

从科学数据的性质来看, 大体可以分为3个级

第一作者简介: 杨晶(1982—), 女, 哲学博士, 博士后, 助理研究员, 主要研究方向为国家创新体系和科技创新政策。

项目来源: 科技部战略研究专项“国家创新体系发展趋势、国际经验与模式研究”(ZLY201822)。

收稿日期: 2019-09-15

别,一是涉及国家安全、个人隐私的保密数据;二是涉及政府政务、影响国家竞争力的内部数据;三是公共财政资金支持产生的、完全不涉及国家秘密和个人隐私的公开数据。按照开放对象来看,可以

分为政府相关部门、国内和国外。按照数据的加工程度与科学数据产品价值可分为无偿和有偿两种形式(见图1)。

科学数据根据不同的分类标准可划分为不

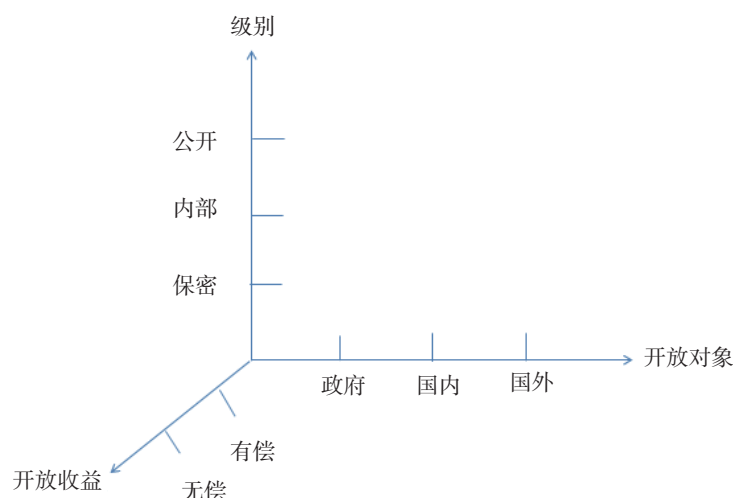


图1 科学数据的分级分对象开放示意图

同类型。例如,从“复用数据”FAIR原则来看,科学数据需满足“可发现(Findable)、可访问(Accessible)、可互操作(Interoperable)和可重复(Reusable)”的基本原则,根据这一原则将科学数据依开放状态分为6类,其中元数据、有限开放数据、开放数据和增强版开放数据4类数据被认为是开放数据的主要形式^[4]。从管理主体来看,主要由专业机构(数据中心)、科研团队以及科学家个

人等行为主体进行管理,不同管理主体的侧重点各不相同。从服务定位来看,包括研究型(科研项目中的数据)、资源型(领域中心的数据)和参考型数据(基础性数据)^[5]。从行业领域来看,包括国防军事、食品安全、智慧交通等。在不同分类标准下产生的各类科学数据的开放程度也不同(见图2)。

科学数据管理主体从个人、到团队再到专业机构,专业程度越高,开放程度也越高。研究型数据

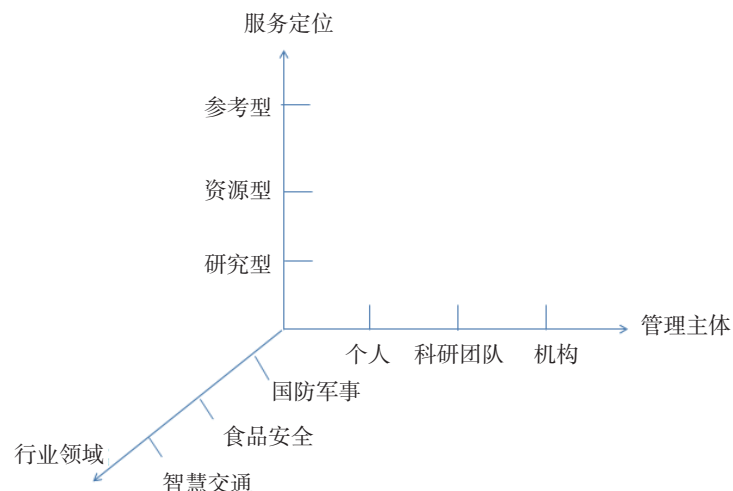


图2 科学数据的分类分领域开放示意图

的规模和覆盖范围有限,只能为少量特定的对象开放共享;资源型数据可以对某一领域或学科的组织 and 机构开放共享;参考型数据是基础性数据资源且规模庞大,开放范围更加广泛。从国防军事、食品安全到智慧交通,行业领域数据的开放程度逐渐提高。

1.2 科学数据资源及其分布

目前,中国已经形成了科研院所、高校及国家有关部门为主体和互补,同时囊括企业社会力量的多主体参与的科学数据资源生产格局。中国从20世纪80年代开始进行数据资源的积累,据不完全统计,截至2017年年底,中国有效管理与保存的科学数据资源总量共计约达83.72PB,这些数据主要分布在生命科学与医学、地球与环境科学、物理与化学、遥感对地观测、天文与空间科学、农业与林业科学等领域^[6](见图3)。其中,以中国科学院为代表的科研院所运用先进的管理方式和现代化保存手段在推动科学数据资源积累方面发挥了重要作用。初步建成了具有超过50PB容量、由分布在全国的13个存储节点组成的大型分布式存储系统,专门为科学数据存储、归档保存和异地容灾备份服务,目前已为500多个科研团队和项目提供数据存储服务,每天可实现20TB高速归档或数据恢复。此外,企业云服务已成为支持科研活动的重要组成部分,阿里云从2014年开始在全球布局,已建成覆盖亚太、北美、中东和欧洲等地区的11个地域节点和44个可用区的云服务环境。

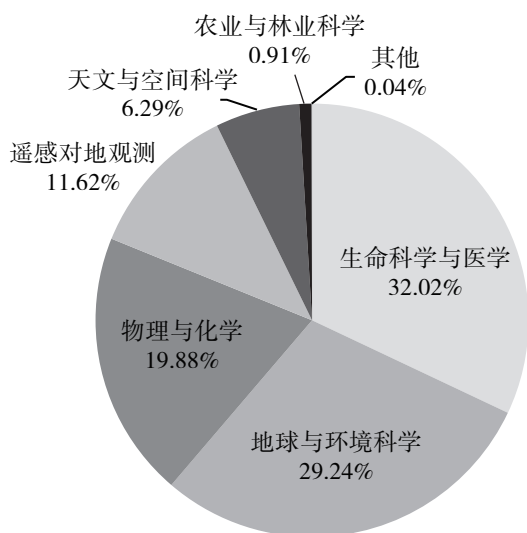


图3 现有科学数据资源领域分布情况

1.3 科学数据管理的成效与问题

科学数据管理是“一只看得见的手”。通过构建科学数据开放共享的政策体系、加强基础设施建设、推动国际合作等方面对科学数据进行有效管理,中国在提升科学数据的使用价值,支撑科学研究方面取得了重要进展,但仍存在一些需要解决的问题。

第一,形成全方位科学数据开放共享政策体系,具体数据开放机制尚需完善。中国对科学数据开放共享的政策及法规日益完善,形成了以各级行政主管部门、机构和行业领域等为主导的科学数据管理体系。自21世纪初开始,科技部就积极推动科学数据管理与共享工作,2001年向国务院提出“实施科学数据共享工程,增强国家科技创新能力”建议,现已形成一批纳入国家基础条件平台体系的科学数据中心。

近年来,国家对科学数据管理与开放共享的重视程度进一步加强。2015年国务院发布《促进大数据发展行动纲要》,提出加快政府数据开放共享,推动资源整合^[7]。2016年国家层面相继发布的多项文件中也明确包含了与加强科学数据管理、推动科学数据开放共享和促进科学数据开发利用有关的内容,科学数据开始被提升至国家战略高度。2018年,国务院办公厅发布《科学数据管理办法》,首次立足国家高度、面向多个领域科学数据,提出开放为主的指导原则,从科学数据采集、汇交与保存,共享与利用,保密与安全等方面对科学数据管理与共享进行规范^[8]。目前,北京、江苏、浙江等地区也陆续出台了促进大数据产业和政府信息资源管理的政策。大数据产业相关政策为科学数据发展勾画出蓝图,政务数据则可看作多种社会科学数据的综合体,这些政策对有效管理和繁荣科学数据具有重要意义。现阶段来看,行业领域政策侧重于宏观数据的整体管理,而计划项目政策则侧重于具体单项数据的管理。

经过十多年的努力,中国出台的政策及法律文件为科学数据的开放与共享提供了规范与保障,但仍存在一些具体问题需要进一步加以解决。其一,我国相关法规默认政府数据为公共资源并要求推动其开放,但缺乏政府和公众对数据权利、责任和义务的规定。其二,没有对数据开放的时限问题做出明确规定,影响了数据重复使用的效率,导致公共

科学数据开放滞后。其三, 没有对科学数据的分级分类开放做出可以参考实施的详细操作指南。

第二, 科学数据中心和平台发展迅速, 数据管理与应用能力尚待提升。自 2001 年以来, 中国开始重视科学数据中心建设, 截至目前初步形成了一批有影响力的科学数据中心。2002 年开始实施“科学数据共享工程”, 陆续启动气象、林业、农业等科学数据共享中心的建设与服务试点。2004 年, 启动国家科技基础条件平台建设专项, 目前国家科技资源共享服务平台作为促进科学数据开放共享的重要载体, 已在资源环境、农业林业、人口健康等多个学科领域建设了科学数据资源共享服务平台并持续开展大规模科学数据服务。与此同时, 科学数据平台发展与管理工作陆续展开。2011 年 23 个科技平台被认定为国家首批科技基础条件平台; 2017 年, 28 个国家科技资源共享服务平台通过考核评估^[4], 科学数据服务规模稳步扩展。中国科学院计算机网络信息中心启动国家发展改革委促进大数据发展重大工程项目“科学大数据公共服务平台与创新应用示范”^[4]。近年来, 进一步向数字化和网络化发展, 建立了大批的科学数据库并上网服务。据不完全统计, 在各级财政支持下, 已建成具有一定规模的科学数据库达 1 775 个。

虽然已在多学科领域建设了一批科学数据中心和大量科学数据库, 但不同数据中心所属行业类别、层级机构等纷繁多元, 不同管理主体所管理的数据中心侧重点各不相同, 服务成效评价仍显粗放。而且, 现有的数据管理与应用能力还比较有限, 在国际上具有权威优势或较高知名度的科学数据中心和数据库还较少。

第三, 以国际数据组织为纽带参与全球交流合作, 数据主权建设任重道远。中国通过参与国际数据组织平台建设, 在世界科技数据领域日益发挥越来越重要的作用。1984 年, 加入国际科技数据委员会 (CODATA) 并成立中国委员会及其管辖的 10 个科学数据协作组^[9]。1988 年, 加入世界数据系统 (WDS, 前身为世界数据中心 WDC), 截至 2016 年, WDS 共收录了 19 个国家和地区的 70 个普通数据中心 / 系统^[10]。其中, 中国共有 9 个数据中心 / 系统加入, 数量仅次于美国 (22 个), 这表明了中国的科技实力得到国际认可并逐步在世界科

技领域扮演日益重要的角色。2007 年, 启动了由中国科学院领衔的“促进发展中国家科学数据共享与应用全球联盟”计划^[11], 后来又促进了“发展中国家数据共享原则”的颁布。2014 年, 第一届“中国科学数据大会”召开并形成年届惯例。

但总体来看, 中国在国际上的科学数据开放共享活动仍以交流研讨居多, 尚且缺乏深入的国际合作实践。与此同时, 科学数据主权的流失现象突出, 鉴于西方发达国家利用其先发优势并掌握了各种科研成果发表平台, 对科学数据造成全球“虹吸”效应, 中国科研人员为了在国际学术期刊上发表论文不得不遵循这些国际期刊的要求, 纷纷将论文支撑数据提交至国外数据库中。

2 对科学数据开放的认识

科学研究的发展迫切需要科学数据开放共享, 但科学数据的开放并非一蹴而就, 需要重点考虑以下几个方面的问题。

2.1 科学数据的有效保存与积累是实现开放共享的前提

科学数据汇交理念在科技界已经形成共识, 完善数据的积累和长期保存机制是促进科学数据开放共享的良好基础。欧美发达国家和国际组织通过建立数据平台、应用数据管理软件、构建数据存储库等方式, 有力加强了科学数据的保存和积累。一是发达国家通过构建大批权威、高质量的科学数据平台, 保存了大量科学数据。例如, 英国数据存档库 (UKDA)、美国国立卫生研究院 (NIH) 支持建设的蛋白质数据库 (PDB) 和澳大利亚国家数据管理服务平台等, 这些国家级数据基础设施为保存和共享数据提供了重要支撑。二是“数据管理计划” (DMP) 工具的开发应用为促进数据管理、分享与长期保存, 提供数据全生命周期的管理服务带来了便利。截至 2017 年底, DMTTool 已经在 216 个组织机构中应用, 越来越多的科研人员和科研单位开始制定数据管理计划, 及时保存了一批来源清晰、质量可靠的数据资源^[6]。三是数据存储库成为对数据进行集中存储、管理并提供稳定服务的共性平台。为了满足数据存储库建设和发展的要求, 国际科学理事会 (ICSU) 下属的世界数据系统 (WDS) 推出数据可信印章 (DSA) 认证标准, 提供初步的可

信数据库评价标准。

2.2 明确科学数据的产权归属是实现开放共享的保障

科学数据从生产、管理到共享等不同阶段的权利应该合理划分,明确数据的权属关系和加强产权利益保护是科学数据开放的基本保障。但是由于数据与其他财产不同,其全生命周期存在多个参与者,且每个参与者在各自环节都对数据价值做出了相应的贡献,因此赋予某一参与者专属、排他性所有权不可行,数据的各类具体权利需要在各参与者之间进行协商划分。例如,德国的《基本法》确立了“科学研究自由”的原则,科学研究中产生的数据属“自有产权”范畴,归相关研究机构或/和相关研究人员所有。德国在国家层面没有一个中央机构来统筹管理科学数据,反之数据主权拥有者可在相关法律

框架内行使对科学数据的最终处置权,这一原则普遍适用于国家财政支持或经济界资助的研究项目所产生的科学数据。

2.3 识别科学数据开放共享的范围边界至关重要

科学数据的开放有其一定的范围边界,既应开放共享使各创新主体从中获益,又需保护数据所有者的利益不受侵犯。特别是涉及个人隐私、国家安全和机密,以及商业力量主导产生的科学数据,有必要尊重数据保密等合法要求,平衡、适度地开放。英国皇家学会研究报告《科学是一项开放的事业》中明确指出,科学开放边界的制约因素包括商业利益、隐私保护和公共安全等(见图4)。

2018年5月25日起正式实施的欧盟《通用数据保护条例》(GDPR),旨在保护数据时代的欧洲公民免于隐私数据泄露。该条例核心内容为确认

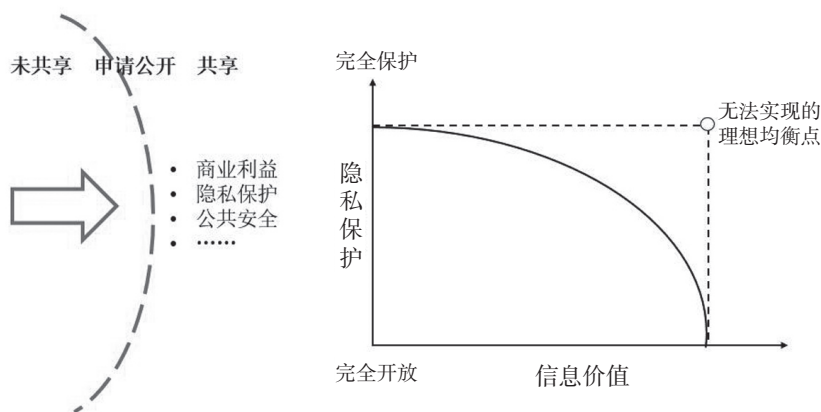


图4 隐私与科学数据价值的价值平衡曲线^[12]

数据主体“知情权、访问权、反对权、个人数据可携权、被遗忘权”等5种公民权利,被视为近20年来最严格的数据保护法案。可以说,科学数据的开放边界是数据从封闭到开放的动态博弈,如何寻找合理的平衡点将持续成为科学数据开放共享的焦点和难点。

2.4 建立广泛的国际合作是科学数据开放共享的必要途径

科学数据的管理与共享需要加强国际交流与合作。对于符合隐私、安全与社会公共利益诉求的数据资源,应该加大开放共享的范围,建立广泛的区域及国际合作,通过多样化用户群体的深入应用

和交互反馈来不断提升数据质量。世界上已有一些国际性的科学数据合作组织和计划,如:1996年,国际科技数据委员会成立,它是国际科学理事会下属一级学术机构,目前主要在数据政策研究、跨学科数据互操作与集成、非洲开放科学平台以及数据培训等方面发挥重要作用;2013年3月,欧盟委员会、美国国家科学基金会(NSF)和美国国家标准与技术研究院、澳大利亚创新部等共同组建了科学数据联盟(RDA);欧盟委员会地平线2020战略发布的《开放数据:创新、增长和透明治理的引擎》要求欧盟及其成员国建立相关的法律机制,促进成员国在开放数据领域的广泛合作^[13]。

3 关于构建中国科学数据开放共享体系的建议

面向中国庞大的科学研究体系,充分考虑科学数据各利益相关群体合理需求,亟须推动科学数据管理和开放政策向纵深发展,构建各类创新主体广泛参与的科学数据开放共享体系,有力支撑新时代国家创新体系建设。

3.1 加强科学数据的保存、积累和分析挖掘

在合理借鉴发达国家数据保存经验的基础上,加强公共财政支持产生科学数据的保存和规范管理。国家财政资金支持科研机构应在项目管理政策中引入数据政策,要求项目申请书包括数据管理计划和对项目所产生数据进行规范保存和适时开放的承诺。将数据保存和管理计划以及开放共享承诺的执行情况纳入科研项目管理政策制度。进一步探索科学数据存储、挖掘、分析等新技术的应用,推动不同学科之间科学数据的综合交叉使用,提升科学数据重复使用的效率和水平。此外,相关部门需要加快制定科学数据的有关标准,探索建立符合中国国情的科学数据标准体系。

3.2 对科学数据权利进行合理界定和有效保护

对个人数据、公共数据和商业数据的权利进行分类界定和保护,构建一个安全、自由的科学数据流通环境。将公共科学数据设定为国家所有模式,通过法律法规明确政府对公共科学数据的管理、使用权利,保护和开放的责任和义务,以及公众对公共科学数据的使用权利和义务,使得公共科学数据的权责利清晰且具可操作性,有效促进公共科学数据最大化开放。赋予非个人商业科学数据生产者所有权,明确规定数据企业的科学数据权利,即对企业合法创造的科学数据享有支配权利,鼓励数据企业收集、存储和利用科学数据,促进科学数据的流通和应用。通过深入研究,进一步明确数据使用、数据隐私保护、数据引用等关键问题,促进共享主体、共享范围多元化。

3.3 逐步推动科学数据的分级分类开放

探索科学数据的分级分类开放模式,对科学数据进行标准化、精细化管理,提升开放科学数据的质量。具体来说,对于涉及国家秘密和个人隐私的数据,建立数据保护指导性规范,采取禁止开放的

方式;对于公共财政支持产出的、不涉及国家秘密和个人隐私的科学数据,采取无偿向国内外开放的方式;对于深度分析和再加工、具有更高价值的衍生数据,采取有偿开放的方式;对于公众迫切需要的、对国家产业发展不形成竞争威胁的领域如交通、环境、农业等科学数据采取立即或在短期内开放的方式;对于参考型、资源型、专业机构产生的以及Web科学数据逐步加大开放力度。

3.4 提升科学数据开放共享的国际合作实践

要进一步拓展科学数据国际合作方式和渠道,主动参与开放共享活动,把握数据开放共享的主动权。通过参与相关国际组织和会议及国际科学数据平台建设,加强中国在科学数据资源方面的贡献程度,扩大中国科研成果的竞争力和影响力。采取多种激励措施,促进科研人员与出版商、数据中心、图书馆之间的合作,推动跨境科学数据流动制度接轨。加强具有自主知识产权的高质量第一手数据及高水平衍生数据的生产、管理与开放共享,提升国际化交流与推广水平。探索与国际科学数据基础设施及服务网络的合作,实现与国际数据服务网络的互联互通、资源交换与服务共享。■

参考文献:

- [1] 国务院办公厅.科学数据管理办法[EB/OL].(2018-04-12)[2019-05-06].http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm.
- [2] Hey T, Tansley S, Tolle K.第四范式:数据密集型科学发现[M].潘教峰,张晓林,等译.北京:科学出版社,2012:xxiii.
- [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,2018:1.
- [4] 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.
- [5] 诸云强,朱琦,冯卓等.科学大数据开放共享机制研究及其对环境信息共享的启示[J].中国环境管理,2015(6):38-45.
- [6] 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,2019:8.
- [7] 邢文明,洪程.开放为常态,不开放为例外——解读《科学数据管理办法》中的科学数据共享与利用[J].图书

- 馆论坛, 2019 (1): 117-124.
- [8] 杨咏梅. 科研数据开放驱动下高校图书馆学科服务转型研究 [J]. 图书馆工作与研究, 2019 (3): 73-77.
- [9] CODATA 中委会秘书处. CODATA 中国全国委员会 [EB/OL]. [2019-07-27]. http://www.codata.cn/zgwyh/201609/t20160906_4523013.html.
- [10] 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.
- [11] 王伟, 华夏, 王建梅. 国内外科学数据管理与共享研究 [J]. 科技进步与对策, 2013 (14): 126-129.
- [12] The Royal Society. Science as an Open Enterprise[R]. London: The Royal Society Science Policy Centre, 2012.
- [13] 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.

Thoughts and Enlightenment on Promoting the Open Sharing of Scientific Data

YANG Jing, KANG Qi, LI Zhe

(Chinese Academy of Science and Technology for Development, Beijing 100038)

Abstract: Under the background of digital transformation, scientific data has become the basic strategic resource for scientific research and innovation development. Promoting the open sharing of scientific data has a positive and important significance for the evolution of innovation system. This paper summarizes the development status and problems of scientific data in China, and shows that China has achieved certain results in the accumulation and management of scientific data resources, which provide important support for science research. However, there are still some problems to be solved in the storage accumulation, property rights, scope boundaries, and international cooperation of scientific data. It is recommended to strengthen the preservation, accumulation and analysis of scientific data; strengthen the rational definition and effective protection of scientific data rights; promote the open sharing of scientific data by classification gradually; and enhance the international cooperation practice of open sharing of scientific data.

Key words: scientific data; open sharing; digital transformation; data management; open data