

科技计划形成的科学数据汇交国家标准实践分析与思考建议 ——以 20 个国家科学数据中心为例

朱艳华^{1,2} 高瑜蔚^{1,2*} 李坤^{1,2}

(1.中国科学院计算机网络信息中心; 2. 国家基础学科公共科学数据中心)

摘要: 科技计划项目数据汇报是持续整合科学数据的重要方式,也是确保国家科学数据安全的关键举措。本文对20个国家科学数据中心落实科技计划形成的科学数据汇交国家标准,开展数据汇交工作进行比较分析。结果显示,目前所有国家科学数据中心均按照国家标准要求,制定了科学数据汇交制度和标准规范,开发了汇交数据管理服务平台。同时,在实际数据汇交过程中,不同国家科学数据中心在汇交流程和汇交内容等方面存在差异。最后,本文从提前制定数据汇交计划、提升汇交数据质量、推动汇交数据的再利用价值等方面提出了建议和对策。

关键词: 科学数据, 数据汇交, 国家标准, 标准实践

DOI编码: 10.3969/j.issn.1674-5698.2024.07.002

Analysis and Suggestions on the Practice of National Standards for Scientific Data Archiving Formed by Science and Technology Plans —A Case Study of 20 National Science Data Centers

ZHU Yan-hua^{1,2} GAO Yu-wei^{1,2*} LI Kun^{1,2}

(1.Computer Network Information Center, Chinese Academy of Sciences; 2. National Basic Science Data Center)

Abstract: The data archiving of science and technology plans is an important means of continuously integrating scientific data, which is also a crucial measure to ensure the security of national scientific data. This paper conducts a comparative analysis of the scientific data archiving formed by science and technology plans at 20 national science data centers. The results indicate that all national data centers have established regulations and standard specifications related to data archiving, and have developed data archiving service platforms. However, differences exist in the actual data archiving process in the aspects of archiving procedures and entities. Finally, recommendations and measures are proposed such as establishing a data archiving plan in advance, improvement of archiving data quality, and promoting the reuse of archiving data.

Keywords: scientific data, data archiving, national standards, standard practice

基金项目: 本文受中国科学院网络安全和信息化专项咨询研究项目“科学数据要素化治理及智能化应用理论与技术”(项目编号: CAS-WX2023ZX01-11)资助。

作者简介: 朱艳华, 高级工程师, 硕士, 主要从事大数据技术与标准规范、数据应用服务研究。

高瑜蔚, 通信作者, 工程师, 博士, 主要从事大数据技术与标准规范、数据应用服务研究。

李坤, 工程师, 硕士, 主要从事科学数据服务与应用研究。

0 引言

科学数据是科学研究和创新发现的重要基础性战略资源。开展科技计划项目形成的数据汇交是持续整合科学数据的重要途径,也是保障国家科学数据安全的重要举措。欧美等发达国家高度重视政府资助科技计划项目产生的数据汇交系统化、规范化管理,项目资助机构产生的数据已成为欧美国家科学数据资源建设和持续积累的重要组成。早在2000年,美国典型的科研资助机构,如:国立卫生研究院(National Institutes of Health, NIH)^[1]、国家科学基金会(National Science Foundation, NSF)^[2]、能源部(Department Of Energy, DOE)^[3]、航空航天局(National Aeronautics and Space Administration, NASA)^[4]等陆续制定了数据管理共享政策,要求项目申请者在立项阶段,就要提交数据管理计划。2020年10月,美国国立卫生研究院NIH发布了新的《数据管理与共享政策》,并于2023年1月生效,目的是促进NIH资助研究所产生的科学数据管理和共享。新政策将科学数据定义为:“科学界普遍接受的记录事实材料,其质量足以验证和复制研究结果,不论数据是否用于支持学术出版物。”与原政策相比,新政策增加了数据质量要求,还规定了项目承担单位应尽快共享数据,不迟于相关出版物发表或支持期结束,以先到者为准^[5]。

我国已经意识到投资大量经费所产生的科学数据资源,尚未在国内得到充分的利用,甚至有的数据流向了国外^[6]。近年来,我国制定的系列数据政策都有涉及数据汇交内容。2018年3月,国务院办公厅印发《科学数据管理办法》,明确提出“各级科技计划(专项、基金等)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”^[7]。2018年12月,科技部办公厅印发《国家重点研发计划项目综合绩效评价工作规范(试行)》,明确要求项目绩效评价材料中应包括科技资料汇交方案,应提交由有关方面认可的科学数据中心出具的汇交凭证^[8]。2019年6月,科技部、财政部对原有国家平台开展了优化调整工作,形成20个

国家科学数据中心^[9]。2019年7月,科技部基础研究所、资源配置与管理司印发《关于开展科技基础性工作专项项目综合绩效评价工作的通知》,将科学数据汇交列入综合绩效评价程序中,并明确项目科学数据汇交的流程和汇交的内容要求^[10]。同年12月,科技部办公厅正式印发《科技计划项目科学数据汇交工作方案(试行)》,明确科学数据汇交原则、管理主体与职责、主要内容及汇交流程,进一步规范科技计划项目科学数据汇交工作。2022年8月,中国科学院印发《中国科学院战略性先导科技专项科学数据汇交管理实施细则(试行)》,指导先导专项有效、保质地完成数据汇交工作,使科学数据作为国家战略性资源得以安全保存和利用^[11]。此外,为了积极推动政府预算资金资助的各级科技计划项目所产生的科学数据,向以国家科学数据中心为代表的科学数据管理机构汇交,我国一直致力于推动数据汇交标准规范研制工作。2021年3月,科技计划形成的科学数据汇交3项国家标准正式发布,全面指导科学数据汇交工作。

同时,我国已有不少学者围绕科学数据汇交展开了多方面的研究工作,其研究主题主要集中在数据汇交政策分析以及汇交进展跟踪。如:(1)在数据汇交政策分析方面,胡吉明等^[12]融合政策目标、政策工具、政策领域3个维度,评价我国科学数据汇交政策。支凤稳等^[13]从国家、地方、具体领域3个层面分别选取数据汇交政策,总结分析其汇交主体、汇交计划、汇交流程等重点内容。徐波等^[14]结合国外科研项目数据汇交管理经验,提出科技计划项目科学数据汇交的通用流程及汇交的主要内容。(2)在数据汇交落地实施跟踪方面,丁晓芹等^[15]通过调研分析国家层面、省(区、市)层面以及重要机构的科学数据汇交情况。徐墨庚等^[16]提出了基于区块链的科学数据汇交存证解决方案,结合海洋科学数据汇交系统设计存证智能合约,实现汇交全流程可追溯。陶毅等^[17]分析国内外科学数据管理情况,提出基于元数据理念建立分层架构的计量科学数据汇交系统构建。吴思竹等^[18]系统分析了国家人口健康科学数据中心数据仓储在科研项目数据汇交方面的功能设计和实施方法。白

燕等^[19]基于2019年度91个科技基础性工作专项项目数据汇交实践,从规范性、完整性、一致性、数据质量、数据说明文档5个方面,系统总结数据汇交的质量评价内容及其存在的典型问题,并给出具体建议。然而,目前针对科学数据汇交标准进展跟踪研究不足。特别是2021年3项科学数据汇交国家标准发布以后,标准落地实施进展如何,目前尚未开展分析研究。

本文调研了作为高质量科学数据资源汇交管理基地的国家科学数据中心,在执行科技计划形成的科学数据汇交标准,开展数据汇交工作实践现状,并从标准视角对科学数据汇交工作提出建议 and 对策。本研究旨在为其他科学数据管理机构提供数据汇交工作参考和借鉴,以便精确追踪、记录和管理科技项目所产生的海量科学数据。这将有助于挖掘汇交数据的潜在价值和信息,从而加速科研成果的传播和共享,促进科技的发展和进步。需要特别强调的是,本文所称的科学数据汇交专指科技计划项目所形成的科学数据汇交。

1 科学数据汇交国家标准研制情况

2021年3月,由国家科技基础条件平台中心牵头研究制定的GB/T 39908-2021《科技计划形成的科学数据汇交 通用代码集》^[20]、GB/T 39909-2021《科技计划形成的科学数据汇交 通用数据元》^[21]和GB/T 39912-2021《科技计划形成的科学数据汇交 技术与管理规范》^[22]3项国家标准正式发布,并于同年10月实施。标准起草单位涵盖了7家科技计划项目管理专业机构、20个国家科学数据中心以及中国科学院相关研究所、中国标准化研究院等研究机构。

3项国家标准的发布实施,对于规范科技计划项目科学数据汇交,具有重要的基础性作用。标准明确界定了数据汇交的管理主体及其责任、数据汇交的主要内容和范围、数据汇交的通用流程,同时规范了通用数据元和通用代码集。其中,《科技计划形成的科学数据汇交 通用代码集》规定了科技计划形成的通用数据元和通用代码集;《科技计划

形成的科学数据汇交 通用数据元》规定了科技计划形成的科学数据汇交通用数据元描述规范、数据元目录;《科技计划形成的科学数据汇交 技术与管理规范》规定了科技计划形成的科学数据汇交的原则、管理的主体与职责、主要内容及流程。

目前,科技计划项目形成的科学数据主要汇交至20个国家科学数据中心。为推动国家重点研发计划项目科学数据汇交工作,国家科技基础条件平台中心积极组织国家科学数据中心对接项目管理专业机构,细化数据汇交流程,研发数据汇交在线系统。同时,依托中国科技资源共享网开设“数据汇交”栏目,发布数据汇交的政策要求,为项目管理和项目承担单位提供支撑与服务^[23]。基于科学数据汇交国家标准和相关数据政策,20个国家科学数据中心遵循科技部等相关部门的部署和要求,积极制定学科领域科学数据汇交细则,并着手开发领域科学数据汇交管理平台。这一举措旨在使科研工作者更便利地管理和共享已有的研究成果和数据资源。

2 科学数据汇交国家标准实践分析

2.1 数据汇交制度规范

目前所有国家科学数据中心均已制定或发布了与数据汇交相关的政策法规和标准规范。这些文件大致可分为3种类型:第一类是数据汇交管理办法,第二类是数据汇交指南,第三类是数据汇交标准规范。具体而言,有11个国家科学数据中心制定了数据汇交管理办法,如:《高能物理科学数据汇交管理办法》^[24]《国家材料腐蚀与防护科学数据中心数据汇交管理办法》^[25]等。6个国家科学数据中心制定了数据汇交指南和汇交细则,如:《国家微生物科学数据中心数据汇交协议》^[26]《极地领域国家重点研发计划项目数据汇交与共享实施细则(暂行)》^[27]《农业数据汇交工作方案》^[28]。另外,3个国家科学数据中心制定了数据汇交标准规范,如:《国家生态科学数据中心国家重点研发计划项目汇交数据整理方法》^[29]。其中一些文件已在国家科学数据中心的网站上公开发布,可供获取;而另一

部分文件尚未对外公布, 仅供内部使用。

深入分析发现, 数据汇交管理办法内容较为宏观, 通常包含数据汇交的基本原则、汇交种类及范围、汇交主体与职责、以及汇交流程等通用要求。而数据汇交指南和汇交细则相对详尽, 并提供了数据汇交模板, 例如: 《国家基础学科公共科学数据中心科技计划项目科学数据汇交工作指南》在其附录中列出了7个模板: 科技计划项目科学数据汇交计划(模板)、科技计划项目科学数据集元数据(模板)、科技计划项目科学数据质量报告(模板)、科技计划项目科学数据使用说明(模板)、科技计划项目工具软件元数据(模板)、科技计划项目科学数据关联成果(模板), 以及科技计划项目科学数据汇交凭证(模板)。而数据汇交标准规范则更侧重操作要求, 如: 国家生态科学数据中心所制定的《国家重点研发计划汇交数据整理方法》规定了数据集元数据描述字段填写、汇交文件存储目录结构以及命名规范等具体要求。

2.2 数据汇交流程

GB/T 39912-2021《科技计划形成的科学数据汇交 技术与管理规范》规定, 科技计划项目形成

的科学数据的通用汇交流程, 通常包括科学数据汇交计划制定、科学数据制备、科学数据提交、科学数据审核、科学数据汇总、科学数据发布与共享, 以及科学数据使用与维护更新等7个阶段。详见表1。

所有国家科学数据中心均遵循标准所规定的科学数据汇交流程开展数据汇交工作。不过在实操中, 存在一定差异。有些国家科学数据中心的汇交流程超过7个阶段, 主要是对标准中所规定的汇交流程进行了细化。例如: 国家空间科学数据中心增设了“项目数据汇交方案制定”环节; 国家高能物理科学数据中心添加了“数据汇交系统用户注册、填写数据汇交意向以及签订数据汇交协议”等环节。而国家天文科学数据中心则引入了“汇交计划审核、提交元数据信息以及元数据审核”等环节。

尽管当前国家科学数据中心的首要任务, 是管理数据汇交过程并开具汇交凭证, 但同时也开展了诸如数据开放共享和持续更新维护等后续工作。毕竟科技计划项目数据汇交的最终目标, 是促进数据的再利用。将科学数据提交给科学数据中心只是任务的一部分, 科学数据中心还需要对接收到

表1 科学数据汇交流程^[22]

序号	阶段名称	基本要求	阶段性成果
阶段一	科学数据汇交计划制定	科学数据提交方应在项目立项阶段按照科技项目管理方的相关规定, 编制科学数据汇交计划, 提出质量控制的手段和方法, 明确科学数据开放时间及其共享方式等内容	《科学数据汇交计划》
阶段二	科学数据制备	科学数据汇交计划通过审查后, 科学数据提交方应遵照科学数据汇交计划和相关标准规范, 进行科学数据实体采集生产与加工整理, 按规定格式形成科学数据的数据元信息	规范化数据实体、数据元描述信息
阶段三	科学数据提交	科学数据提交方应按照科学数据汇交计划, 对计划汇交的科学数据进行汇总整理, 对科学数据质量进行自查, 编制科学数据质量信息报告, 并将科学数据提交至科学数据管理方	可共享数据实体、数据集元数据、科学数据质量报告等
阶段四	科学数据审核	科学数据管理方按照科学数据汇交计划和科学数据质量控制体系要求进行形式审查, 通过组织开展科学数据质量同行评议等方式对科学数据质量进行评估。审核通过时, 科学数据管理方出具审查报告作为汇交凭证, 并将其提交至科技项目管理方, 由科技项目管理方审批后汇总	《科学数据汇交凭证》
阶段五	科学数据汇总	科学数据管理方在接收科学数据后, 应对通过审核的科学数据进行分类、编目、标识、保存、加工、整理、管理与维护	数据资源目录
阶段六	科学数据发布与共享	科学数据管理方负责保障科学数据安全, 经由科技项目管理方审批后对外公布科学数据及目录。按照开放为常态、不开放为例外的原则, 根据科学数据的开放条件、开放对象和审核程序等, 在保障科学数据安全的前提下向社会开放共享	数据在线发布与共享服务
阶段七	科学数据使用与维护更新	科学数据管理方应做好科学数据的管理与维护工作, 对科学数据共享应用情况进行统计, 并将统计结果反馈至科技项目管理方	数据应用情况跟踪与统计

的数据进行分类编目、标识、保存、加工整理、发布共享以及维护更新等持续工作。

2.3 数据汇交内容

GB/T 39912-2021《科技计划形成的科学数据汇交 技术与管理规范》明确规定,“科学数据汇交内容包括:科技计划项目执行过程中产生的科学数据实体、科学数据描述信息和科学数据辅助工具软件。”

2.3.1 科学数据实体

科学数据实体是科学数据汇交的核心内容。标准规定:“科学数据实体包括在科技计划形成的原始数据、以及基于原始数据或研究分析数据所形成的完整数据文件或数据库”。

根据《科学数据管理办法》和汇交标准规定,重点研发计划项目汇交的科学数据应同时具备以下特征:(1)受到重点研发计划项目资金支持,由项目方自主取得的;(2)通过基础研究、应用研究、试验开发、观察观测、考察调查、检验检测等科学研究活动产生的原始数据及其衍生数据;(3)数据的形式包括数据库与数据文件,类型包括但不限于科学数据、证明科学数据质量的(第三方)检测报告等成果,以及具有再利用或保存价值的音视频资料、照片图表等;(4)按照国家相关法规经由资质机关、单位定密的秘密数据,以及数据内容涉及个人隐私且未经必要技术处理的科学数据,在对应科学数据中心取得涉密资质之前暂不汇交。在数据汇交过程中,国家科学数据中心对提交项目成果文件的具体要求不同。一些数据中心要求项目方提交所有相关项目成果文件,包括论文、专利、标准等,以确保记录的完整性。而一些数据中心则并不要求提交项目成果文件,而是以标准定义的科学数据实体为重点,将文章、标准、专利和研究报告等成果视为可选择提交的辅助材料。

举例而言,国家基础学科公共科学数据中心连续多次邀请专家进行研讨,以确定每个项目应提交的科学数据实体^[30]。具体包括:(1)考核指标或研究内容中的数据库,其汇报数据量应确保不低于“完时指标值”或考核结论中提及的数值。(2)科技计划生成的原始数据,主要包括项目实施过程中

产生的观测数据、探测数据、试验数据、调查数据等。(3)分析数据,指项目针对科研目标进行数据分析挖掘时产生的结果。(4)关键考核指标的支撑数据,例如:论文、发明专利、标准以及支撑数据。这些分类采用不同视角,可能存在一定程度的重叠,但基本不影响实际数据汇交工作。

2.3.2 科学数据描述信息和数据辅助工具软件

标准规定“科学数据应提供相应的元数据描述信息”“并根据实际需要提供比元数据更详细的数据说明文档及过程资料”。同时GB/T 39909-2021《科技计划形成的科学数据汇交 通用数据元》规定了9类数据元:与科技计划项目、科学数据有关的通用数据元;与机构、人员以及通信联系有关的数据元;与空间坐标、地理位置有关的数据元;与日期、时间、期限有关的数据元;与数量、计量单位等有关的数据元;与科学数据汇交过程有关的数据元;与科学数据描述与标识等有关的数据元;备用数据元;其他数据元。

基于数据在线服务和再利用的考虑,这些描述信息可分为3类:首先,帮助用户在国家科学数据中心网站上查找到目标数据库的描述信息。这类元数据应遵循GB/T 30523-2023《科技资源核心元数据》,并根据需要扩展领域特色内容。其次,帮助用户了解数据的描述信息。最后,提供帮助用户使用数据的说明文档及过程资料。考虑到各个学科领域数据特点,数据中心可根据实际需要扩展学科领域描述元数据。例如:国家冰川冻土沙漠科学数据中心规定:“默认支持科学数据提交方按照GB/T 30523-2014《科技平台资源核心元数据》(已更新为:GB/T 30523-2023《科技资源核心元数据》)提交数据集元数据,如果提交方已有符合其他标准的元数据,则可按原格式提交^[31]。”

关于数据辅助工具软件,标准规定,“科学数据辅助工具软件,是指科技计划形成的用于科学数据处理、加工和分析的专门辅助软件工具等。”要确定软件是否应该进行数据汇交,关键在于遵循规定提到的“专用性”原则,即只有那些项目团队为汇交数据开发的专用软件工具才需要汇交。此外,一些数据中心还提出了额外要求。如:国家空

间科学数据中心规定：“汇交的科学数据辅助工具软件应配备相应的描述信息，提交的科学数据辅助工具软件必须符合相关的软件安全规定^[32]。”

为了提高科技计划项目资源的汇交与共享效率，国家科学数据中心充分利用信息化手段，开发了基于网络服务的数据汇交服务平台，以规范数据汇交流程，加强汇交数据管理。各国家科学数据中心的数据汇交入口和联系信息，详见中国科技资源共享网“数据汇交”栏目^[33]。此外，一些国家科学数据中心还探索采用了区块链等技术，以确保汇交数据的安全性和可追溯性。举例来说，2022年11月，国家海洋科学数据中心推出了全国首个科技计划项目数据汇交区块链系统。该系统实现了数据文件提交、审查、保存和出具汇交证明全流程上链留痕，初步建立了覆盖涉海科技计划项目汇交全流程的区块链存证体系。2023年1月，国家基础学科公共科学数据中心正式推出基于区块链的基础科学数据可信服务，科技项目所形成的科学数据资源上链可追溯便查询，支持可信平台查询；同时集成了国家授时中心时间服务，开展汇交凭证节点“时间戳”可信认证，提升数据存储安全性、可靠性和隐私保护。

3 科学数据汇交标准思考与建议

3.1 提前制定数据汇交计划，规范数据汇交过程管理

根据GB/T 39912-2021《科技计划形成的科学数据汇交 技术与管理规范》的规定，“科学数据提交方应在项目立项阶段，依照科技计划项目管理方的相关规定编制科学数据汇交计划”。然而，在实际操作中尚未全面实施该规定。我国科研数据的管理和提交，通常延至项目结题验收阶段，尽管这种模式有利于项目整体管理需要，却不利于科研过程中数据监管和质量控制。许多科研项目研发周期较长，项目结束后，数据的归档备份和长期服务难以执行，导致数据质量追踪和溯源变得困难。究其原因，科学数据汇交工作涉及新内容和新任务，要在不影响现有科技活动的前提下，逐步形

成社会共识。

国外科研资助机构将数据管理视为项目申请和研究的起点，在项目立项申请书中明确要求提交数据管理和共享计划，以确保科研过程中产生的数据得到有效管理和利用。例如：美国国家科学基金会NSF项目管理指南、国立卫生研究院NIH数据共享计划以及能源部DOE数据管理声明，均要求项目申请人提交数据管理计划，明确数据共享范围和管理规划。目前，我国也有部分机构开始采取类似的做法。例如：2023年起，中国生物技术发展中心要求申报国家重点研发计划项目需在申报立项阶段完成科学数据汇交计划，其采用的模板与目前的数据汇交计划一致。这种前置数据汇交计划的举措与国际数据管理实践相似，值得其他项目管理机构借鉴和推广。

可以考虑在国家重点研发计划项目中率先实施在“立项阶段编制科学数据汇交计划”，并逐步推广至所有项目。为了规范项目管理，项目组应及时制定数据汇交管理计划，有序实现数据生产、管理和共享，以确保科学数据安全有效利用。建议项目组在立项后，尽早考虑与目标国家科学数据中心合作，获取数据汇交标准模板，提前做好数据汇交准备。

3.2 发展高质量数据治理策略，全面提升汇交数据质量

当前，科学数据汇交工作正在有序稳步进行，每个国家科学数据中心也积极建立数据质量审核机制，邀请学科领域数据质量专家，开展数据质量审核。然而，总体而言，项目汇交数据质量提升面临重大挑战。首先，十三五立项的项目并没有同步完成项目数据汇交计划，项目研究人员对科学数据汇交的认识和理解程度不够，数据采集和加工过程缺乏标准规范，导致一些过程数据未能完整保留，影响了数据整体质量。其次，超大规模科学数据质量审查在操作层面面临实际困难。以国家基础学科公共科学数据中心为例，已汇交的数据超过1.49PB，组织领域专家即使只抽取一定比例的样本进行审查，难度依然很大。最后，缺乏高质量数据治理策略和方法。尽管国家科学数据中心已经将数据质量控制要求纳入数据汇交模板，但目前数据

质量控制软件只解决部分问题,领域专家知识规则仍需进一步挖掘,非结构化数据的质量自动化识别程度也需要进一步提高。

国家重点研发计划项目涵盖众多学科领域,项目数据质量控制体系和质量方案存在差异。在尚未建立成熟数据质量控制方案之前,可借鉴ISO 9000系列质量管理、ISO 8000系列数据质量、GB/T 36073《数据管理能力成熟度评估模型》以及GB/T 34945《数据溯源描述模型》等通用质量标准的全流程闭环理念。通过规范数据集、数据活动、参与人员、软件工具等相关实体,提升数据质量。此外,在科学数据汇交过程中涉及的数据元定义和描述,建议遵循GB/T 32843《科技资源标识》、GB/T 30523《科技资源核心元数据》等相关标准。目前,这几项国家标准已在国家科学数据中心数据汇交工作中得到落实。同时领域元数据的扩展应符合GB/T 30522《科技平台 元数据标准化基本原则与方法》中规定的标准化流程。

2024年1月,17个部门联合印发了《“数据要素×”三年行动计划(2024-2026年)》,以充分发挥数据要素的乘数效应,为经济社会发展赋能^[34]。在该行动计划中,针对数据要素在科技创新中的作用,明确提出依托国家科学数据中心等平台,加强高质量科学数据资源建设和场景应用。为此研制贯穿数据全生命周期的高质量数据治理标准体系至关重要。通过高水平标准推动高质量科学数据资源体系建设,重点围绕科学数据质量政策、标准规范、共性技术、服务模式、监测评估与效能评价等方向,开展系统研究工作。通过国家科学数据中心与项目科学家的密切合作,科学数据汇交有望在解决数据质量难题方面率先采取行动。

3.3 推动汇交数据规范化加工和利用,带动项目结题后数据的持续汇交和共享

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^[35]。科学数据汇交的目标是发挥数据的再利用价值。为此科学数据中心需要迅速对汇交来的数据进行规范化整理加工,及时发布和共享。同时,

《科学数据管理办法》明确提出,项目/课题验收后产生的科学数据也应进行汇交,明确要求主管部门和法人单位应积极推动科学数据出版和传播工作,科研人员整理发表产权清晰、准确完整、共享价值高的科学数据。数据出版通过一定的公共机制发布科学数据集,使得公众根据一定规则发现、获取、评价和应用这些数据,其中数据论文出版是得到业界广泛认可的重要形式^[36]。

2022年12月,《中共中央 国务院关于构建数据基础制度更好发挥数据要素作用的意见》(以下简称“数据二十条”)正式公布,从数据产权、流通交易、收益分配、安全治理等多方面建立数据基础制度,并提出了20项政策措施^[37]。该文件明确指出,鼓励公共数据在维护个人隐私和保障公共安全的前提下,按照“原始数据不出域、数据可用不可见”的要求,以模型、核验等产品和服务等形式向社会提供。依据“数据二十条”的相关规定,亦须考虑加强推动汇交科学数据形成标准化的数据产品,建立可确认数据来源、明确使用范围、可追溯流通过程,以及防范安全风险的可信数据流通体系,最大程度释放数据要素的价值。

值得注意的是,2024年2月,美国总统拜登正式签署了《关于防止受关注国家获取美国公民大量敏感个人数据和美国政府相关数据的行政命令》,限制包括中国在内的国家科研人员访问和使用美方的“敏感数据”,其中个人健康数据和人类基因组数据是被关切的重点^[38]。科学数据作为一类重要的数据,与国家发展息息相关,如何在安全和发展中取得平衡,建立科学数据跨境流动过程中的互信合作机制和标准规范,应对复杂多变的国际环境挑战,也是我国在科学数据汇交管理和共享利用过程中关注的方向。

4 结语

3项科学数据汇交国家标准的发布和实施,对于规范科技计划项目形成的科学数据汇交管理具有重要的指导和规范作用。自科学数据汇交工作启动以来,已有超过3000个国家重点研发计划项目向

国家科学数据中心提交数据。目前,十三五重点研发计划项目数据汇交工作已经逐步完成,同时部分十四五项目也已经开始启动数据汇交工作。这些汇交的数据实体及其元数据信息,部分已经在国家科学数据中心门户网站上发布。不过要真正推动项目承担单位在项目立项阶段,按照规定编制数据汇交计划,并将数据管理贯穿到其科研活动

全过程之中,全面提升汇交数据质量,充分发挥汇交数据再利用价值,仍然有许多问题值得进一步的思考和研究。

致谢: 本研究得到中国科学院监督与审计局的支持,在此一并感谢。

参考文献

- [1] NIH Data Sharing Policy and Implementation Guidance[EB/OL].[2024-1-14].http://grants1.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm#ex.
- [2] NSF Data Sharing Policy[EB/OL].[2024-1-14].<http://www.nsf.gov/bfa/dias/policy/dmp.jsp>.
- [3] Statement on Digital Data Management of DOE[EB/OL].[2024-1-14].<http://science.energy.gov/funding-opportunities/digital-data-management/>.
- [4] National Aeronautics and Space Administration Directories Science[EB/OL].[2024-1-14].<http://science.nasa.gov/>.
- [5] Final NIH Policy for Data Management and Sharing[EB/OL].[2024-1-14].<https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html>.
- [6] 科技部举行《科学数据管理办法》新闻通气会[EB/OL].[2024-1-14]. http://www.scio.gov.cn/xwfb/bwxwfb/gbwfbh/kxjsb/202207/t20220715_206870.html.
- [7] 国务院办公厅关于印发科学数据管理办法的通知[EB/OL].[2024-1-14]. https://www.most.gov.cn/xxgk/xinxifenlei/fdzdgknr/fgzc/gfxwj/gfxwj2018/201804/t20180404_139023.html.
- [8] 科技部办公厅关于印发《国家重点研发计划项目综合绩效评价工作规范(试行)》的通知[EB/OL].[2024-1-14].https://www.gov.cn/zhengce/zhengceku/2018-12/31/content_5446186.htm.
- [9] 科技部 财政部发布国家科技资源共享服务平台优化调整名单[EB/OL].[2024-1-14]. https://www.gov.cn/xinwen/2019-06/11/content_5399105.htm.
- [10] 科技部基础研究司 资源配置与管理司关于开展科技基础性工作专项项目综合绩效评价工作的通知[EB/OL].[2024-1-14]. https://www.gov.cn/zhengce/zhengceku/2019-12/03/content_5457753.htm.
- [11] 中国科学院印发战略性先导科技专项科学数据汇交管理实施细则[EB/OL].[2024-1-14]. http://www.ecas.cas.cn/yjgdt/202209/t20220914_4939387.html.
- [12] 胡吉明,陈恬,杨云. 基于政策特征的我国科学数据汇交一致性评价分析[J/OL]. 情报科学. <https://link.cnki.net/urlid/22.1264.G2.20240126.1851.014>.
- [13] 支凤稳,赵宇晨. 我国科技计划项目科学数据汇交政策分析[J]. 图书馆学研究, 2021, (23): 45-52.
- [14] 徐波,王瑞丹,王卷乐,等. 科技计划项目科学数据汇交共性机制研究[J]. 中国科技资源导刊, 2021, 53 (01): 9-14.
- [15] 丁晓芹,汤怡洁,徐雯. 我国科学数据汇交管理现状及面临的问题[J]. 科技管理研究, 2023, 43 (23): 63-69.
- [16] 徐墨庚,姜晓轶,童心,等. 基于区块链的科学数据汇交存证模型研究与实现[J]. 海洋信息技术与应用, 2023, 38 (04): 238-248.
- [17] 陶毅,苏爽,赵正宜,等. 基于元数据的计量科学数据汇交系统研究[J]. 中国科技资源导刊, 2022, 54 (02): 1-12+92.
- [18] 吴思竹,钱庆,周伟,等. 面向人口健康领域科研项目数据汇交的数据仓储设计与实现[J]. 数据分析与知识发现, 2020, 4 (12): 2-13.
- [19] 白燕,杨雅萍. 科技基础性工作专项数据汇交实践与启示[J]. 中国科技资源导刊, 2020, 52 (04): 70-79.
- [20] 全国科技平台标准化技术委员会. GB/T 39908-2021, 科技计划形成的科学数据汇交通用代码集[S]. 北京: 中国标准出版社, 2021:3.
- [21] 全国科技平台标准化技术委员会. GB/T 39909-2021, 科技计划形成的科学数据汇交通用数据元[S]. 北京: 中国标准出版社, 2021:3.
- [22] 全国科技平台标准化技术委员会. GB/T 39912-2021, 科技计划形成的科学数据汇交技术与管理规范[S]. 北京: 中国标准出版社, 2021:3.
- [23] 平台中心牵头研究制定的《科技计划项目形成的科学数据汇交技术与管理规范》等3项国家标准正式发布

- [EB/OL]. [2024-2-19]. https://www.gov.cn/xinwen/2021-04/03/content_5597642.htm.
- [24] 高能物理科学数据汇交管理办法[EB/OL]. [2024-2-19]. <https://www.nhepsdc.cn/standard/7>.
- [25] 国家材料腐蚀与防护科学数据中心数据汇交管理办法[EB/OL]. [2024-2-19]. <https://www.corrddata.org.cn/corrosion/index.php>.
- [26] 国家微生物科学数据中心数据汇交协议. [EB/OL]. [2024-2-19]. <https://nmcdc.cn/submit/agreement>.
- [27] 极地领域国家重点研发计划项目数据汇交与共享实施细则(暂行)[EB/OL]. [2024-2-19]. <https://www.chinare.org.cn/regulations/detail?id=0daa80e4-4a9c-4860-aec7-6b17c4374be6>.
- [28] 农业数据汇交工作方案 [EB/OL]. [2024-2-19]. <https://www.agridata.cn/gather.html>.
- [29] 国家生态科学数据中心国家重点研发计划项目汇交数据整理方法[EB/OL]. [2024-2-19]. <http://www.nesdc.org.cn/data>.
- [30] “十四五”重点研发计划新项目科学数据汇交工作的几点思考[EB/OL]. [2024-2-19]. <https://mp.weixin.qq.com/s/UEAiup5-ttiF90jx1z78Vg>.
- [31] 国家冰川冻土沙漠科学数据中心数据汇交[EB/OL]. [2024-2-19]. <http://www.nodc.ac.cn/portal/collect>.
- [32] 国家空间科学数据中心数据汇交范畴[EB/OL]. [2024-2-19]. <https://pdat.nssdc.ac.cn/login>.
- [33] 中国科技资源共享网[EB/OL]. [2024-2-19]. <https://www.escience.org.cn/data-archive/contact>.
- [34] 十七部门关于印发《“数据要素×”三年行动计划(2024—2026年)》的通知[EB/OL]. [2024-2-19]. http://www.cac.gov.cn/2024-01/05/c_1706119078060945.htm.
- [35] 苏靖. 大数据时代加强科学数据管理的思考与对策[J]. 中国软科学, 2022 (09): 50-54.
- [36] 高瑜蔚,朱艳华,孔丽华,等. 数据论文及关联科学数据集出版元数据标准研究[J]. 中国科技期刊研究, 2023,34(10): 1270-1282.
- [37] 中共中央 国务院关于构建数据基础制度更好发挥数据要素作用的意见[EB/OL]. [2024-2-19]. https://www.gov.cn/zhengce/2022-12/19/content_5732695.htm.
- [38] Executive Order on Preventing Access to Americans' Bulk Sensitive Personal Data and United States Government-Related Data by Countries of Concern[EB/OL]. [2024-3-19]. <https://www.whitehouse.gov/briefing-room/presidential-actions/2024/02/28/>.