

科研机构数据管理体系构建研究 ——来自 NIH 和 CNRS 的经验启示

刘思彤^{1,2}, 游玗怡¹, 陈光³, 温珂^{1,2}

(1. 中国科学院科技战略咨询研究院, 北京 100190;
2. 中国科学院大学公共政策与管理学院, 北京 100049;
3. 中国科学院学部工作局, 北京 100190)

摘要: 对任何一个科研活动主体而言, 如何汇集、存储、共享、开发和利用科学数据, 基于数据构建开放协同的科研组织模式, 已成为迎接数据密集型科学范式到来所面临的现实挑战。与大学和企业相比, 科研机构具有开展科学数据管理的组织优势, 有条件做出先行探索。本文尝试以美国国立卫生研究院和法国国家科学研究中心的科学数据管理工作为对象, 采用案例分析和对比分析的研究方法, 剖析二者围绕数据生命周期建立的数据管理体系特征, 以期为我国科研机构构建数据管理体系提供借鉴。

关键词: 科研机构; 科学数据; 数据生命周期; 数据管理体系

中图分类号: G321 **文献标识码:** A **DOI:** 10.3772/j.issn.1009-8623.2022.06.006

目前, 科学研究正迈入吉姆·格雷 (Jim Gray) 提出的“第四范式”时代——数据密集型科研范式时代。科学数据不再仅仅是研究活动的结果, 而且是科学研究活动的投入要素, 成为重要的科研基础设施。经济合作与发展组织 (OECD) 将科学数据定义为: 科学研究基本来源的实时记录, 包括数值、文本记录、图像和声音, 是科学团体共同接受的对研究结果有用的数据^[1]。我国《科学数据管理办法》规定, 科学数据是指: 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^[2]。科学数据包括科研过程和结果的各项记录, 具有典型的大数据特征: 规模巨大、多源多样和价值待挖掘等。因此, 对任何一个科研活动主体而言, 如何汇集、存储、共享、开发和利

用科学数据, 围绕数据构造开放协同的科研组织模式, 已成为迎接科学研究“第四范式”时代到来所面临的严峻挑战。

科研机构与大学和企业相比, 具有建制化和任务牵引的组织优势, 有条件有机会在科学数据管理体系化建设上做出先行探索。美国国立卫生研究院 (National Institutes of Health, NIH) 作为全球最大的生命医学研究机构, 是最早倡导科学数据管理的组织之一; 法国国家科学研究中心 (Centre national de la recherche scientifique, CNRS) 作为欧洲最大的基础科学研究机构, 也是科学数据管理的先行者之一。身处不同科技体制中的两家科研机构在建设科学数据管理体系中都面临着主体多元、类型多样和促进共享等挑战, 从组织结构的适应性变革到数据标准化建设、分析工具开发, 再到数据安全和质量的管控, 两家综合性科研机构建立科学数据管理体

第一作者简介: 刘思彤 (1998—), 女, 在读硕士研究生, 主要研究方向为科学数据管理。

通讯作者简介: 温珂 (1977—), 女, 研究员, 教授, 主要研究方向为科技创新政策、研发组织与创新系统等。邮箱: wenke@casisd.cn

项目来源: 国家自然科学基金面上项目“知识生产新模式下公立科研机构构建创新网络的动力机制、行为策略与政策激励研究” (71974185)。

收稿日期: 2022-04-08

系的努力为我国提供了有益经验。本文尝试从数据生命周期管理的角度梳理NIH和CNRS的科学数据管理体系现状,总结构建科学数据管理体系的关键机制,以期为我国科研机构加快建设和完善数据管理体系提供借鉴。

1 构建科学数据管理体系的制度和组织准备

美国和欧洲秉持的不同数据监管模式,在NIH和CNRS设计数据管理工作的制度和组织准备中得以体现。NIH的数据监管是以自律为基础的部门模式,CNRS则是在欧盟和法国统一性的数据管理规范内行事。

1.1 制定科学数据管理规划

自主型规划。NIH既是美国生物医学的重要研究机构,也是美国政府最主要的医学研究资助机构,具有国家研究机构和政府科学基金资助组织的双重重要属性。在科学数据管理方面,NIH做出了不少先行实践。2003年,NIH发布了《研究数据共享的最终声明》;2014年,NIH专门就基因组数据管理发布了《基因组数据共享政策》,在保护相关研究者隐私的同时,促进基因组研究数据的临床转化和应用;2018年,NIH制定《数据科学战略计划》,阐述其数据管理的战略目标和实施策略;2020年4月,面对新冠肺炎疫情的肆虐,NIH专门发布了《COVID-19研究战略规划(2020—2024)》,同时启动用于追踪相关神经系统症状的“COVID-19神经系统数据库”(NeuroCOVID)项目,在及时开放共享与COVID-19相关的研究数据和研究成果的同时,倡议并支持科学界基于临床数据情况联合开展战略性研究^[3]。

响应型规划。从科研机构的角度出发,CNRS以实践经验支持国家数据政策的编写,并在国家政策的大背景下,立足于机构使命和愿景提出自身的数据管理发展规划。2016年法国政府颁布《数字共和国法》,其中关于数据开放的相关条例(研究人员有权在较短的开放获取滞后期之后发表由公共资金资助的研究文章等),即是由CNRS结合自身实践支持编写。在法国政府《数字化路线图》(2013年)和《数字共和国法》(2016年)等政

策指导下,CNRS颁布了《开放科学路线图》(2019年11月)及《研究数据计划》(2020年11月)文件,积极响应国家数据管理的政策号召^[4]。

1.2 组织结构的适应性变革

NIH和CNRS都将FAIR原则(可发现 Findability、可访问 Accessibility、可互操作 Interoperability、可重用 Reuse)贯彻到数据战略规划中,在该原则指导下,两家机构均展开了适应性组织创新。NIH先后任命了数据科学副主任和首席数据战略家,并设有数据科学战略办公室(The Office of Data Science Strategy, ODSS)以及科学数据委员会(NIH Scientific Data Council, SDC)^[5]和数据科学政策委员会(NIH Data Science Policy Council, DSPC)两个内部委员会。数据科学战略办公室主要负责领导NIH数据科学战略计划的实施,科学数据委员会和数据科学政策委员会则分别从发展机遇和政策法规方面提供相应的指导建议^[6]。

从2020年开始,CNRS的科学技术信息部(The Department of Scientific and Technical Information, DIST)和数学计算任务部(MiCaDo)合并为开放研究数据部(Open Research Data Department, DDOR)^[7],从事开放科学战略的制定与执行工作,并关注与数据研究相关的所有问题,包括数字基础设施建设等。其中,科学技术信息部主要负责CNRS的数据管理工作落地,包括三个研究单元,分别为科学技术信息研究所(Institute for Scientific and Technical Information, INIST),负责科技信息的获取与传播、分析工具开发;科学交流中心(Centre for Direct Scientific Communication, CCSD),负责开放获取期刊出版物的平台(Hyper Articles Online, HAL)建设工作;Persée^①,负责数字化传播科学历史工作^[8]。此外,CNRS还在筹备新的数据研究部门,主要负责数据开放程度的界定工作。层级式模块化的管理结构设计,让CNRS拥有很强的数据管理执行力。

2 基于数据生命周期的管理体系建设

从过程来看,科学数据管理涵盖了数据的获取、描述、存储、共享和重用等环节,从多源数据

① Persée 通常不翻译为中文,是电子学术期刊法文缩写,最初是一个项目,目前是一个隶属于里昂高师和法国国家科学研究中心的研究支撑单元。感谢中科院科技战略咨询研究院陈晓怡提供此条解释。

产生到汇集数据、对数据进行命名及统一数据格式,再到对数据进行存储并在此基础上进行开放共享和重复利用等等。数据生命周期理论,即是对上述数据管理各环节进行阶段特征分析,提出了链型、矩阵型、环型和层次型等模型^[9-11]。英国国家数据档案馆(UK Data Archive)结合自身管理实践,将数据生命周期界定为六个阶段,包括数据创建、数据处理、数据分析、数据存储、数据访问和数据重用^[12]。国内学者基于对不同科学数据管理实践的案例观察,也分别提出了五阶段^[13-14]、六阶段^[15-16]

的划分,认为收集、保存、处理、共享、分析等是数据生命周期共有属性^[17]。

综合已有数据生命周期理论的相关分析,本文采纳的科学数据管理生命周期模型(见图1),包括获取、描述、存储、共享和重用五个环节。运用这一模型,以下重点比较分析NIH和CNRS的科学数据管理体系现状与特征。

2.1 以规制和补贴等方式多途径汇集科学数据

NIH和CNRS主要通过政策规制和补贴的方式要求或鼓励科学数据的汇交。一方面,对于利用政

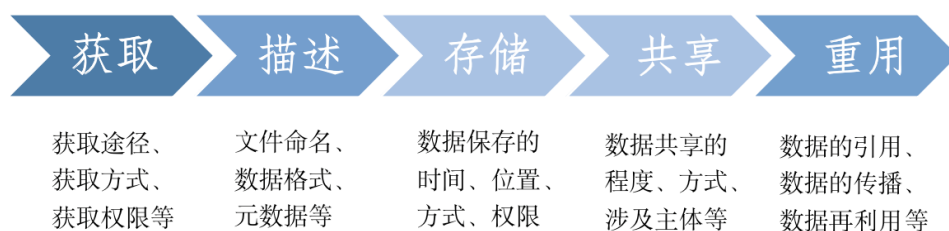


图1 科学数据管理生命周期模型图

府资助产生的研究数据,要求汇交。例如,NIH要求“年度预算超过50万美元”的大额资助项目必须公开研究数据。具体的研究数据范围涵盖用于证明研究发现的、科学界公认的真实数字化资料,不包括样本、实物资料、音频、视频等内容。并且为降低数据共享成本,研究者可以基于自己的数据量选择恰当的共享渠道。例如,针对少量且访问量不高的数据,研究者可以选择“自主提供数据”的方式;而对于访问需求量或数据体量较大的数据,研究人员可以选择将数据提交公共数据库的共享方式^[18]。

CNRS要求接受其资助的研究人员将研究成果在其所属的HAL数据库进行存储和公开。针对可共享的研究数据,研究人员需要提交原始或重新处理的数据的所有格式、文本和文档、软件、算法、协议和工作流情况。为遵循CNRS“尽可能开放,必要时尽可能保留”的开放数据原则,研究人员需要与知识产权事务、数据保护等部门共同确定后续数据的具体开放程度^[19]。

另一方面,对于科学家个人拥有的数据,机构通过补贴或创建交流网络等方式鼓励汇交。例如,NIH鼓励个人、团队、科研机构通过数据平台上传数据,并给予数据提交者一定的补贴。NIH还开发了Eye Wire项目,以游戏的形式联系起130多个

国家的约7万名玩家。玩家可以使用真实的电子显微镜图像绘制老鼠视网膜上神经元的三维结构,“游玩”过程所产生的数据信息可以帮助研究人员发现神经元是如何连接起来处理视觉信息的^[20]。

2.2 建立数据标准化体系

为便于数据存储和共享,对数据管理工作进行质量把控,汇集过程中的标准化工作倍受重视。NIH和CNRS要求从数据类型、应用工具、应用标准等方面对数据进行描述,并将相关信息与数据一并提交。两家科研机构要求数据上传者按照标准内容和格式提交的数据信息如表1所示。

2.3 建设高质量数据库

数据库是数据存储的载体,肩负了“数据中转场”的责任。NIH和CNRS目前均建成了多个数据库,为研究人员提供数据库参考建议,鼓励研究人员将数据存储到合适的高质量数据库。为了更好地统一存储需求,促进数据共享,CNRS在研究者提交数据之始强调了数据存储和数据归档的应用差别,即存储包括数据识别、索引和频繁访问的长期化管理,而归档则是出于法律或历史原因对数据的保存管理。

迄今,NIH已建成涵盖文献、基因、基因组、蛋白质类、化学物质、健康等方面的多个高质量数

表1 NIH与CNRS关于数据信息的提交要求

	NIH ^[21]	CNRS ^[22]
数据类型	科学数据的体量;数据模式(例如成像、基因组、移动、调查);聚合级别(例如个体、聚合、汇总);原始数据的处理方式等	数据标题、创作者、主题、发布者、贡献者、日期资源标识符、数据相关所有权
应用工具	访问和使用科学数据的专用工具和软件等	数据可互操作的工具和系统等
应用标准	科学数据和关联的元数据应用的标准;针对并未发布通用标准的领域内的数据,则需提供其相关的共识性数据标准	元数据应用文档、数据交换协议的规范和标准等

据库^[23],并以需求为导向,根据不同类型的疾病或项目特点,分级分类地增设新的数据库。例如,新冠肺炎爆发时,NIH启动用于追踪相关神经系统症状的“COVID-19 神经系统数据库”项目,旨在从临床医生手中收集与新冠肺炎神经系统症状相关的各类信息,加速研究并发症、疾病情况以及新冠肺炎对已有神经系统疾病的影响^[24]。CNRS则针对不同类型数据的特点,不断探索更优的数据分类存储方式。例如,CNRS正以地理领域为试点,为数据量级较小的“长尾数据”建设通用数据存储库^[25]。

在指导研究人员选择合适的数据库方面,NIH发布了《NIH数据管理和共享政策的补充信息:为NIH支持的研究结果选择数据存储库》政策计划,帮助研究人员高效存储数据,并鼓励研究人员尽可能使用已建立的并且更适合的存储库来保存和共享相应的科学数据,以确保数据的质量和可长期存储性。为帮助研究人员更好地选择数据库进行数据存储,CNRS下属的科学技术信息研究所也在其门户网站中公开了旗下的数据库清单,涵盖法国工程学院的博士论文数据库、集成式书目科学数据库以及超1700万条文献的PASCAL和FRANCIS数据库^[26]。同时,CNRS也正在筹备开放一个更加详细完善的专题数据中心清单。

2.4 开发云平台和数据分析工具促进共享

为促进高效高质量的数据共享,让科学数据创造更大的社会价值,开发云平台成为科学数据共享的重要方式。一方面,云平台可以链接起数据产生方和数据需求方,帮助研究人员快速且无缝访问、使用科学数据,另一方面,云平台也可以大大降低

科学数据的基础设施建设和运维成本。NIH主张使用大规模云计算平台(用于数据存储、访问和计算的共享环境),通过分布式数据存储资源来实现可访问性和规模经济。以NIH Data Commons为例^[27],其主要职责是开发和测试云平台,研究人员可以在该平台上存储、共享、访问生物医学和行为生成的数字对象(数据、软件等),通过数据的便捷共享加速生物医学发现。目前NIH也在与战略伙伴共同努力创建一个可操作的服务平台(PaaS)环境,推动整体的数据生态建设^[28]。

CNRS在欧盟委员会的欧洲开放科学云(European Open Science Cloud, EOSC)计划中积极行动,为研究人员提供共享服务清单,促进国家范围内的云平台建设。CNRS旗下的科学技术信息研究所也在开放科学的目标下,开发建设了Connect Sciences(一个可通过英语、意大利语、法语、西班牙语四种语言进行检索的门户网站),逐步打破数据共享的国别限制,并不断汇集科学技术信息及医学信息等,形成完善的知识云平台^[29]。

因获取和分析数据的技术水平不统一而导致的“技术壁垒”会对数据共享的程度和范围产生较大阻碍。对此NIH和CNRS均开发了帮助研究人员高效挖掘和分析数据的线上工具,并向外界开放获取分析工具的渠道和使用方式,将数据共享主体拓展至非专业性研究的大众用户。数据共享工具的开放不仅可以减少数据污染的情况发生,还可以通过工具的普及有效地降低因技术导致的“数据鸿沟”现象。NIH的国家卫生服务研究和卫生保健技术信息中心(NICHSR)网站会提供数据库和相关统计分析工具包^[30]。CNRS在其官方网站上开放了数

据分析工具 GarganText 以及数据可视化工具 Lodex 等, 帮助研究人员提取数据和进行数据的可视化操作^[31]。为了不断优化数据分析工具, NIH 还设立项目资助私营部门的系统工程师, 不断将原型工具和算法更好地应用至生物医药研究领域, 对现有工具进行改良迭代和优化升级, 如增加更具安全性数据接口等。

2.5 促进数据重用的互动迭代机制

为促进数据重用, NIH 和 CNRS 采取扩大数据再利用范围、提供良好的数据服务等方法, 与数据使用者搭建良好的互动关系, 提升数据重用效率。在扩大数据再利用范围方面, NIH 广泛邀请数据领域专家参与数据科学项目解决方案和计划立项工作。例如, 数据科学战略办公室启动数据和技术进步国家服务学者计划, 参与计划的数据科学家和计算机工程师可获得公共健康领域中的生物医学数据; 数据科学战略办公室同时提供相关的潜在生物医学问题, 以促成多领域的科学家共同解决“如何加速人工智能在医学成像中的临床应用”等问题^[32]; 数据引用规范方面, 除了发布相关引用规范文件供研究人员参考外, 隶属 NIH 的美国国立医学图书馆 (National Library of Medicine, NLM) 正将一些新的索引办法发表在数据期刊上, 如《科学数据》, 帮助研究人员更好地掌握数据索引的使用^[33]; 数据服务方面, CNRS 的科学技术信息研究所会针对用户需求, 提供领域及数据专家的数据监测和提取服务, 帮助用户整理及总结所需的数据信息。科学技术信息研究所还会聚焦当前的热门研究话题, 如禽流感 and 人类、生物技术和药品等主题, 在官方网站上将上述话题的评论文件进行集中发布, 以推动交叉领域内学者的研究互动。

3 运行数据管理体系的关键要素分析

在建立数据生命周期五阶段数据管理工作的基础上, NIH 和 CNRS 还通过与外部多主体的互动合作, 拓展数据存储和共享工作的边界, 加大基础实施建设和人才培养投入, 重视数据安全隐患问题, 稳定支持数据管理工作的开展。这些举措为确保数据管理体系的高效运作提供了进一步的保障, 激发了机构数据生态的活力。

3.1 推动多主体合作

3.1.1 与其他机构合作, 不断探索数据交互新机制

在数据管理流程的各阶段, 积极推进数据的交互十分重要。如数据收集阶段, NIH 与各类组织机构合作, 促进数据协同以解决疾病难题。为改善对致病细菌和食源性疾病的监测, NIH 与疾病预防控制中心 (Centers for Disease Control and Prevention, CDC) 和美国食品药品监督管理局 (Food and Drug Administration, FDA) 合作, 实施了病原体检测项目和食品与饲料安全基因组学跨机构研究项目 (Gen-FS)。通过该项目, 美国和国际上的许多公共卫生机构从食物、环境和人类患者中收集样本, 并将获得的细菌病原体的基因序列数据提交至 NIH^[34]。类似的还有结核病门户项目, 负责耐药结核病的临床医生和科学家组成联盟, 与数据科学家和信息技术专业人员合作, 收集多领域的结核病数据, 并向临床和研究界提供这些数据。在数据存储的互联互通方面, 由于拥有的数据库体系庞大, NIH 致力于加强整合, 改进知识库和数据库的互操作性。其以美国国家生物技术信息中心 (National Center for Biotechnology Information, NCBI) 为中介, 在云平台与目前广泛使用的 NIH 数据库之间建立连接。

在科学技术信息研究所的主导下, CNRS 与法国高校的图书馆和文献中心达成了合作关系。例如, 在与法国高等教育编目中心 (ABES) 的良好合作下, 科学技术信息研究所能够实现与大学体系资料库 (SUDOC) 的互通, 构建了馆际互借的良好互动体系。同时, 科学技术信息研究所在与国际上重要科技信息机构的合作方面投入了大量精力, 目前已经和英国图书馆、加拿大科技情报所以及德国文献服务系统 Subito 达成合作协议。与此同时, CNRS 积极探索与美国计算机联图书馆中心 (Online Computer Library Center, OCLC)、美国剑桥科学文摘 (ProQuest-Cambridge Scientific Abstracts, ProQuest-CSA) 的合作路径。如若科学技术信息研究所能促成与上述机构的合作, 格式化的文献就可以通过 FTP 或电子邮件实现订单的跨国传递, 或者以机器可读文件的形式交换目录。

3.1.2 与公众合作, 开放科学边界

一方面, 激励公众为数据库提供数据。NIH 积

极招募志愿者以及通过补贴鼓励公众提交健康信息等数据,以便为研究人员提供研究资源。每年有近3 500名健康志愿者参与NIH的研究^[35]。除此之外,每天有超过3 000个不同的团体和个人通过美国国家生物技术信息中心系统上传数据,数据包括人类和研究生物的基因组序列、基因表达数据、化学结构和性质(安全性和毒性数据)、有关临床试验及其结果的信息等。许多个人和团体,如联邦机构、出版商、州公共卫生实验室、基因检测实验室、生物技术和制药公司等,积极主动地为生物医学研究数据生态贡献数据。

另一方面,帮助公众了解科学数据和支持科学研究。如CNRS开发了崭新的交互式数字媒体Dorandum软件并在其官网发布,通过远程培训的方式帮助公众了解数据管理计划和共享方面的知识,并不定期地开展数据知识研讨会^[36]。公众可以在网站自由报名,与嘉宾共同探讨数据管理的相关内容。

3.2 保障对基础设施建设的资金投入

NIH的《数据科学战略计划》明确提出要支持建设高质量的生物医学研究数据基础设施、实现数据资源生态系统的现代化。2020年,NIH请求增加1亿美元投资内部的信息技术基础设施,保障数据隐私安全,以及不断研发和更新数据处理、共享、分析的工具和方法等。

为促进数据信息共享,CNRS同样为基础设施建设投入大量心血。2019年,CNRS为下属的法语学术文献开放网站,即HAL项目,额外拨款65万欧元予以支持,通过研究工具的改进增加HAL存储量,加强与其他国际开放档案库的互操作性等。同时,CNRS正在筹备一份基础研究的设备规章,以扩大FAIR原则在所有学科中的应用,并承诺所有的基础研究和数据存储设备均将采用FAIR惯例和质量标准。

3.3 重视数据安全和隐私问题

当前,各管理主体在推进数据开放共享的实践中,都遭遇了数据安全以及隐私保护等挑战。通过对NIH的“注重数据开放的前提”和CNRS的“尽可能开放,必要时尽可能不开放”等相关数据政策进行梳理,可以发现二者对数据安全问题的重视。NIH积极探索通过技术升级等方式尽可能地保证数

据获取过程的安全性。如鼓励研发人员开发和采用更适合移动设备与数据接口的工具,确保该信息工具可以获得相关认证,以及认证的电子健康记录和其他临床数据能够安全合法地应用于医学研究等。NIH特别重视隐私保护,要求促进基因组研究数据的临床转化和应用必须是在保护相关研究测试人员隐私基础上进行下一步研究。

比较而言,CNRS由于涉及领域更加广泛、不同学科领域之间存在异质性和复杂性,因此更多的是仅做出原则性规定,如科学成果需要在不挑战个人数据或知识产权保护的情况下获取和公开等。而对于数据隐私安全和知识产权的归属并未给出统一界定,号召各领域形成各自的具体要求规范。

3.4 培养管理人才,保持管理体系活力

随着数据与其他领域交叉的问题涌现,科研机构也在不断将目光聚焦于数据科学的人才培养和队伍建设。人才招聘方面,NIH启动了“数据研究员计划”等项目,为积极建设数据科学人才队伍提供支撑。NIH主要采用数据驱动研究的理念招聘相关背景的科研人员,并将招聘的数据科学家和其他在项目管理等领域有专长的人纳入NIH的一系列数据科学项目,比如“All of Us”项目等,通过人才的知识多样化增强项目研究专业性^[37]。CNRS数据管理部门的管理层人员也是领域内具有数据类专业背景的管理人才。同时,为打破社会对数据科学领域女性研究人员的刻板印象,CNRS在网站上专门发布了12幅女性数字科学研究人员的肖像和漫画^[38],分享优秀女性数据科学研究人才背后的故事,为实现数字科学研究的多样性做出了巨大努力。

人才培养方面,CNRS凭借持续积累的文献加工与数据库管理的丰富经验,为有意愿提升信息检索和数据管理方法的相关研究人员或者信息专业学者提供上述内容的培训课程^[39],旗下的科学技术信息研究所还为此专门建设了一个线上的培训网站,以便研究人员进行线上学习。

人才评价方面,CNRS大力倡导对数据研究人员的评价方式改革。考虑到目前主要是通过文献计量的方式进行评估,CNRS签署了《数字运营弹性法案》(Digital Operational Resilience Act, DORA),承诺机构各部门采用更定性的评估方式,并且在评估时应考察各种类型的研究成果。

4 对我国科研机构建设数据管理体系的启示

在开放共享的目标下, NIH 和 CNRS 基于数据生命周期的管理体系和开放式数据生态系统建设有力推动了科学数据的流动和价值创造, 为我国科研机构的数据管理工作提供了有益启示。

(1) 重视顶层设计, 建立基础制度和组织管理架构。

在建设数据管理体系之初, 首先要做好顶层设计, 明确机构数据管理工作的原则和定位, 如 NIH 和 CNRS 始终坚持 FAIR 原则, 并强调要构建开放共享的数据管理体系。聚焦战略目标, 科研机构应结合领域数据管理的特征制定相应的管理政策, 统筹规划数据管理工作的层级和要素, 且要有专业的数据管理领导团队牵头推进数据管理工作。科研机构的各部门需要展开相应的数据管理流程建设, 并加强部门之间的互联互通。领导团队与各部门数据管理负责人之间紧密合作, 形成科研机构数据管理的基本组织架构。

(2) 建设专业数据库, 多主体合作构建科学数据库网络。

在业务范围内, 科研机构应着力聚焦建设领域数据库。依托领域数据库, 再逐步拓展至交叉领域的数据库链接, 如 NIH 以美国国家生物技术信息中心为中介, 连接起关联领域的数据库和数据资源, 为数据的共享增加可操作性和便捷性。在建设数据库的过程中, 科研机构要增进跨领域多主体间的合作交流, 为后续的数据资源流动和共享夯实基础, 最终参与到更大范围的科学数据库网络建设。

(3) 技术和管理并重, 注重科学数据安全和标准化工作。

在技术上, 首先要重视数据分析和工具的开发。科研机构应加大数据库软件研发投入, 开发数据检索、分析等工具, 在使用中扩大数据规模并迭代数据服务, 促使科学数据价值流动替代科学数据流动。其次, 加强区块链等技术在科学数据保密和隐私保护等方面的应用, 为安全前提下的科学数据开放共享提供技术支撑。

在管理上, 重视科学数据标准化工作, 着力搭建科学数据分级分类管理制度体系, 出台科学数据安全使用各项规定。科研机构应制订和完善科学数

据提交、描述等的标准格式, 明确访问控制要求。科研机构应着力推进建立科学数据分级分类管理制度, 为推进科学数据安全使用和共享提供制度基础。依据国家数据安全管理相关规定, 积极探索制定科学数据安全管理的职责和程序, 形成具有可操作性的实施范例。

(4) 加强数据管理人才培养, 完善人才成长激励制度。

科研机构应重视科学数据管理人才培养, 给予研究项目支持和人员培养培训等机会, 为人才成长提供实践土壤。研究制定适用于科学数据管理人才的岗位设置与晋升办法, 推动出版科学数据论文纳入职称晋升和工作绩效等评价内容, 畅通数据管理人才的职业发展路径。建立有竞争力的薪资管理制度, 吸引具有领域知识背景和信息化管理技能的复合型人才积极投身科学数据事业。■

参考文献:

- [1] Pilat D, Fukasaku Y. OECD principles and guidelines for access to research data from public funding[J]. Data Science Journal, 2007, 6(0): OD4-OD11.
- [2] 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 - 国办发〔2018〕17号 [EB/OL]. (2018-04-02) [2022-04-06]. http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm.
- [3] Trans-NIH BioMedical Informatics Coordinating Committee. NIH data sharing policies[EB/OL]. (2022-04-05)[2022-04-06]. https://www.nlm.nih.gov/NIHbmic/nihd_data_sharing_policies.html.
- [4] Centre National de la Recherche Scientifique. Sharing scientific information[EB/OL]. (2017-12-01)[2022-04-06]. <https://www.cnrs.fr/en/sharing-scientific-information>
- [5] The National Institutes of Health. About the council[EB/OL]. (2021-05-25)[2022-04-06]. <https://datascience.nih.gov/scientific-data-council>.
- [6] The National Institutes of Health. About the office of data science strategy[EB/OL]. (2019-08-07)[2022-04-06]. <https://datascience.nih.gov/about/odss>.
- [7] Centre National de la Recherche Scientifique. CNRS Research Data Plan[M]. Paris: CNRS IFSeM, 2021: 1-13.
- [8] Open Research Data Department. The actors of open

- science at the CNRS[EB/OL]. [2022-04-06]. <https://www.science-ouverte.cnrs.fr/ddor-cnrs-direction-des-donnees-ouvertes-de-la-recherche/>.
- [9] 师荣华,刘细文.基于数据生命周期的图书馆科学数据服务研究[J].图书情报工作,2011,55(1):39-42.
- [10] 李伟绵,崔宇红.研究数据管理生命周期模型及在服务评估中的应用[J].情报理论与实践,2015,38(9):38-41.
- [11] 索传军,王涛,付光宇.国内外信息生命周期管理研究综述[J].图书馆杂志,2008(7):14-20.
- [12] UKDA. Research data life cycle[EB/OL]. (2018-10-29)[2022-04-06]. <https://www.ukdataservice.ac.uk/manage-data/lifecycle>.
- [13] 黄源,施栩婕,李晨英.基于科学数据管理流程的科研机构职责分析[J].数字图书馆论坛,2020(1):20-26.
- [14] 魏悦,刘桂锋.基于数据生命周期的国外高校科学数据管理与共享政策分析[J].情报杂志,2017,36(5):153-158.
- [15] 鲍静,范梓腾,贾开.数字政府治理形态研究:概念辨析与层次框架[J].电子政务,2020(11):2-13.
- [16] 陈恬,余亚荣,张照余,等.基于数据保全思想的科学数据全流程管理研究[J].档案与建设,2020(12):4-9.
- [17] 杨林,钱庆,吴思竹.科学数据管理生命周期模型比较[J].中华医学图书情报杂志,2016,25(11):1-6.
- [18] 汪俊.美国科学数据共享的经验借鉴及其对我国科学基金启示:以NSF和NIH为例[J].中国科学基金,2016,30(1):69-75.
- [19] Centre National de la Recherche Scientifique. Research data plan[EB/OL]. [2022-04-06]. https://www.science-ouverte.cnrs.fr/wp-content/uploads/2021/04/Cnrs_Research-Data-Plan_mars21.pdf.
- [20] The National Institutes of Health. NIH 3D print exchange[EB/OL]. [2022-01-04]. <https://3dprint.nih.gov/collections/neuroscience>.
- [21] The National Institutes of Health. NIH releases strategic plan for data science[EB/OL]. (2018-06-04)[2022-01-04]. <https://www.nih.gov/news-events/news-releases/nih-releases-strategic-plan-data-science>.
- [22] Par l'Atelier Données. Guide de bonnes pratiques de gestion des données de la recherche[EB/OL]. (2021-12-07)[2022-04-06]. <https://mi-gt-donnees.pages.math.unistra.fr/guide/03-collector.html>.
- [23] 唐志立.NCBI所有数据库简介[EB/OL]. (2019-08-07)[2020-10-06]. <https://www.renrendoc.com/paper/97254776.html>.
- [24] 张唯.新冠肺炎影响神经系统 美国NIH启动数据库项目追踪症状[EB/OL]. (2021-01-07)[2022-01-04]. <https://baijiahao.baidu.com/s?id=1690038409231503823&wfr=spider&for=pc>.
- [25] Centre National de la Recherche Scientifique. Roadmap[EB/OL]. [2022-04-06]. https://www.science-ouverte.cnrs.fr/wp-content/uploads/2019/11/CNRS_Roadmap_Open_Science_18nov2019.pdf.
- [26] 张志刚.法国国家科技信息研究所发展现状和思考[J].数字图书馆论坛,2009(12):38-42.
- [27] The National Institutes of Health. New models of data stewardship[EB/OL]. (2019-05-03)[2022-01-04]. <https://commonfund.nih.gov/commons>.
- [28] The National Institutes of Health. NIH releases strategic plan for data science[EB/OL]. (2018-06-04)[2022-01-04]. <https://www.nih.gov/news-events/news-releases/nih-releases-strategic-plan-data-science>.
- [29] 宋筱璇,王延飞,钟灿涛.国内外科研数据安全政策比较研究[J].情报理论与实践,2016,39(11):10-16.
- [30] The National Institutes of Health. National information center on health services research & health care technology[EB/OL]. (2021-10-21)[2022-01-04]. https://hsric.nlm.nih.gov/hsric_public/topic/datasites/.
- [31] Centre National de la Recherche Scientifique. Exploration des données[EB/OL]. [2022-01-04]. <https://www.science-ouverte.cnrs.fr/exploration-des-donnees/>.
- [32] The National Institutes of Health. Office of data science strategy: 2020 year in review[EB/OL]. (2021-01-20)[2022-01-04]. <https://datascience.nih.gov/news/office-data-science-strategy-2020-year-review>.
- [33] The National Institutes of Health. National institutes of health plan for increasing access to scientific publications and digital scientific data from NIH funded scientific research[EB/OL]. (2015-02-01)[2022-01-04]. <https://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf>.
- [34] 药明康德. \$2.15 亿! NIH 与 11 家药企携手推进免疫

- 疗法[EB/OL]. (2017-10-13)[2022-01-04]. https://med.sina.com/article_detail_100_2_34782.html.
- [35] The National Institutes of Health. All of US research program[EB/OL]. [2022-01-04]. <https://allofus.nih.gov/>.
- [36] Centre National de la Recherche Scientifique. Formation-Ouverture des données de la recherche[EB/OL]. [2022-01-04]. <https://www.science-ouverte.cnrs.fr/formation-science-ouverte/formation-ouverture-des-donnees-de-la-recherche/>.
- [37] The National Institutes of Health. Big data integration to better health for all of us[EB/OL]. (2021-02-22)[2022-01-04]. <https://datascience.nih.gov/big-data-integration-better-health-all-us>.
- [38] Centre national de la recherche scientifique. Les décodeuses du numérique[EB/OL]. [2022-01-04]. <https://www.ins2i.cnrs.fr/fr/les-decodeuses-du-numerique>.
- [39] Centre National de la Recherche Scientifique. Action nationale de formation|Exploration documentaire et extraction d'information[EB/OL]. (2021-07-06)[2022-04-06]. <https://www.science-ouverte.cnrs.fr/actualite/action-nationale-de-formation-exploration-documentaire-et-extraction-dinformation/>.

Research on Data Management System Construction of Research Institutes: Experience from NIH and CNRS

LIU Si-tong^{1,2}, YOU Ding-yi¹, CHEN Guang³, WEN Ke^{1,2}

(1. Institutes of Science and Development, Chinese Academy of Sciences, Beijing 100190;

2. School of Public Policy and Management, University of Chinese Academy of Sciences, Beijing 100049;

3. Working Bureau of Chinese Academy of Sciences, Beijing 100190)

Abstract: For any scientific research entity, how to identify, store, share, use and protect scientific data, and construct data-based, open and collaborative research and innovation mode has become the biggest challenge to meet the arrival of data-intensive science era. Compared with universities and enterprises, scientific research institutions have the organizational advantages to carry out scientific data management. This paper takes the scientific data management work of National Institutes of Health and Centre national de la recherche scientifique as the object, and uses case analysis and comparative analysis to analyze the characteristics of the data management system on the basis of data life cycle, in order to provide reference for the construction of data management system in China's scientific research institutes.

Keywords: research institutes; scientific data; data life cycle; data management system