

我国科学数据管理相关政策解读与人口健康科学数据管理的启示^{*}

王安然 吴思竹 钱 庆 李赞梅 钟 明

(中国医学科学院/北京协和医学院医学信息研究所 北京 100020)

〔摘要〕 解读我国发布的两项科学数据管理政策《科学数据管理办法》和《国家科技资源共享服务平台管理办法》，阐述其中关于我国科学数据管理和国家科学数据中心建设的要点，分析人口健康领域科学数据管理现状以及存在的问题，对后续人口健康科学数据管理工作提出建议。

〔关键词〕 数据管理政策；科学数据管理平台；科学数据中心；数据共享

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2019.12.001

Interpretation of Relevant Policies of Scientific Data Management and Enlightenment of Scientific Data Management of Population Health in China WANG Anran, WU Sizhu, QIAN Qing, LI Zanmei, ZHONG Ming, Institute of Medical Information, Chinese Academy of Medical Sciences/Peking Union Medical College, Beijing 100020, China

〔Abstract〕 The paper interprets the *Scientific Data Management Methods* and *Management Methods for National Sharing Service Platform of Scientific and Technological Resources*, expounds the key points on scientific data management and building of the National Center for Scientific Data in China, analyzes the present situation of and existing problems with scientific data management in population health, and puts forward suggestions on subsequent scientific data management of population health.

〔Keywords〕 data management policy; scientific data management platform; scientific data center; data sharing

〔收稿日期〕 2019-02-20

〔作者简介〕 王安然，硕士，实习研究员，发表论文 10 篇；通讯作者：吴思竹，博士，副研究员，发表论文 40 余篇。

〔基金项目〕 国家人口与健康科学数据共享服务平台标准宣贯实施与推广后补助项目（项目编号：NC-MI-MCE02-201801）；国家重点研发计划精准医学研究重点专项“重大疾病精准医学数据库群”课题（项目编号：2016YFC0901602）。

1 引言

近年来随着开放科学的不断发展，科学数据已成为世界多个国家的重要战略资源。在各类开放数据政策与规划的指引下，政府、科研院所、企业、国际组织等都积极参与科学数据资源建设，开展各类研究与项目。英、美等发达国家走在科学数据资源建设与共享的前列，发布多项政策支持资源建设与开放共享^[1-3]。2018 年我国先后发布两个关于科学数据的管理办法。2018 年 1 月中央全面深化改革领导小组第 2 次会议审议通过《科学数据管理办法》并于同年 3 月 17 日由国务院办公厅正式印

发^[4]。《科学数据管理办法》首次从国家层面明确要求各相关部门应围绕国际发展战略和科技创新需求, 统筹规划和建设本行业的科学数据中心。同年 2 月科技部、财政部印发《国家科技资源共享服务平台管理办法》^[5], 以规范管理国家科技资源共享服务平台, 深入发展科技创新, 推动数据共享。这一系列举措可见我国对科学数据管理的重视。

本文首先对《科学数据管理办法》和《国家科技资源共享服务平台管理办法》这两项政策进行分析, 阐述两项管理办法针对科学数据管理的要点, 同时简要分析我国人口健康领域科学数据管理现状, 为我国人口健康科学数据管理相关工作提出建议。

2 科学数据管理政策要点

2.1 建立多层次的数据管理机制

长期以来, 国外针对科学数据共享实行“政府先行”的政策, 制定一系列的科学数据开放共享政策法规^[6], 而我国在数据共享方面起步较晚, 在科学数据管理政策法规方面有一定的缺失。两项《办法》的发布首次从国家层面建立关于科学数据的管理政策, 从多方面对科学数据管理进行规范。《科学数据管理办法》规定由国务院科学技术行政部门负责全国科学数据的统筹和宏观管理; 国务院和省级人民政府相关部门等作为主管部门协助分工负责; 由主管部门委托有条件的科研院所等法人单位成立科学数据中心, 承担科学数据的汇交、管理、共享和安全保障等任务; 同时确定法人单位的主要责任以及科学数据中心的地位与职责。《国家科技资源共享服务平台管理办法》也分别对国家宏观管理部门、主管部门和平台依托单位的主要职责进行详细阐述。两项《办法》均明确指出宏观管理部门、主管部门、法人单位、科学数据中心等各级部门的职责, 使各级责任有了明确界定。这种由国家主导、层层分级负责的责任制度弥补现今我国科学数据管理与应用的不足。

2.2 组建国家级科学数据管理平台

为推动我国科学数据的开放共享, 《科学数据

管理办法》指出应统筹规划和建设科学数据中心, 《国家科技资源共享服务平台管理办法》更加细致地阐述国家平台(国家平台将统一规范命名为“国家 XX 科学数据中心”)的建立、管理办法, 明确说明宏观管理部门、主管部门、依托单位等各层级部门在平台建设中的具体管理责任。规定宏观管理部门应制定相关政策规范, 统筹协调国家平台建设并开展考核工作; 主管部门应制定部门规划标准, 推动国家平台规范建设; 依托单位作为责任主体应建立国家平台标准规范, 保障国家平台运行管理。

2.3 施行强制性的数据汇交政策

近年来我国对科研创新的投入不断增加, 同时也产生大量有质量的科学数据。但由于没有统一标准对各项目科学数据进行管理整合, 数据无法有效共享、利用, 造成科学数据的严重浪费。《科学数据管理办法》明确指出主管部门应建立科学数据的汇交制度, 开展科学数据汇交工作。其第 13 条要求由政府预算资金资助的各级科技计划项目所形成的科学数据都要汇交到科学数据中心并在汇交后由科学数据中心出具汇交凭证; 要求各级科技计划管理部门实行先汇交再验收的制度并对全生命周期的数据进行汇交。《国家科技资源共享服务平台管理办法》也规定由财政性资金资助的各类科技项目的科技资源要汇交到平台。两项《办法》从各方面详细阐述科学数据的强制性汇交政策, 以确保科学数据的整合保存, 加强数据积累。

2.4 秉持数据开放共享原则

科学数据的开放共享早已是国际共识。长久以来, 我国科学数据管理工作缺乏清晰的管理机制, 各大高校、科研院所之间不能有效沟通科研进展, 科学数据无法共享和重复利用, 严重制约科学发展。《科学数据管理办法》规定我国科学数据的开放共享要秉持“开放为常态、不开放为例外”的原则, 实现最大限度的开放。除国家法律法规有特殊规定的情况外各类科学数据应与国家共享交换平台进行及时交接并面向公众开放共享, 实现科学数据的最大化利用。同时两项《办法》均指出法人单

位和国家平台等应为政府决策、公共安全、国防建设、环境保护、防灾减灾、公益性科学研究等无偿提供科学数据,以充分发挥科学数据的重要作用。

2.5 保障数据安全和知识产权

大数据时代科学数据成为各个国家重要的战略资源,同时数据的安全问题也接踵而来。国外陆续推出诸多针对信息网络安全建设的战略举措^[7],我国新出台的两项《办法》也将保证数据安全放在首要位置。《科学数据管理办法》对涉及国家秘密、国家安全、公共利益等科学数据的开放方式做了原则性、政策性规定,各级部门应建立健全相关制度,做好数据的保密审查,加强数据监管。两项《办法》还强调知识产权保护的重要性。《科学数据管理办法》要求科学数据的使用者对使用科学数据的来源和使用情况进行明确标注,保护数据提供者的权益。《国家科技资源共享服务平台管理办法》也规定国家平台应建立符合国家知识产权保护和保密等有关规定的制度,保护科技资源提供者的知识产权和利益。

2.6 建立分级评价考核制度

在平台监督评价方面,英国于2012年发布《开放数据白皮书》,建立一套对公共部门开放数据程度的评价体系,对各公共部门完成开放数据任务情况进行审计,以促进英国公共服务数据的开放性。我国《科学数据管理办法》中也提到主管部门和法人单位应建立完善的评价考核制度。《国家科技资源共享服务平台管理办法》详细阐述评价考核的具体措施。宏观管理部门对国家平台进行两年一次的分类评价考核;主管部门对所属国家平台进行年度自评;宏观管理部门还将委托国家科技基础条件平台中心对国家平台进行评价考核,最终根据考核结果动态调整国家平台的运行。建立这种系统、体现导向性的评价体系有利于规范各级参与方的职责和行为。将评价结果与绩效考核、经费申请、奖励相结合,建立激励机制,有利于规范科学数据的建设和运行。

3 人口健康领域的科学数据管理现状

3.1 国外

国外针对人口健康领域的科学数据管理建设起步较早,已开展大量工作,取得一定经验。欧美等国家在政府的引导下已建立大批具有国际知名度和影响力的数据中心和数据平台。早在19世纪80年代,美国成立国立生物技术信息中心(NCBI),负责收集、存储生物医学文献资料以及生物分子序列、结构数据^[8]。欧洲分子生物学实验室建立核酸序列数据库并于1992年成立欧洲生物信息研究所(EMBL-EBI)^[9]。2007年哈佛大学定量社会科学研究所数据科学团队研发Dataverse研究数据管理系统,已被多个研究机构采用^[10]。2015年美国、欧洲和日本共同建立国际核苷酸序列数据库联盟(INSDC)^[11],为世界范围内的研究人员提供高效、便捷的生物信息资源获取方式。2017年12月美国国立医学图书馆(NLM)发布2017-2027战略规划——生物医学发现和数据驱动健康平台^[12],培育数据驱动团队,加强数据研究与传播,加速研究发现。

3.2 国内

我国人口健康领域科学数据管理建设与发展主要受政府财政支持,相比国外起步较晚,但多年来已开展大量工作。2002年我国启动国家科学数据共享工程,人口健康领域是最早的试点项目之一,目前建成的国家人口与健康科学数据共享平台已投入使用。“十二五”期间我国卫生系统初步建立全国人口信息、电子健康档案、电子病历等数据库,全国已建立27个省(区、市)级人口健康信息平台。2016年我国将精准医疗计划列入国家“十三五”科技发展重大专项并上升为国家战略。专项要求构建国家级的精准医学大数据管理共享技术平台,用于汇集各组织、各部门的医学大数据。我国现有的人口健康科学数据管理平台多通过政府主导资助,由高校、科研院所等机构平台构建。国家人口与健康

科学数据共享平台^[13]是首批由科技部支持建设的国家级科学数据中心,项目的总体目标是建立一个物理上分布、逻辑上高度统一的医药卫生科学数据管理与共享服务系统。目前平台已集成 20 多家单位的近 800 个数据集,共享资源总量超 200 亿条、资源容量达 67TB。2015 年中国科学院北京基因组研究所开发并构建组学原始数据存储归档系统(GSA)^[14],专注于组学原始数据收集与整合并提供免费的数据存储、共享与访问服务。该系统是国内首个被国际期刊认可的组学数据发布平台,可接收世界各国的科研数据。北京大学和北京大学医学部构建多个健康医疗领域的科学数据管理平台,如中国健康医疗数据共享平台^[15]、中国队列共享平台^[16]、中国肾脏疾病数据网络^[17]。这些平台旨在整合健康医疗领域的数据资源,实现数据共享,推动健康医疗大数据领域关键共性问题的解决。此外,社会力量也积极参与我国人口健康领域数据资源建设。创办于 1999 年的华大基因目前已发展成为全球最大的基因组学研发机构,创建并汇聚海量生物医学数据,为我国人口与健康领域科学数据资源建设做出重大贡献。2012 年华大基因创办开放型的在线期刊 *GigaScience*^[18],其不仅提供文献全文,还允许用户直接通过文献所提供的数据和分析工具对结果进行测试和验证,实现数据的透明、公开及可重现性。

3.3 存在的问题

近年来我国对人口健康领域的科学数据管理十分重视,陆续出台一系列法规政策,将科学数据管理提升到战略水平,针对管理平台的建设工作也取得很大进步。但是与国外科学数据管理平台和相应政策、标准相比还存在一定的问题。第一,我国现有人口健康领域科学数据管理的数据来源较为固定,多来源于高校自身或下级数据中心的科研产出,有待进一步扩展更多资源渠道,提升数据资源发现能力,实现多方资源联合,合作共赢。第二,人口健康科学数据管理平台的管理内容多以科学数据存储和共享为主,主要对科研项目中后期产生的结果数据、文档资料进行收集整理和发布,并未实

现整个生命周期的科学数据管理。第三,缺乏人口健康科学数据汇交和共享的标准,导致现有开放数据的整体质量和可复用性都亟待提高。需进一步改善科学数据管理的总体环境,完善政策法规,统一数据标准。第四,我国数据共享建设正处于起步阶段,在国际的影响力还十分欠缺,尚未建成具有国际竞争力的科学数据中心,导致我国人口健康领域的科学研究数据多流向海外的权威数据库,科学数据流失情况严重。第五,我国科学数据管理起步较晚,缺少针对科学数据管理方面的人才培养,研究人员的数据安全、数据道德和知识产权意识相对薄弱。

4 启示与建议

4.1 注重发挥政府主导作用

政府占据着大量的数据资源,因此也常常成为科学数据资源建设与共享的主体。我国人口健康领域科学数据管理平台建设主要是由政府主导,以及由政府引导高校、科研院所、企业等进行一系列合作,大部分资源建设活动通过政府投资、项目驱动的形式进行。以往由于相关政策法规的缺失,各大资源占有单位的共享意识不足,导致资源封闭且分散、影响力明显不足。随着两项《办法》的推出,在后续科学数据管理平台的建设中应注意进一步加强国家政策引导,增强政府的主导力度,建立强有力的全国性、跨行业的资源建设与管理协调机制。具体包括政府应对每一层级部门(宏观管理部门、主管部门、科学数据管理平台、法人单位等)的管理责任做出具体指示,使各层级能够进行更有效的工作对接。针对发放科研项目的宏观管理部门和主管部门,在项目申报初期应向科研机构等法人单位落实具体的科学数据汇交标准并明确指出接收该项目数据的科学数据管理平台。同时各层级应建立良好的关联,宏观管理部门对平台总体工作进行统筹规划,主管单位对法人单位的数据工作进行相应评价考核。

4.2 加强各级单位共享合作

我国人口健康领域的科学数据管理平台一般由

各高校、科研院所等进行独立建设,如北京大学和北大医学部的中国健康医疗数据共享平台,中国科学院北京基因组研究所的组学原始数据存储归档系统等。而这些管理平台通常只针对院校内的科研项目产出数据进行接收和管理,缺乏合作伙伴的参与,各层级的共享合作观念十分淡薄。两项《办法》中强调我国科学数据的开放共享应秉持“开放为常态、不开放为例外”的原则,因此在后续平台建设中应加强科学数据管理平台与各单位间的互联互通。平台主管部门应组织发布科学数据资源目录,促进数据共享和重复利用。同时应加强区域内、系统内的共享合作,以及科研院所、图书馆、信息技术公司等机构之间的合作,综合利用各单位资源、服务、人才和技术优势打造更好的资源建设环境与网络。

4.3 建立全生命周期数据管理原则

《科学数据管理办法》指出科学数据的管理应覆盖到科学研究的整个过程。英国的 Data Archive 将科学数据时代生命周期分成 6 个阶段,即数据创建、数据处理、存储数据分析、数据存储、数据共享和数据重用^[19]。国外很多数据仓储平台要求科研人员从研究开始就提交数据管理计划,对全生命周期的科学数据进行管理。而我国目前的人口健康科学数据管理重心多为数据汇交、存储和共享。建议国内科学数据管理平台建设过程中应针对数据计划、生产、采集、汇交、加工、存储、保存、分析、出版、再利用、安全、更新等具体环节制定详尽的数据管理标准,使科研项目与科学数据管理平台能互通互认,数据汇交流程能够透明化,项目组能通过科学数据管理平台跟踪数据的汇交状态,合作完成全生命周期的数据管理。

4.4 加强平台标准化建设

由于标准的不统一,不同单位的数据资源采用的数据标准也不尽相同,导致我国人口健康领域的科学数据整合难度较大,严重影响科学数据的共享和复用。《科学数据管理办法》提出应制定国家科学数据管理政策和标准规范,规定各法人单位的科

学数据生产者应按照相关标准规范进行科学数据的生产、采集、整理和加工,使生成的数据库或数据集便于统一管理,使数据可利用、可分析。在后续平台建设中应建立包括资源唯一标识符、元数据规范、知识组织规范、长期保存规范、数据质量控制、资源评价等在内的能够适应新环境下大数据服务需求的资源标准规范体系。重点加强数据加工、描述、长期存储与服务类的标准规范建设,促进资源在更深层次、更广范围内的共享、互操作和复用,实现科学数据的价值最大化。

4.5 构建平台综合评价体系

国外大型科学数据管理平台已建成成熟的评估考核体系,而我国人口健康领域的科学数据管理平台缺乏一体化的统筹管理,也缺乏体系化、行之有效的量化考核指标。两项《办法》也提出应对数据中心建立完善的评价考核制度。在后续建设中,应以资源管理视角,建立相对科学、客观且具备一定可操作性的平台资源评价指标体系。融合多维度定量定性指标,全面反映资源主题的代表性、权威性 & 行业影响力、资源质量、规模、服务能力及服务成效等内容。基于平台资源评价体系,持续开展数据质量评估,发布年度评价报告,使平台资源评价常态化。

4.6 加强专业人才培养

科学数据管理与共享的各个环节都离不开专业人员的操作,但我国对专业数据管理和共享方面的人才培养重视度并不足够。因此大力加强科学数据管理专业人才培养队伍建设势在必行。国内高校、科研院所、相关企业都应加强培养研究人员的数据素养,为研究人员树立正确的科学数据管理意识。尤其应注重培养研究人员的数据安全、数据道德和知识产权意识。同时还应加强国际交流,学习国外科学数据管理经验和成果。建立全面、行之有效的人才培养机制,组建专业的建设队伍,促进我国科学数据管理、分析、利用、共享等工作的有效开展,对提高我国科研产出效率、促进研究成果的转换有着十分重要的意义。

5 结语

综上所述,为提高我国科学数据管理平台的建设水平,需进一步加强国家政策引导、建立健全的平台建设与共享机制、转变平台建设与共享观念、完善标准体系、加强技术支持与人才培养,建设覆盖全球的科学数据管理平台。

参考文献

- 1 陈美. 美国开放政府数据的保障机制研究 [J]. 情报杂志, 2013, 32 (7): 148 - 153.
- 2 完颜邓邓, 高峰. 英美澳科学数据存储与共享平台建设现状调查及启示 [J]. 图书馆建设, 2016 (3): 29 - 34.
- 3 李俊俊. 英国科学数据的管理及启示 [J]. 经济管理: 全文版, 2016, 3 (6): 128.
- 4 国务院办公厅. 科学技术期刊管理办法 [S]. 2018 - 3 - 17.
- 5 科技部, 财政部. 国家科技资源共享服务平台管理办法 [S]. 2018 - 2 - 13.
- 6 温芳芳. 国外科学数据开放共享政策研究 [J]. 图书馆学研究, 2017 (9): 91 - 100.
- 7 陈明奇, 姜禾, 张娟, 等. 大数据时代的美国信息网络安全新战略分析 [J]. 信息网络安全, 2012 (8): 32 - 35.
- 8 Bao Y, Federhen S, Leipe D, et al. National Center for Biotechnology Information Viral Genomes Project [J]. Journal of Virology, 2004, 78 (14): 7291 - 7298.
- 9 Brooksbank C, Cameron G, Thornton J. The European Bioinformatics Institute's Data resources: towards systems biology [J]. Nucleic Acids Research, 2005, 33 (Database issue): D46 - D53.
- 10 The President & Fellows of Harvard College. Harvard Dataverse [EB/OL]. [2019 - 01 - 26]. <http://dataverse.harvard.edu/>.
- 11 Stevens H. Globalizing Genomics: the origins of the international nucleotide sequence database collaboration [J]. Journal of the History of Biology, 2017, 51 (1): 1 - 35.
- 12 The National Library of Medicine. A Platform for Biomedical Discovery and Data - powered Health: National Library of Medicine Strategic Plan 2017 - 2027 [EB/OL]. [2019 - 01 - 30]. https://www.nlm.nih.gov/pubs/plan/lrp17/NLM_StrategicReport2017_2027.pdf.
- 13 刘德培. 人口与健康科学数据共享 [C]. 南京: 中华医学会医学工程学分会学术年会暨中华临床医学工程及数字医学大会, 2009.
- 14 Wang Y, Song F, Zhu J, et al. GSA: genome sequence archive [J]. Genomics Proteomics & Bioinformatics, 2017, 15 (1): 14 - 18.
- 15 北京大学健康医疗大数据国家研究院. 中国健康医疗数据共享平台 (C - MEDS) [EB/OL]. [2019 - 01 - 30]. <http://101.200.36.59/index.html/>.
- 16 北京大学公共卫生学院. 中国队列共享平台 (CCC) [EB/OL]. [2019 - 01 - 30]. <http://chinacohort.bjmu.edu.cn>.
- 17 北京大学. 中国肾脏疾病数据网络 (CK - NET) [EB/OL]. [2019 - 01 - 30]. <http://www.kidney.net>.
- 18 Sneddon T P, Li P, Edmunds S C. GigaDB: announcing the GigaScience database [J]. GigaScience, 2012, 1 (1): 11.
- 19 UK Data Service. Research Data Lifecycle [EB/OL]. [2019 - 02 - 05]. <http://data-archive.ac.uk/create-manage/life-cycle>.

《医学信息学杂志》版权声明

(1) 作者所投稿件无“抄袭”、“剽窃”、“一稿两投或多投”等学术不端行为,对于署名无异议,不涉及保密与知识产权的侵权等问题,文责自负。对于因上述问题引起的一切法律纠纷,完全由全体署名作者负责,无需编辑部承担连带责任。(2) 来稿刊用后,该稿包括印刷出版和电子出版在内的版权、复制权、发行权、汇编权、翻译权及信息网络传播权已经转让给《医学信息学杂志》编辑部。除纸质载体形式出版外,本刊有权以光盘、网络期刊等其他方式刊登文稿,本刊已加入万方数据“数字化期刊群”、重庆维普“中文科技期刊数据库”、清华同方“中国期刊全文数据库”、中邮阅读网。(3) 作者著作权使用费与本刊稿酬一次性给付,不再另行发放。作者如不同意文章入编,投稿时敬请说明。

《医学信息学杂志》编辑部